
Cambridge Assessment
Research Matters
Issue 25 / Spring 2018

Cambridge Assessment

Proud to be part of the University of Cambridge

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

Citation

Articles in this publication should be cited using the following example for article 1:
Shaw, S., Kuvalja, M., and Suto, I. (2018). An exploration of the nature and assessment of student reflection. *Research Matters: A Cambridge Assessment publication, 25*, 2-8.

Credits

Editorial and production management:
Karen Barden: Research Division, Cambridge Assessment
Additional proofreading: David Beauchamp, Research Division, Cambridge Assessment
Cover image: John Foxx Images
Design: George Hammond
Print management: Canon Business Services

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

The full issue of *Research Matters 25* and all previous issues are available from our website: www.cambridgeassessment.org.uk/research-matters

Research *Matters* / 25

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

The topics covered in this issue of *Research Matters* seem like a microcosm of education, which is forever seeking stability and yet permanently in transformation. We have theory which explains this state of affairs: Critical realism tells us that in social systems, such as education, things will only happen on a predictable basis when all the factors impinging on it are stable. Which, given the shifts in youth culture, the economy, families and so on, seldom holds true for long. This necessary feature of social systems commits policymakers to constant evaluation, fine-tuning, innovation and optimisation. Unintended consequences and collateral impacts are legion.

In this issue we have articles which focus on the impact and effect of core aspects of contemporary policy as well as potential innovations which go beyond common assumptions. Balancing stability and innovation is a constant challenge. We know from the historical record that stability in assessment has assets: Public confidence can accumulate; understanding of qualifications can grow in society; and learning programmes can carefully be refined and enhanced. Conversely, repeated changes can undermine confidence; can cause the hard work behind lesson plans and resources to be redundant; and introduce confusion into how assessments should be used and interpreted. In various papers over the past decade, researchers at Cambridge Assessment have argued that, too frequently, qualifications are seen as 'the thing to change' as a means of implementing wider policy aspirations – not least because they are relatively easy to change, compared with other key factors in education and training arrangements.

Undue change indeed decreases capacity in education. But holding on too long to things which are known to be problematic or defective has a bad history – the '5 grade A*-C' performance measure for GCSEs; some vocational equivalents to GCSE; high levels of coursework in a setting of hyper-accountability. Sound research – well constructed in its focus, method, scheduling, and reporting – is an essential foundation to policy which can achieve this balance between innovation and stability.

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

Editorial

There is currently much interest in the '21st century' or 'transversal' skills that young people need to acquire in order to be ready for the workplace and life in general. Among these is 'reflection'. Like many of the transversal skills, it is difficult to define and even more difficult to assess. In the first article of this issue, Stuart Shaw, Martina Kuvalja and Irenka Suto describe how reflection has been conceived in the education and assessment literature, and show how it can be assessed, at least in part, in a high-stakes context. In the second article, I consider possible justifications for using fixed pass marks. They have the advantages of simplicity and transparency, but can these outweigh potential unfairness when tests vary in difficulty?

Vicki Crisp in the third article investigates experimentally the process by which schools ensure that their teachers are marking non-examined assessments to the same standard ('internal moderation') prior to the work being externally moderated by the awarding body.

The fourth article by Carmen Vidal Rodeiro contributes to the current debate on whether students who do not achieve a 'good pass' in GCSE Mathematics and English should have to retake them in the sixth form. She finds some evidence that those who retake do slightly worse in their Level 3 qualifications than comparable students who do not. In the fifth article, Tom Benton shows how to go about predicting the number of students who will achieve 'straight' grade 9s in the reformed GCSEs. It is not straightforward! Many readers will probably be content just to wait and see….

The final article addresses an important aspect of reform to A levels: What is the effect on the transition from school/college to higher education? Simon Child and colleagues use observation of additional support classes in Biology at three universities (carried out prior to the reforms) plus interviews with A level teachers, undergraduates and lecturers, to shed light on the issues from different perspectives.

Tom Bramley *Director, Research Division*

An exploration of the nature and assessment of student reflection

Stuart Shaw Cambridge Assessment International Education, **Martina Kuvalja** OCR, and **Irenka Suto** Research Division

(The study was completed when the second author was based at Cambridge Assessment International Education)

Introduction

Reflection is often considered to be one of the so-called '21st century' or 'transversal' skills, or 'life competencies'. Many societies value people who can reflect upon their own beliefs and experiences in the classroom and beyond, and learn from them. It is also important to be able to contemplate the work of others at a deep level. In this article, we review some of the academic literature on reflection and explore ways in which it is assessed in educational contexts. Cambridge Assessment International Education offers the General Certificate of Education Advanced Subsidiary level (GCE AS level) Global Perspectives and Research: This serves as a case study for how reflection can be assessed as part of a taught curriculum.

An early definition of reflection

The American philosopher and educational reformer John Dewey (1859–1952) was one of the first to articulate the idea of reflective thinking. He is often regarded as the father of experiential learning, famously observing, "We do not learn from experience. We learn from reflecting on experience." Dewey defined reflection as "Active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it and the further conclusions to which it tends" (p.118).

Dewey's work has been studied widely by philosophers and has proven particularly popular with educationalists in his home country. Rodgers (2002), for example, deconstructs Dewey's concept of reflection as:

- a "meaning-making process that moves a learner from one experience into the next, with deeper understanding of its relationships with, and connections to, other experiences and ideas …" (p.845);
- a "systematic, rigorous and disciplined way of thinking" (p.845);
- a social phenomenon which happens in the community, through interaction with others;
- requiring "attitudes that value the personal and intellectual growth of oneself and others" (p.845).

The breadth of Dewey's definition of reflection (and its characteristics) has facilitated its adoption in multiple disciplines, where it has been used to construct different models of development.

Self-reflection versus reflection upon other material

In education, it is useful to distinguish self-reflection from reflection upon other material. Students can reflect upon their own learning, which includes their personal experiences, perspectives, beliefs and claims. Alternatively, but often additionally, they can reflect upon the experiences, perspectives, beliefs and claims of others, and on study material presented as factual knowledge. Hereafter, we refer to this second type of reflection as 'reflection upon other material'.

Reflection and critical thinking

Reflection upon other materials is sometimes, but not always, regarded as an element of critical thinking. For example, McPeck (1981) defines critical thinking as: "The propensity and skill to engage in an activity with reflective skepticism" (p.8). Ennis (1985) describes critical thinking as "reflective and reasonable thinking that is focused on deciding what to believe or do" (p.45). The relationship between reflection and critical thinking is arguably somewhat circular, however, as it is also possible to use other critical thinking skills such as analysis and evaluation during both self-reflection and reflection upon other material.

This is evident within Mezirow's concept of 'critical reflection'.[1] In his influential theory of 'transformative learning' (Mezirow, 1997), we are encouraged to view learning as a process of (i) becoming aware of our own assumptions and (ii) revising them. Among transformative learning theorists, critical reflection is "The means by which we work through beliefs and assumptions, assessing their validity in the light of new experiences or knowledge, considering their sources, and examining underlying premises" (Cranton, 2002, p.65).

Mezirow (1997) claimed that our frames of reference can be transformed through both 'subjective reframing' (which entails critical self-reflection) and 'objective reframing' (which entails critical reflection upon other material). For example, both types of reframing could occur when a student explores an historical period from the perspective of another nation, or when a student is introduced to a new method of solving a mathematical problem. There may be a single learning event that serves as a catalyst for transformation. Alternatively, the process may be much more gradual, occurring through a series of events both within and beyond the taught curriculum.

1. The origins of this concept can be traced back to *Critical Theory* which was developed by Adorno (1998/1969) amongst others (Marcuse, 1969; Horkheimer, 1972). This school of thought emphasises the reflective assessment and critique of society and culture by applying knowledge from the Social Sciences and the Humanities.

Self-reflection, self-regulation and metacognition

There are at least two further traditions within educational and developmental psychology which include self-reflection in their conceptualisations of learning. The first of these is the Vygotskyian tradition: This takes a 'socio-cultural' approach to exploring the self-regulation of learning and the internalisation of regulatory processes (Vygotsky, 1986). In the context of a goal-directed activity, self-regulation is considered to comprise: planning, monitoring (keeping track of the activity), updating progress, control (retaining or changing an action as needed), and contemplation of the outcomes (Pintrich, 2004; Pintrich & Zusho, 2002; Schunk & Zimmerman, 1994; Schunk, 2005). Self-reflection represents the final stage of this self-regulatory cycle, when a student reviews and evaluates his or her own performance in relation to the original goal. Vygotsky believed that psychological development emerges through interpersonal connections and interactions with the social environment, with language playing a crucial role in this process. Self-regulation, including self-reflection, would therefore be evident in joint learning activities.

The link between various manifestations of cognitive self-regulation and academic achievement is well documented for secondary school students (Zimmerman, 2002). However, neo-Vygotskians also emphasise that in addition to regulating their own cognitive processes, students need to regulate their own emotional responses, motivational states, and the contexts in which their learning occurs. Behaviours which evidence emotional and social self-regulation include delaying gratification (inhibitory control), persevering with tasks, and displaying appropriate manners.

In the second tradition, cognitive psychologists are attempting to expand a fine-grained understanding of a set of executive functions which enable the successful metacognitive regulation of one's own performance. Metacognition is the process of thinking about one's own cognitive, emotional, motivational and social functioning (Efklides, 2008). It plays a crucial part in critical thinking (Magno, 2010). For example, if a student is asked to analyse, evaluate, and synthesise material on a topic of interest, he or she needs to be able to do so with as little bias as possible. This is possible through a constant monitoring process (metacognitive monitoring – Flavell, 1981a; 1981b) which involves self-reflection. If bias is detected, then the student can engage in self-control (self-regulation) and re-evaluate his or her own conclusions on the studied topic. Critical thinking can therefore involve both self-reflection and reflection upon other material concurrently. Cambridge Assessment's own definition of critical thinking includes self-reflection in this sense (Cambridge Assessment, 2007).

Findings are consistent on the positive influence of both naturally emerging and taught metacognitive and self-regulatory behaviours from an early age through to undergraduate level study; such behaviours lead to better academic performance (Chemers, Hu, & Garcia, 2001; Forman & Cazden, 1985; Palinscar & Brown, 1984; Siegler, 2002). An effective way of encouraging self-regulation and metacognitive thinking in students is through providing them with opportunities to practise these aspects of learning, and to reflect further upon that practice (Nicol & Macfarlane-Dick, 2006).

Assessing reflection

When we speak of reflection, we are referring to opaque higher-order thinking processes that are, by their very nature, difficult to assess. As with assessments in traditional academic subjects, we can design tasks to elicit behaviours that require students to use these internal thought processes. We must also then define clearly the indicators of such processes.

When assessing reflection, it could be argued that there is a risk of assessing merely the ability to remember and report (that is, memory and writing or oral skills) rather than all of the mental processes that constitute reflection. This could lead to confusion for students in knowing the criteria on which they are being assessed (Wilson, 2013). It is also difficult to establish whether reflections (irrespective of how they are captured) resemble students' authentic experiences (Ryan & Ryan, 2013). There is an argument that reflection should not be assessed in an educational setting at all, although it should be established and nurtured in the classroom. For example, Ixer (2016) claims that by attempting to assess reflection we are distorting the construct.

However, there are several good reasons why assessing reflection remains desirable. Firstly, students often focus on assessment in their learning, and their learning becomes motivated by assessment (Watkins, Dahlin, & Ekholm, 2005). The assessment of reflection may therefore increase the value of reflection in the eyes of students. Secondly, a related function of assessment can be to make student learning visible. A third important function of assessment is diagnosis. In this regard, assessment can be part of a process used to determine students' strengths and weaknesses. It may therefore be needed to identify students who struggle with reflection.

Arguably, many examinations in traditional subjects include the covert assessment of reflection upon other material because they assess critical thinking skills. For example, History and English Literature examinations frequently require students to reflect upon sources and literary excerpts and evaluate them in multiple respects. In these subjects and others, examination questions that begin with the classic opener 'Compare and contrast…' usually require students to reflect in this sense. Similarly, Science examinations may require students to reflect upon the outcomes of experiments when interpreting their findings. Perhaps more explicitly, the OCR awarding body offers General Certificate of Education Advanced Subsidiary and Advanced level (GCE AS and A level) Critical Thinking, which are skills-based, rather than content-based.[2] Cambridge Assessment International Education assesses critical thinking explicitly within its AS and A level Thinking Skills.

When it comes to self-reflection, the most widely used method of assessment is reflective writing (e.g., Barney & Mackinlay, 2010; Carrington & Selva, 2010; Fitzgerald, 2009; Ghaye, 2007; McGuire, Lay, & Peters, 2009; Moon, 2013). This can take a variety of forms. For example, learning portfolios have been found to encourage reflective thinking per se (Scott, 2009) and can be used in assessment. They enable students to document, store and review their work. The portfolios can then be used by teachers to analyse students' strengths and weaknesses in depth, particularly for formative purposes (Fernsten & Fernsten,

2. Over the past few years AS and A levels were redeveloped nationally. Unfortunately it was not possible to develop Critical Thinking content that met the national regulator's principles for reformed AS and A levels. The final assessment session opportunity for first time candidates is therefore Summer 2018, with resits available in 2019.

2005). Within learning portfolios, reflective journals and log books are often used to record self-reflection in regard to the overall learning experience (Ghaye, 2007). Recording reflective thinking in such a manner has been found to have a positive impact on students' overall metacognitive and other critical thinking skills (Naber & Wyatt, 2014), which might in turn have a positive 'knock-on' effect on students' learning performances (McCrindle & Christensen, 1995; Mauroux, et al., 2015; Nückles, Hübner, & Renkl, 2009).

When writing in their reflective journals, students are usually encouraged to record their reflections as they occur, or as soon as possible afterwards. In this way students avoid relying on their memory and retrieving this information after the internal authentic reflective process has already happened. This approach should also reduce a student's temptation to 'fill in' their memory gaps with false information and inauthentic experiences. This may reduce a key threat to validity for this kind of assessment.

Reflective papers provide a means of assessing both reflection upon other materials and self-reflection. Students are often given a topic and stimulus materials upon which they have to critically reflect (framing their reflection 'objectively'). They are also expected to demonstrate reflections upon their own initial and (potentially) changed perspectives, which occur as a result of researching a topic ('subjective framing'). Reflective papers and essays can be highly structured or unstructured, giving students the opportunity to engage in reflection in a unique way.

Operationalising the construct of reflection for assessment purposes: a case study

In this section we explore an example of the 'reflective paper' approach to assessment. The AS level Global Perspectives and Research is a skills-based programme of study offered by Cambridge Assessment International Education. It assesses reflection as part of a taught curriculum, and its aim is to encourage students to think about and explore issues of global significance. Global Perspectives and Research students are expected to engage in metacognitive and critical thinking with regard to their own perspectives and understanding of a topic, as well as those of others. Schunk and Zimmerman (1994) and Whitebread, et al. (2009) have argued that such skills are crucial for the development of independent and self-regulated individuals who are capable of collaborating and co-operating with others.

AS level students are assessed via three compulsory components:

1. A written examination
2. An essay
3. A team project.

For the team project, students work in teams to identify a local problem which has global relevance. Individual team members research the issues and suggest solutions to the problem based on their research findings. Working together, a set of proposed team solutions to the problem is agreed. While the focus is on teamwork, each student within a team prepares two pieces of work for individual submission: an 8-minute presentation of their individual research and proposed solutions to the problem (which is delivered to an audience), and an 800-word reflective paper.

The reflective paper gives students the opportunity to consider the process they have undertaken in researching and producing their individual presentation as part of a team. As such, it is their chance to provide evidence for reflection, which is Assessment Objective 2 (AO2), and the collaboration aspect of Assessment Objective 3 (AO3). For AO2 (reflection), students are assessed on their ability to:

- research and consider alternative perspectives objectively and with empathy;
- consider the ways in which personal standpoints may have been affected by the research process;
- evaluate the impact of alternative perspectives and conclusions on personal standpoint; and
- identify the need for further research in light of the research findings.

The reflective paper is assessed externally. It accounts for 10 marks of the total of 100 for the whole AS level: There are 5 marks for AO2 and 5 marks for AO3. There are two assessment criteria: the first relates to considerations of one's own perspective, belief and knowledge (Mezirow's [1997] 'subjective framing'), and the second relates to considerations of other's perspectives, beliefs and knowledge ('objective framing'). Each criterion is assigned a level from 1 to 5 when marking (see Table 1).

Table 1: The Cambridge Assessment International Education Global Perspectives and Research Reflective Paper mark scheme

Level	*Marks*	*Indicative descriptors*
5	9–10	• The candidate engages in a **probing and critical** evaluation of their own practice in working with others to identify a local problem and explore possible solutions. • The candidate reflects **fully** on how their personal standpoint and scope for future research have been affected by alternative team and research perspectives.
4	7–8	• The candidate engages in **some effective** evaluation of their own practice in working with others to identify a local problem and explore possible solutions. • The candidate undertakes **some clear** reflection on how their personal standpoint and scope for future research have been affected by alternative team and research perspectives.
3	5–6	• The candidate evaluates to **some extent** their own practice in working with others to identify a local problem and explore possible solutions. • The candidate undertakes **some** reflection on how their personal standpoint and scope for further research have been affected by alternative team and research perspectives.
2	3–4	• The candidate **attempts** to evaluate their own practice in identifying a local problem and exploring possible solutions, but **may lack** consideration of their work with others. • The candidate **attempts** to reflect on how their personal viewpoint and scope for further research, but **may lack** a consideration of alternative team or research perspectives.
1	1–2	• The candidate shows **limited** evaluation of their own practice and **lacks** consideration of their work with others. • The candidate shows **limited** reflection on their personal viewpoint and scope for further research and **lacks** any consideration of alternative team or research perspectives.
0	0	No creditworthy material has been submitted.

The reflective paper needs to be understood as a separate and intellectually demanding piece of work, where students undertake two distinct tasks. Firstly, they need to evaluate the effectiveness of the

way in which the group worked together in undertaking their research. Secondly, they also need to consider how their own views were challenged or developed by engaging with the alternative perspectives suggested by other team members (or other perspectives and solutions they located in the research that they undertook). More able students are expected, therefore, to evaluate and make judgements on their performances, going beyond just descriptions of what they did. This makes the activity a rewarding yet challenging task to accomplish successfully. The two assessment criteria can be conceptualised as questions against which the reflective paper is judged, and it is expected that stronger performances ensure that both questions are addressed, using discrete sections. For example:

1. How well has the student evaluated their own practice in working with others to identify a problem and explore possible solutions?

The focus here is on the student's evaluation of their own practice in working with others. This should go beyond what the group did and focus on areas that worked well and/or were less successful before making a reasoned judgement on the success of the group work. Thus, the reflective paper affords an opportunity for self-reflection leading to personal transformation (Mezirow, 1997). The following is an example from a student's reflective paper of a simple but effective approach to outlining an aspect of teamwork. The student then highlights strengths and weaknesses, and the actions taken as a result:

> *Within our group we partnered into pairs and assigned each pair with two of the four components we wanted to cover. Then the two members within those pairs would assign one of the components to each other. By doing this every group member had one of the aspects that they were responsible for researching and after all the information was gathered. We shared that information amongst one another. This was a very effective strategy because everyone in our group executed their assigned job with sufficiency and managed to provide everyone with useful information and resources needed to successfully complete our assignment in a timely manner. However, there was one minor issue that came across our group when using this method. Being able to copy and paste information was simple but being able to paraphrase and combat text chunkiness called for a bit more effort. Several of us struggled trying to avoid gathering twelve pages of information. We came to the conclusion that we needed to do better with only gathering the most important and helpful information.*

2. How well has the student reflected on the extent to which their own standpoint and the scope for future research have been affected by alternative perspectives from within their team and from additional research?

The focus for this second part of the assessment is on the impact of alternative perspectives. Students need to identify what those alternative perspectives are, and to assess the extent to which they have made an impact on their own point of view. The following is an example of the clear identification of how other team members have affected the student's position:

> *My point of view was strengthened because through research I discovered that strict immigration laws would be the most simple and most easy to follow. But, with further analysis into other perspectives such as unilateral immigration, presented by [Student B], made me realize that the strictness of laws may not be the best way to handle the solution. In that way was how I determined that [Student C]'s solution would be the most appropriate.*

The purpose of the reflective paper is to evaluate, not just describe, the student's experiences. Simply listing alternative perspectives, or the different aspects of the research, or the solutions reached, is not sufficient to be awarded a high mark. What is required is a reflection on how these things impacted on the student's own work. The following extract makes a clear transition between the two, demonstrating how the formulation of the team solution also developed their own understanding in specific ways:

> *After creating our group solution I felt that I had learned a lot about the economic, political and ethical themes within the subject of homelessness. Previously I only thought that people became homeless due to problems with drug addictions and a lack of money. However, now I am aware of the legal demands and other governmental requirements people have to go through before receiving a house.*

This is a clear illustration of how the student has critically reflected on the study material and on others' point of view with consequent self-regulation.

Able students appreciate the difference between evaluation and narration when it comes to writing about their practice in working with others. An account of what happened is not the same as an identification of working practices and a judgement on their strengths and weaknesses. In the following extract, the student begins by identifying the benefits of the high level of agreement among team members:

> *This level of cooperation was a welcome experience, however, I feel that the lack of any dissenting opinions and an effective devil's advocate possibly weakened the collective brainpower used in selecting our issue. When a group is so readily agreeable then there is the possibility of a stagnation of perspectives, which also limits possible conversation about solutions and paths to take.*

Here, strengths are weighed up against weaknesses in order to reflect on the wider implications for the effectiveness of collaboration and to make a judgement on it.

Strong reflective papers are clear about the specific strengths and weaknesses of the contribution made by other team members and the student's own experience of working with them:

> *I found working with [Student A] was good but also had its challenges. We were able to come together well and decide on a good topic and question. We were also able to connect on an intellectual level both of us being well educated students. I struggled a little bit at the start to form some points and find good information for my argument but [Student A] was able to suggest some idea as well as a few sources that could assist me. The only challenges I had with [Student A] were our ability to clearly communicate with each other and the fact that he or I were away from class frequently. Sometimes I found that he was a bit unclear with his arguments and so I was unable to form strong counter arguments. I also found it difficult for us to understand each other's standpoints as we were away quite a few times and so could not explain our perspectives and reasoning.*

One element of self-regulation is the evaluation of the progress and the outcome of the individual or joint goal-directed activity. It is

important for students to specifically identify and assess the impact of other perspectives on their own learning and views (Forman & Cazden, 1985). These perspectives may come from research that they have undertaken or from the findings of other team members, as in this extract related to a team project on the internet:

> *My individual standpoint about the effects of the internet has been affected by both my own and my teammates' perspectives. I knew the internet had several negative effects before I started researching the topic. However I did not know the details about the many negative effects on our social lives and the many negative effects of the internet on both our mental and physical health. Therefore, my own findings for the social perspective affected my view on the topic. I was definitely astonished by the findings of my teammate who covered the medical perspective. I did not know much about the medical problems the internet can cause, and felt that it was very interesting and important to know, as it affects a lot of people almost every day.*

Successful students have knowledge about themselves as students, including their strengths and weaknesses (Flavell, 1981b). In this extract, we observe metacognitive processing: The student identifies what they knew or thought before, the new information they have acquired, and how their understanding has changed. Metacognitive insights of this kind inform understanding of self and the learning process, and change and improve personal behaviour accordingly.

The next extract provides a good example of the 'place' of reflection. After evaluating the strengths and weaknesses in the team's work together, the student reflects upon the relative strengths of different team members' solutions to problems in the prison system in New Zealand:

> *We came to the decision that the best solutions were [Student B's] solutions. It was fairly obvious that these solutions were the best from the beginning as they were the solutions that appeared to produce the best results and dealt with the roots of the crimes as opposed to dealing with the prisoners after they had already committed the crime and been put in prison. By taking an approach that looked at the core issues resulting in a higher rate of crimes and then finding a solution, [Student B] was able to develop three key resolutions to stopping crimes in the first place.*

In this example, the effectiveness of the student's ability to reflect is shown in their willingness to acknowledge and precisely articulate the greater strength of another team member's solution, making a desirable outcome more likely (Halpern, 2003). The student has reflected critically on the reasons why a solution is effective. Strong reflective papers go beyond simply saying what everyone's role was, to thinking more closely about how individuals exploited their strengths, added breadth and depth to the arguments, and considered others' views before coming to a group solution.

It is important to note that reflective papers can only score Level 3 or higher when they evaluate the process of collaboration. This means identifying strengths and weaknesses, then reaching a judgement. Reflective papers which simply provide a narrative of what the team did, however fluently this is expressed, will not be able to do this. This conclusion to a student's evaluation of their team's collaboration is a good example of how this can be done in a straightforward but effective way:

> *Thus, our strengths were we made use of our time when needed, and creative thinking played a good part and our weaknesses were the inability to exchange ideas efficiently and lack of motivation during some periods of the completion of the project; these factors altered the rate at which things were completed. As a result, I think the next time, as a group, we should hold more after school and weekend meetings to completely discuss the ins and outs of the problem along with each solution.*

In the final extract, a reflective paper has been reproduced in full. It is a good example of how the quality of a response can benefit if a student is focused and detailed in evaluating their experiences of teamwork. The student takes care to explain the factors which had a negative impact on the team working effectively. They also do well in evaluating the impact of this on their project as a whole. This means that the response meets the requirements for Level 4 on the first criterion (see Table 1).

However, there appears to be no reference at all to their personal standpoint on the topic itself, or how that standpoint was affected by the other team members or their research. Therefore, the student has not demonstrated that they have critically reflected upon the assumptions and alternative perspectives others have taken in proposing their theories about the phenomenon under investigation. As a consequence, the paper can only receive a mark of 0 for the second criterion. This inconsistent profile of performance leads to a Level 2 achievement overall. The quality of the student's evaluation of the teamwork is such that a mark at the top of Level 2 would be the best fit.

> *In all honesty, I feel as if the communication between my partner and I were* (sic) *not the best. Of course we have talked through how we would structure our presentation, who would write and present about what and have supported each other throughout the whole project, when we needed it. This group project was not the hardest, yet again, it was not the easiest. In comparison with other groups in our Global Perspectives class, there was only two of us in the AS level class, which limited our options to choose a partner and limited our numbers in a group. In a way I do not feel that it is fair to say that we had a disadvantage just because our group was made up only two people. Although it was definitely a lot more work that what others had to do. Or at least I feel as if it was like that. The project itself did take some time to work on for both of us. The paragraphs I have written are from both our perspectives of the good things, the bad things and the things we could have improved in. There were a couple of things that we had agreed on that we did well and a couple of things we agreed on what we did good and what we did wrong and things we agreed on which could have been improved. It was not easy.*
>
> *I think that we gave each other some good ideas of what we could write and talk about. As well as that, I think that we've been in on track of what we've been needing to do. The support and the understanding of each other was just fine. As struggles of not having such a big group in comparison to other groups in the class, I honestly think that we worked quite well (individually). It is honestly not easy having only one other person in the group, giving us almost double the work for the both of us. Having such a big topic – Global Warming – and for only two people takes a lot of effort and a lot of time, but I believe that we worked through it with good hope for the outcome.*
>
> *Although, there were a couple of disadvantages to this project. I do not think that we've had enough communication. That we sat in the same*

room but did not explain to each other everything about what we actually wrote down in details. Even though we talked a bit about what we have written down but I do not feel as if it was enough. It was difficult to actually work together or talk about the project together outside school. We do not live in the same house or part of school so it was difficult to figure out hours to which we would be working together. As well as figuring out when we could work together, we've been getting quite a load of homework and quite a lot of activities, so it takes up a lot of our free time. I do not feel that Facebook or Skype would have helped a great deal but of course it would have made a some kind of difference. I think that making a schedule of when we're going to work together, and how many hours would have made our planning and our work more efficient and we may not have been in a hurry. Although, even if I hadn't started on my group work right away, I had finished my part of the presentation before my partner did, but that was because I had actually worked on it in the time we had, using up my hours wisely, unlike my partner.

There is always room for improvement! I can think of a couple of things that we have agreed on that we could have done better. Looking back at what i've written in the above paragraphs, I can say that lack of communication could be improved. Togetherness could be improved as well (using time out of school to work on the group of project). Though, with the struggle of only being two in the group, I think that learning to cope with double as much work is a good strategy that we will have to grow with. This would have given us an advantage of polishing our work with hopes for our reader's satisfaction. I admit that I; and my partner, did not start right away. It did take me, at least, a couple of weeks to start writing, though for the couple of weeks before actually starting the project, I was wondering what would be in it, what I would write, what I would say as well as the improvement and the ups and downs of this project.

Conclusions

Ensuring that students succeed in the 21st century requires fresh thinking about what knowledge and competencies are, and how they should be supported throughout education. In this article, we have looked briefly at key conceptualisations of reflection within the academic literature. Whilst there is a degree of circularity in the definitions of some key terms, it is clear that skills in both self-reflection and reflection upon other material are valued highly in several schools of thought which have not always aligned historically. Despite the difficulties of assessing reflection as authentic student experience, we have considered the reasons why it is nevertheless important to do so, and have offered a practical example of how it can be done. It is hoped that studies of this kind will bring greater clarity to test designers and developers when they are defining and operationalising the construct of reflection and to teachers and students who wish to focus on reflection within their curricula.

Acknowledgements

The authors would like to thank Lorna Stabler, formerly of Cambridge International Examinations, for her initial review of the literature.

References

Adorno, T. W. (1998/1969). *Critical models: Interventions and catchwords*. NY: Columbia University Press.

Barney, K., & Mackinlay, E. (2010). Creating rainbows from words and transforming understandings: Enhancing student learning through reflective writing in an Aboriginal music course. *Teaching in Higher Education, 15*(2), 161–173.

Cambridge Assessment. (2007). *Critical Thinking: Deriving the Definition. Factsheet 1*. Retrieved from http://www.cambridgeassessment.org.uk/Images/109971-critical-thinking-factsheet-1.pdf

Carrington, S., & Selva, G. (2010). Critical social theory and transformative learning: Evidence in pre-service teachers' service-learning reflection logs. *Higher Education Research & Development, 29*(1), 45–57.

Chemers, M., Hu, L., & Garcia, B. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology, 93*(1), 55–64.

Cranton, P. (2002, Spring). Teaching for transformation. In J.M. Ross-Gordon (Ed.), *New directions for adult and continuing education: No. 93. Contemporary viewpoints on teaching adults effectively* (pp.63–71). San Francisco, CA: Jossey-Bass.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educational process*. Lexington, Massachusetts: D.C. Heath and company.

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*(4), 277–287.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership, 43*(2), 44–48.

Fernsten, L. & Fernsten, J. (2005). Portfolio assessment and reflection: Enhancing learning through effective practice. *Reflective Practice, 6*(2), 303–309.

Fitzgerald, C. (2009). Language and community: Using service learning to reconfigure the multicultural classroom. *Language and Education, 23*(3), 217–231.

Flavell, J. H. (1981a). Monitoring social cognitive enterprises: something else that may develop in the area of social cognition. In J. Flavell, & L. Ross (Eds.), *Social Cognitive Development: Frontiers and Possible Futures* (pp.272–287). Cambridge: Cambridge University Press.

Flavell, J. H. (1981b). Cognitive monitoring. In W. P. Dickson (Ed.), *Children's oral communication* (pp.35–60). New York: Academic Press.

Forman, E. A., & Cazden, C. B. (1985). Exploring Vygotskian persepctives in education: The cognitive value of peer interaction. In J. W. Wertsch (Ed.), *Culture, Communication and Cognition: Vygotskian Perspectives* (pp.323–347). Cambridge, UK: Cambridge University Press.

Ghaye, T. (2007). Is reflective practice ethical? (The case of the reflective portfolio). *Reflective Practice, 8*(2), 151–162.

Halpern, D. F. (2003). *Thought and Knowledge: An Introduction to Critical Thinking*. London: Lawrence Erlbaum Associates.

Horkheimer, M. (1972). *Critical theory: Selected essays (Vol. 1)*. New York: Continuum.

Ixer, G. (2016). The concept of reflection: is it skill based or values? *Social Work Education, 35*(7),809–824.

Magno, C. (2010). The role of metacognitive skills in developing critical thinking. *Metacognition and Learning, 5*(2), 137–156.

Marcuse, H. (1969). *An essay on liberation* (Vol. 319). Beacon Press.

Mauroux, L., Zufferey, J. D., Rodondi, E., Cattaneo, A., Motta, E., & Gurtner, J. L. (2015). Writing reflective learning journals: Promoting the use of learning strategies and supporting the development of professional skills.

In M. Bétrancourt, G. Ortoleva, & S. Billett (Eds.), *Writing for professional development* (pp.107–128). Leiden, The Netherlands: Brill.

McCrindle, A. R., & Christensen, C. A. (1995). The impact on learning journals on metacognitive and cognitive processes and learning performances. *Learning and Instructions, 5*(2), 167–185.

McGuire, L., Lay, K., & Peters, J. (2009). Pedagogy of reflective writing in professional education. *Journal of the Scholarship of Teaching and Learning, 9*(1), 93–107. Retrieved from http://files.eric.ed.gov/fulltext/EJ854881.pdf

McPeck, J. E. (1981). *Critical Thinking and Education*. Toronto: Oxford University Press. pp.v1, 170.

Mezirow, J. (1997). Transformative learning: Theory to practice. *New Directions for Adult and Continuing Education, 74*, 5–12.

Moon, J. A. (2013). *Reflection in learning and professional development: Theory and practice*. Oxon: Routledge.

Naber, J., & Wyatt, T. H. (2014). The effect of reflective writing interventions on the critical thinking skills and dispositions of baccalaureate nursing students. *Nurse Education Today, 34*(1), 67–72.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Nückles, M., Hübner, S., & Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction, 19*(3), 259–271.

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction, 1*(2), 117–175.

Pintrich, P. R. (2004). A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students. *Educational Psychology Review, 16*(4), 385–407.

Pintrich, P. R., & Zusho, A. (2002). The development of academic self-regulation: The role of cognitive and motivational factors. In A. Wigfield, & J. S. Eccles (Eds.), *Development of Achievement Motivation* (pp.249–284). San Diego, CA, US: Academic Press.

Rodgers, C. (2002). Defining reflection: Another look at John Dewey and reflective thinking. *Teachers College Record, 104*(4), 842–866.

Ryan, M., & Ryan, M. (2013). Theorising a model for teaching and assessing reflective learning in higher education. *Higher Education Research & Development, 32*(2), 244–257.

Schunk, D. H. (2005). Self-regulated learning: The educational legacy of Paul R. Pintrich. *Educational Psychologist, 40*(2), 85–94.

Schunk, D. H., & Zimmerman, B. J. (1994). *Self-regulation of learning and performance: Issues and educational applications*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Scott, S. G. (2009). Enhancing reflection skills through learning portfolios: An Empirical Test. *Journal of Management Education, 34*(3), 430–457.

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott, & J. Parziale (Eds.), *Microdevelopment: Transition Processes in Development and Learning* (pp.47–59).

Vygotsky, L. S. (1986). *Thought and language (Newly revised and edited by Alex Kozulin)*. Massachusetts: MIT.

Watkins, D., Dahlin, B., & Ekholm, M. (2005). Awareness of the backwash effect of assessment: A phenomenographic study of the views of Hong Kong and Swedish lecturers. *Instructional Sciences, 33*(4), 283–309.

Whitebread, D. G., Coltman, P., Pasternak, D. P., Sangster, C., Grau, V., Bingham, S., & Demetriou, D. (2009). The development of two observational tools for assessing metacognition and self-regulated learning in young children. *Metacognition and Learning, 4*(1), 63–85.

Wilson, G. (2013). Evidencing reflective practice in social work education: Theoretical uncertainties and practical challenges. *British Journal of Social Work, 43*(1), 154–172.

Zimmerman, B. J. (2002). Achieving academic excellence: A self-regulatory perspective. In M. Ferrari (Ed.), *The pursuit of excellence through education* (pp.85–110). Mahwah, NJ: Erlbaum.

When can a case be made for using fixed pass marks?

Tom Bramley Research Division

Introduction

General Certificate of Secondary Education (GCSEs) and General Certificate of Education Advanced levels (A levels) have sophisticated procedures to ensure that the grade boundaries on examination components are set in places that achieve the goal of maintaining standards over time and between awarding organisations (AOs). Statistical methods currently have a prominent role. The 'comparable outcomes' method of The Office of Qualifications and Examinations Regulation (e.g., Ofqual, 2011; Benton, 2016) produces a target distribution of grades for each examination[1] and the AOs have to set boundaries on the components that result in an overall outcome that does not deviate beyond an allowed tolerance from these targets. Although there are good reasons for using these sophisticated procedures (including the prevention of 'grade inflation', and helping to ensure examinees are not disadvantaged when there is a major or minor system change), they do have drawbacks in terms of the resources required to administer them, both in staff time and in data availability. They are well-suited to the GCSE and A level case where there are only one or two examination sessions a year, large cohorts of examinees of roughly the same age are taking the exams, and large administrative data sets tracking the previous educational achievement of these examinees are available. However, some other high- and low- stakes assessment contexts do not have these advantages. In particular, many vocational and other non-academic assessments (such as the driving theory test) are either available on-demand or have multiple testing sessions, with widely fluctuating cohort sizes and groups of test-takers

1. The target distribution is for those examinees for whom there is a measure of prior attainment: Key Stage 2 score at GCSE, and mean GCSE score at A level.

from a wide range of ages, institutions and educational backgrounds. The AO or testing agency may have no information about the prior or concurrent achievement or ability of the group of test-takers and, in some cases, pre-testing is not possible because of cost or concerns about test security. Furthermore, in many such contexts the pass/fail (or other) decision needs to be made as soon as the test has been marked – and for computer-based tests this can be instantly, which requires the pass mark either to be known before the test is taken, or derivable from the items that were administered (in the cases where tests are compiled 'on-the-fly' or administered adaptively).

In some cases expert judgement can be used to arrive at a pass mark – for example by using a standard-setting method such as the Angoff or Bookmark methods (see Cizek, 2012, for a description of such methods). These methods often involve experts making judgements about the difficulty of test items, and the final decision can involve consideration of the potential impact on pass rates of setting the pass mark at particular scores. However, judgements of item difficulty can be unreliable and, as already noted, in some contexts the pass mark needs to be set before the impact on pass rates is known.

Using fixed pass marks, such as "To pass this test you need to answer 30 out of 40 items correctly" or "To pass this test you need to obtain more than 60 per cent of the available marks" might seem to be a simplistic solution to a complex problem. However, it does have some attractions, (Bramley, 2012), including:

- transparency: Test-takers know before taking the test how well they need to do in order to pass;
- validity of inferences about what test-takers know and can do. If past or example papers are publicly available then stakeholders can inspect these themselves and draw their own conclusions about the capability of someone who has achieved a given percentage of the marks available;
- perceived fairness for the test-taker: They know that their result did not depend on the performance of the other test-takers who happened to take the same test (or the prior attainment of other test-takers). However, this advantage could entirely disappear if different test forms are perceived to differ drastically in difficulty ('my friend got an easy set of questions');
- if the pass mark is fixed at a relatively high level then there is some reassurance that people who pass can actually answer most of the questions of the kind that were asked, which is important for 'consumer confidence' in some cases (e.g., a pass mark of only 50 per cent on knowledge of medical terms or routine procedures might not inspire confidence if it was part of a qualification for surgeons);
- the pass/fail decision can be made instantly (assuming the test is auto-marked); and
- the cost in money and staff time of setting the pass mark by more complex methods could be reduced.

The obvious drawback to using fixed pass marks is that it does not allow for the fact that test forms may vary in difficulty despite best efforts to construct or design them to be similar. The aims of the research described here were to investigate how serious a problem this might be in practice, and to explore the extent to which it could be alleviated by using expert judgement in the test construction process.

How much do tests randomly sampled from an item bank differ in difficulty?

A calibrated item bank[2] of 664 dichotomous items testing a single construct (Thinking Skills) classified into 7 different topic/skill areas was used as the basis for several simulations. The number of items and distribution of difficulties within each topic/skill area are shown in Table 1.

Table 1: Descriptive statistics for item bank (item difficulties in logits)

Topic/skill	*Total # items*	*Mean*	*SD*	*Min*	*Max*
1	122	0.39	0.99	-2.35	2.77
2	102	-0.08	1.10	-2.89	2.84
3	125	0.23	1.02	-2.17	3.64
4	120	0.42	1.15	-3.62	3.29
5	86	0.50	1.00	-2.72	3.73
6	57	0.01	1.13	-2.98	2.09
7	52	-0.10	0.75	-2.37	1.57
Total	**664**	**0.24**	**1.06**	**-3.62**	**3.73**

The simulated scenario was that a 40-item test with a fixed pass mark was to be constructed from this bank, with items from the different topic/skill areas represented according to their proportions in the bank[3]. The bank was 'recentred' by subtracting 0.24 logits from each item's difficulty to make the overall mean zero, and facilitate the interpretation of the minimum ability required to pass. This ability was arbitrarily selected to be 0.7 logits which, according to the Rasch model equation, corresponds to a probability of ≈0.67 of success on the average item. A thousand stratified random samples of 40 items were taken from the bank (stratified to ensure that the correct number of items testing each skill were included) and the 'correct' pass mark was calculated as the expected score that would be achieved by an examinee with an ability of 0.7 logits[4]. The average pass mark across all 1,000 tests was 25.6, so the nearest whole number value for a 'fixed' pass mark of 26 (65%) was taken, and compared with the correct pass mark (rounded to the nearest whole number) on each of the 1,000 tests.

Table 2 shows that 76% of the tests had a pass mark within 1 mark of the fixed pass mark of 26, and that 95% were within 2 marks. Figure 1 shows how the pass marks fluctuated from test to test.

One of the factors that affects how much pass marks fluctuate on tests constructed by sampling in this way is the underlying variability of difficulty in the whole bank. If all the items in the bank were the same difficulty, all tests constructed from it would be too. It is conceivable that different domains of knowledge/skill might differ in the extent to which test items might vary in difficulty. For example, if all the items require straightforward recall of basic factual knowledge gained on the course of study, there might be less reason to expect one item to differ too much from another in terms of difficulty. With that in mind, the entire bank was scaled by a factor of 0.8 to reduce the spread of difficulties and the process previously described was repeated.

2. The items were multiple-choice items calibrated using the Rasch model (e.g., Wright & Stone, 1979).
3. Specifically: 7, 6, 7, 7, 5, 4, 4 items from topic/skill areas 1–7 respectively.
4. This is the sum of expected scores on each item according to the Rasch model.

Table 2: Distribution of (absolute) differences from a fixed pass mark of 26 (full bank of 664 items)

FixedPassMarkDiff	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	319	31.90	319	31.90
1	446	44.60	765	76.50
2	187	18.70	952	95.20
3	40	4.00	992	99.20
4	8	0.80	1,000	100

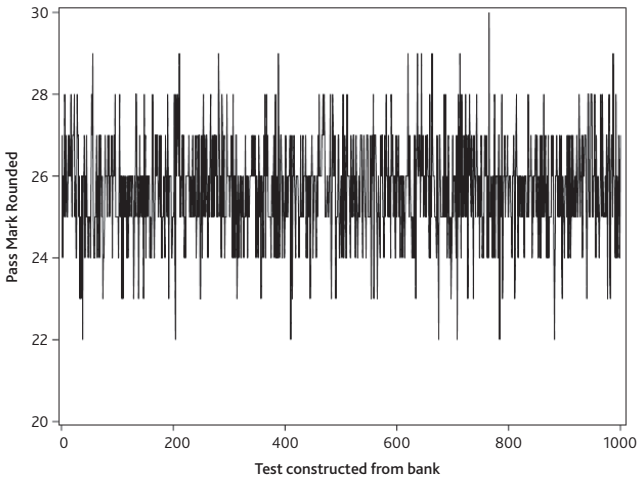

Figure 1: Correct pass marks on the 1,000 tests constructed from the full bank of 664 items

Table 3: Distribution of (absolute) differences from a fixed pass mark of 26 (664 item bank scaled by a factor of 0.8)

FixedPassMarkDiff	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	368	36.80	368	36.80
1	490	49.00	858	85.80
2	126	12.60	984	98.40
3	16	1.60	1,000	100

The scaling reduced the variability of the pass marks – over 85% of tests now had pass marks within 1 mark of the fixed pass mark of 26, and 98% of tests were within 2 marks. Of course, the scaling factor of 0.8 was entirely arbitrary, but this result shows that attempts to reduce the variability of item difficulty could contribute significantly to justifying using fixed pass marks.

The two simulations we have outlined used all the available calibrated items – 664 in total. In some testing contexts (e.g., the development of a new test) there may not be the luxury of such a large pool of items to draw from. A smaller bank of 200 items was therefore created by randomly sampling from topic/skill areas 1 to 6 according to the proportions (20%, 15%, 15%, 20%, 15%, 15%). The new smaller bank therefore had (40, 30, 30, 40, 30, 30) items representing these 6 topic/skill areas. Repeating the sampling process to construct 1,000 new tests gave a new (rounded) mean pass mark of 25, so this was now taken as the fixed pass mark. The resulting pass marks fluctuated in a very similar degree to those from the full bank.

In some contexts there may be rules or reasons preventing the sharing of items across test forms. For example, we could imagine that the 200 items in the smaller bank were constructed with the intention of creating 5 unique 40-item tests. It is therefore interesting to see how much pass marks would vary across sets of five tests (i.e., using every item in the bank) meeting the content specification but containing *no overlapping items*. A thousand such sets of five tests were constructed by random sampling as before (but without replacement). We are now interested in the extent to which the pass marks on each set of 5 tests differ from a set of 5 tests with a fixed pass mark of 25. One way to quantify this is simply to calculate the total absolute deviation across the 5 tests from the pass mark of 25. For example, a set of 5 tests with pass marks of (25, 26, 24, 24, 27) would score a total of 0+1+1+1+2 = 5.

Table 4: Distribution of total absolute deviation (TAD) from a pass mark of 25 across 5 non-overlapping tests in 1,000 sets of 5 tests constructed from the bank of 200 items and 6 topic/skill areas

TAD	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	10	1.00	10	1.00
1	24	2.40	34	3.40
2	102	10.20	136	13.60
3	113	11.30	249	24.90
4	231	23.10	480	48.00
5	149	14.90	629	62.90
6	165	16.50	794	79.40
7	104	10.40	898	89.80
8	61	6.10	959	95.90
9	23	2.30	982	98.20
10	16	1.60	998	99.80
11	2	0.20	1,000	100

Table 4 shows that nearly 63% of the sets had a total absolute deviation of 5 or less. A value of 5 would correspond to being 1 mark away from the fixed pass mark on all 5 (or to other combinations such as 2 above on 1, 3 below on another, and equal on 3). It was very rare (occurring only 1% of the time) for all 5 tests to have the fixed pass mark by chance.

Can expert judgement help to reduce the extent to which test forms differ in difficulty?

In the previous scenario 5 non-overlapping tests were constructed from a bank of 200 items. If the imagined scenario is adapted such that only four tests are needed operationally (with one as back-up for emergencies), then experts could be asked to identify, from the set of five, the four that appear most similar in difficulty (or, conversely,

the test that appears to be most different from the others in difficulty). Table 5 shows that when the most discrepant test from the 5 was removed (using the same data as in Table 4) then the percentage of sets of 4 with a total absolute deviation of 4 or less was nearly 86%, which compares well with the equivalent figure of 63% for the 5 tests. The percentage of sets where all 4 met the fixed pass mark was still low at 3.4%.

Table 5: Distribution of total absolute deviation (TAD) across best 4 non-overlapping tests from a pass mark of 25 in 1,000 sets of 5 tests constructed from the bank of 200 items and 6 topic/skill areas

TAD	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	34	3.40	34	3.40
1	136	13.60	170	17.00
2	223	22.30	393	39.30
3	277	27.70	670	67.00
4	189	18.90	859	85.90
5	85	8.50	944	94.40
6	39	3.90	983	98.30
7	16	1.60	999	99.90
8	1	0.10	1,000	100

Another way of capturing expert judgement of item difficulty might be to ask the item writers to rate individual items (e.g., as being of low-, medium- or high- difficulty. Would tests constructed to be of equal difficulty, in terms of the proportions of items in these three categories, be more likely to be of equal difficulty than tests constructed at random? In order to simulate expert ratings in three categories, a continuous variable was created to be correlated ≈0.7 with the item difficulties. (An average correlation of around 0.6 was reported in Brandon, 2004, between estimates of difficulty in Angoff-type standard-setting exercises and the empirical difficulty values). The top 50 items in the bank according to this variable were assigned a value of '3' (high); the next 75 items '2' (medium); and the bottom 75 items '1' (low). The correlations of this discrete variable with the actual difficulties turned out to be 0.64. This probably represents a slightly optimistic view about what might be achievable with expert judgement.

Figure 2 shows that there was some overlap in the three categories. Nevertheless there was a clear increase in difficulty with the judged category of difficulty. The next step was to construct sets of 5 non-overlapping tests from the bank that not only met the criteria of having the right number of items testing each topic/skill area, but also met the criteria for having the right number of items at each level of judged difficulty (i.e., 10 high, 15 medium, and 15 low). The algorithm written to do this started from a random selection (as before) but then within each test swapped items from over-represented levels of difficulty for items with under-represented levels of difficulty testing the same topic/skill area in the remaining pool of unselected items[5]. This took substantially more computer time to run, so 200 sets of 5 tests were created instead of 1,000.

5. The algorithm was not optimal (in many ways), one way being that the different skills were searched sequentially for items to swap. Thus, 'Skill 1' was always involved in any swapping and 'Skill 6' only very rarely.

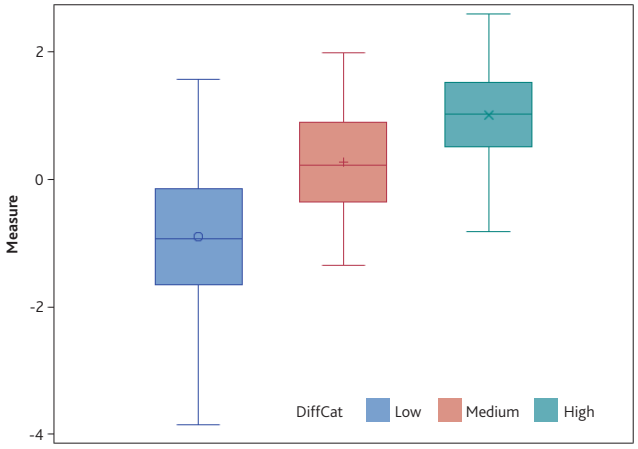

Figure 2: Relationship between 'judged' (simulated) difficulty category (DiffCat) and actual difficulty in the bank of 200 items

Table 6: Distribution of total absolute deviation (TAD) across five non-overlapping tests with the same distribution of judged difficulty

TAD	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	2	1.00	2	1.00
1	7	3.50	9	4.50
2	35	17.50	44	22.00
3	30	15.00	74	37.00
4	58	29.00	132	66.00
5	36	18.00	168	84.00
6	21	10.50	189	94.50
7	9	4.50	198	99.00
8	2	1.00	200	100

Comparing Table 6 with Table 4 shows that there was considerably less deviation of the pass marks. For example, 84% of the sets had a total absolute deviation of 5 or less compared with 63% using random selection.

Table 7 shows that if (after constructing 5 tests with the designated number of items at each level of judged difficulty) it were still possible for experts to identify the one furthest away from the average, then over 95% of sets of 4 would have a total absolute deviation of 4 or less (cf. 86% in Table 5).

Effect of overall ability distribution on fluctuations in pass rate

Finally, the effect on the pass rate of having fixed pass marks (as opposed to pass marks with the 'correct' value according to the bank difficulty) was investigated. The fluctuation in pass rate clearly is likely to depend on the ability (achievement/learning/knowledge) of the examinees in relation to the questions. When setting grade boundaries on A levels, there are usually relatively few examinees around the E boundary, and moving this boundary up or down by a few marks has little effect on the

Table 7: Distribution of total absolute deviation (TAD) across best four non-overlapping tests with the same distribution of judged difficulty

TAD	*Frequency*	*Percentage*	*Cumulative Frequency*	*Cumulative Percentage*
0	9	4.50	9	4.50
1	42	21.00	51	25.50
2	66	33.00	117	58.50
3	45	22.50	162	81.00
4	29	14.50	191	95.50
5	8	4.00	199	99.50
6	1	0.50	200	100

cumulative percentage of examinees achieving grade E. By contrast, there are usually many more examinees near the grade A boundary, and small changes in this boundary can have much larger effects on the cumulative percentage. To investigate the effect of the examinee ability distribution on pass rate fluctuation with fixed pass marks, a 'worst-case scenario' was simulated with a (normal) distribution of ability with a mean of 0.7 logits (i.e., around the pass mark, so 50% of examinees would be expected on average to pass the test) and standard deviation (SD) of 1 logit. Then this distribution was shifted by adding a constant amount such that around 80% of examinees would be expected to pass the test. The scores of 1,000 (different) examinees on each of the (randomly constructed) 1,000 tests from the 200-item bank were simulated using the Rasch model. Figure 3 shows the simulated score distributions for the first of these 1,000 tests.

Table 8: Descriptive statistics for distributions of simulated pass rates

		N	*Mean*	*SD*	*Min*	*Max*
Mean at pass mark	True	1,000	52.44	2.18	45.7	58.3
	Fixed	1,000	52.64	6.42	29.9	69.9
Mean above pass mark	True	1,000	80.31	1.65	75.7	84.6
	Fixed	1,000	80.25	4.54	59.9	90.6

Figure 4 and Table 8 show that there is considerably more variability (SD ≈3 times greater) in pass rates using fixed pass marks from tests constructed at random than from using the correct (true) pass marks, but that, as expected, the variability (and the difference between true and fixed) is less when the bulk of the distribution is some distance away from the pass mark.

Summary and possible rationale for using fixed pass marks

In summary, the simulations have shown that:

- tests constructed by random sampling from an item bank vary in difficulty;
- with a pass mark at around 60–65% of the maximum mark, around 75% of 40-item tests constructed at random from the particular real item bank used as a basis for this work would have a pass mark within 1 mark of the fixed pass mark;
- this percentage would be greater if the items in the bank had a lower spread of difficulty (and vice versa);
- constructing 5 non-overlapping tests (i.e., with no items in common) at random from a bank of 200 items produced around 63% of sets of 5 where the total absolute deviation from fixed pass marks was 5 or less (i.e., an average discrepancy of 1 mark per test);
- this could be increased (to 86%) for 4 tests if experts could infallibly identify the most discrepant test in a set of 4;
- constructing 5 non-overlapping tests to meet criteria of equal difficulty as defined by expert judgement (assumed to correlate around 0.6 with actual difficulty) produced around 84% of sets of 5 where the total absolute deviation from fixed pass marks was 5 or less; and
- the variability in pass rates from tests with fixed pass marks is around three times greater than from tests with the correct pass mark, but the amount of variability (for both) depends on where the distribution of examinee ability is in relation to the pass mark. If the average pass rate is around 80%, the variability (SD) in pass rate is around three-quarters of what it is if the average pass rate is around 50%.

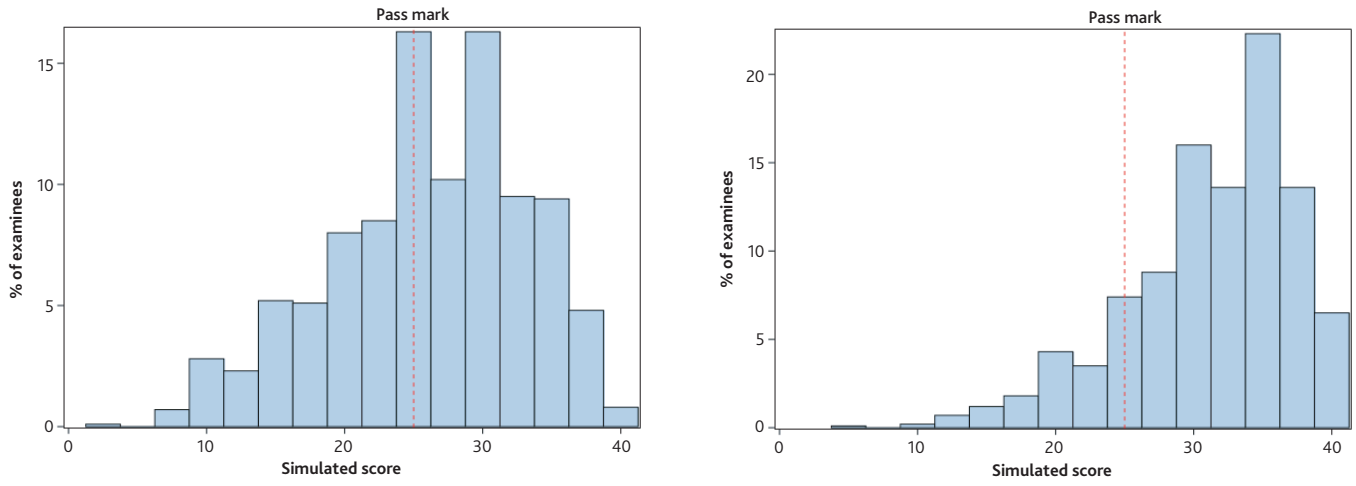

Figure 3: Simulated score distributions on one test for cohort with mean ability around the pass mark (left) and with mean ability above the pass mark (right)

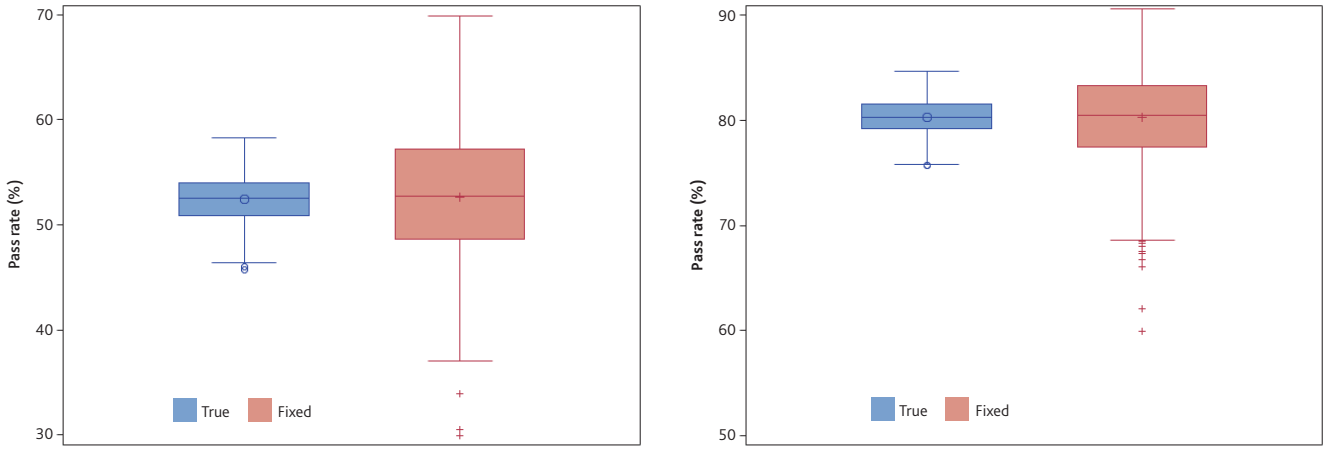

Figure 4: Distribution of pass rate using the true or fixed pass mark for cohorts with mean ability around the pass mark (left) and with mean ability above the pass mark (right)

One possible rationale for using fixed pass marks would be to conceive of the knowledge domain in each subject as a finite set of questions that could possibly be asked. We want to infer that the proportion of the domain known by examinees is above a certain value (e.g., 70%). If a test is constructed by stratified random sampling from the domain, then the proportion they get right is an unbiased estimate of the proportion of the domain that they know. The pass mark on the test could be set at the same percentage as the target domain percentage (i.e., 70%) or it could be adjusted to allow for the cost of making a false positive or false negative error (e.g., if it were deemed more costly to fail someone who knew more than 70% of the domain than it would be to pass someone who knew less than 70%).

The main challenge to this idea would be that individual tests would still differ in difficulty and it would be unfair to examinees not to try to allow for this somehow (as currently happens in GCSEs and A levels). This is of course a good point, but there are some possible responses. Firstly, we could argue that factual knowledge does not fit the concept of a 'latent trait' in the way that, for example, mathematical ability does. That is, there is arguably no real concept of a hierarchy of item difficulty that could define a meaningful continuum of progression. That is not to say that some items will not be answered correctly by more people than other items, but that the factors that make particular items of knowledge 'easy' or 'difficult' to recall will be idiosyncratic to the particulars of the learning experience and interests of different individuals. Tests of factual knowledge are therefore, in a sense, by definition equally difficult.

Secondly, when numbers of examinees are low, attempting to equate tests by statistical methods (e.g., comparable outcomes) can introduce more random error than it removes systematic error. In GCSEs and A levels, the grade boundaries on examination components are often unchanged when very few examinees have taken the component.

Thirdly, in an on-demand testing context (e.g., a test which is computer-delivered and auto-marked) when tests are constructed from a bank such that different individuals take different tests, statistical definitions of equivalent scores based on the performance of large groups could be less relevant. A given individual might have a better chance of passing on Test A than Test B, even if in a large group more would pass B than A.

Fourthly, being a victim of bad luck is not quite the same as being a victim of unfairness. If an individual happens to receive a selection of items that they do badly on when they would have done better on other possible selections of items, this is bad luck for them. The potential for unfairness perhaps resides more in how costly (in terms of time, money, and missed opportunities) it is for the individual to re-take the test.

Finally, this research has shown that there are steps that can be taken to reduce the amount by which different tests fluctuate in difficulty – such as trying to reduce the range of item difficulty, and making use in the test construction process of any information we have in advance about item difficulty, such as expert judgements. In testing contexts where the reuse of items is permitted, accurate empirical data will over time replace the more fallible expert judgements and allow test forms of equivalent difficulty, and hence the same pass marks, to be created with increasing precision.

Returning to the question posed in the title of this article, people will differ in the weight they give to different considerations when reaching a judgement. In my opinion, for on-demand tests that mainly require recall of facts in well-defined domains, with groups of test takers that vary in size and demographic composition, the advantages slightly outweigh the disadvantages.

References

Benton, T. (2016). *Comparable Outcomes: Scourge or Scapegoat?* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/346267-comparable-outcomes-scourge-or-scapegoat-.pdf

Bramley, T. (2012, February). *What if the grade boundaries on all A level examinations were set at a fixed proportion of the total mark?* Paper presented at the Maintaining Examination Standards seminar, London. Retrieved from http://cambridgeassessment.org.uk/Images/459357-what-if-the-grade-boundaries-on-all-a-level-examinations-were-set-at-a-fixed-proportion-of-the-total-mark-.pdf

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

Cizek, G. J. (Ed.) (2012). *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd Ed.). New York and London: Routledge.

Ofqual (2011). *Maintaining standards in GCSEs and A levels in summer 2011.* Retrieved from http://webarchive.nationalarchives.gov.uk/20110718105952/http://www.ofqual.gov.uk/files/2011-05-16-maintaining-standards-gcses-and-alevels-summer-2011.pdf

Wright, B. D., & Stone, M. (1979). *Best Test Design*. Chicago, Illinois: MESA Press.

Insights into teacher moderation of marks on high-stakes non-examined assessments

Victoria Crisp Research Division

Introduction

Where teachers assess their students' work for high-stakes purposes, their judgements are standardised through professional discussions with their colleagues – a process often known as *internal moderation*. This process is important to the reliability of results as any inconsistencies in the marking standards applied by different teachers within a school department can be problematic.

This research explored internal moderation practice for school-based work contributing to high-stakes assessments in England, Wales and Northern Ireland, with a focus on General Certificate of Secondary Education (GCSEs). Since their introduction in the 1980s, GCSEs in some subjects have involved a component of work that students conduct in the classroom rather than in the exam room. The nature of this work and the restrictions around it (e.g., how much time is allowed, whether it can be partly conducted at home) have varied and the number of subjects with a non-examined element has reduced over time. Broadly speaking, the work tends to involve some kind of project or extended piece of work that could not, realistically, be conducted within the time and limitations of an exam situation. Generally, such work is marked by the students' teachers and externally moderated by examiners appointed and trained by the awarding organisation (AO). The procedure for *external moderation* involves the school submitting a list of their students' marks and the 'moderator' selecting, in line with the AO's guidance, a relatively small number of students' work to review. Their selection ensures that students across the ability range are sampled. The moderator reviews the teachers' marking of the student work and determines whether the marking at the school is in line with national standards or requires adjustment (see Crisp, 2017, for insights into the processes involved in external moderation).

Where there is more than one teacher marking the student work in a particular school for a particular qualification and subject, there is a requirement for marks to be internally moderated before submission to the AO (i.e., before external moderation is conducted). Internal moderation can involve one teacher (e.g., the head of department) evaluating the marking of the other teachers against their own, and adjusting marks if needed, or a group meeting of the teachers within a school department where they compare and discuss how they would each mark the same example pieces of student coursework, agree on the marks for these, and subsequently adjust the marks given to other students if needed. The aim is for a school's set of coursework marks across all classes for a particular subject to have been marked to the same standard such that students are placed in an accurate rank order in respect of the specified criteria. Some General Certificate of Education Advanced levels (GCE A levels) also involve a non-examination element and tend to be assessed following a similar model to that for GCSEs.

Where internal moderation is a group activity, the process could be similar to that of consensus moderation described by Sadler in the context of higher education:

> *Consensus moderation starts with a sample of student responses drawn from the course pool. Working independently, all assessors mark all responses in the sample. For each, they record their provisional judgement and their reasons for it. Markers then convene as a group, individually present their decisions and rationales, and deliberate them until consensus is reached.*
>
> (Sadler, 2013, pp.7–8)

The approach involves comparison and alignment of judgements of student work against stated criteria, leading to clarification of interpretations of criteria and the development of shared understandings.

A number of studies have usefully explored the judgements involved when teachers assess student work. For example, Cooksey, Freebody and Wyatt-Smith's (2007) detailed coding and analysis of the influences on teachers' assessments revealed the complexity of their judgements and the varying strategies of different teachers even when applying national benchmark criteria. The knowledge and skills that teachers need have also been discussed. Drawing on Sadler (1998), Klenowski and Wyatt-Smith (2013) proposed that teachers making judgements about student work need to be able to utilise: knowledge of the content to be assessed; deep knowledge of the assessment criteria; and evaluative skills developed from previous experience of judging student work on similar tasks. A newly qualified teacher should already have the first of these, but the other two elements need to be developed through experience. Teachers will be provided with a set of written marking criteria to use but these could be subject to differing interpretations if applied by isolated individuals (Johnson, 2013). The provision of written exemplars may help with understanding the intended meaning of criteria (Sadler, 1998) but involvement in consensus moderation is also likely to be valuable to understanding criteria (Johnson, 2013), and is recognised to be a useful professional learning opportunity for teachers, building their assessment capacity (Harlen, 2005; Klenowski & Wyatt-Smith, 2010; Smaill, 2012). Common interpretations of criteria are developed through discussion of evidence depicting the qualities represented in the criteria (Klenowski & Wyatt-Smith, 2013). In the context of school-based assessment in Queensland, Australia, Klenowski and Wyatt-Smith (2010) argued that moderation meetings provide:

> *… an opportunity to generate new knowledge and new ways of knowing as teachers draw on their individual tacit and individual explicit knowledge and the group's tacit and explicit knowledge, and use this knowledge as a tool of knowing within a situated interaction with the social and physical world*
>
> (Klenowski & Wyatt-Smith, 2010, p.121)

Various examples are given, such as being able to check that similar skills are taught and assessed, increased confidence in the understanding of achievement expected at particular levels, and a shift from individual practice to shared practice and improvement (Klenowski & Wyatt-Smith, 2010). Klenowski and Wyatt-Smith (2013) related this process to Cook and Brown's (1999) notion of 'bridging epistemologies' in which individuals' tacit and explicit knowledge (that would bear on the individual's judgements) are revealed, and ways of knowing are generated through the group process of working together to articulate their understandings of criteria and develop a shared perspective. This is not dissimilar to Wenger's (1998) theory of communities of practice and shared understandings and practices developing through participation.

The role of collaboration amongst teachers has been explored by Allal and Mottier Lopez (2014). They drew on Cobb, Gravemeijer, Yackel, McClain, and Whitenack's (1997) notion that human judgement involves a reflexive relationship between an individual's psychological processes and shared social practices. Within this view, it is argued that meaning is not identical in the minds of all those involved but that interactions between participants allow 'taken-as-shared' meaning to emerge which guides activity (Allal & Mottier Lopez, 2014). Evidence suggests that collaborative assessment activities can facilitate this process of 'deprivatisation' and the construction of shared practices (Allal & Mottier Lopez, 2014; Black, Harrison, Hodgen, Marshall, & Serret, 2011). Such theories and evidence suggest that it is likely to be important that teachers have opportunities to be involved in collaborative assessment activities.

Perhaps the most directly-relevant research to the current study is Wyatt-Smith, Klenowski and Gunn's (2010) analysis of recorded teacher talk during consensus moderation meetings of teachers in the Queensland context. Their research identified that teachers move back and forth between:

> *(1) supplied textual artefacts, including stated standards and samples of student responses;*
>
> *(2) tacit knowledge of different types, drawing into the moderation; and*
>
> *(3) social processes of dialogue and negotiation.*
>
> (Wyatt-Smith, et al., 2010, p.59)

They concluded that the written assessment criteria are 'insufficient to account for how the teachers ascribe value and award a grade to student work in moderation' (p.59), and emphasised the social and cognitive elements of moderation practice. A tension was found between criteria that teachers carry 'in their head', developed through experience, and the stated criteria. The former was influential in judging ability but essentially unstated, though assumed to be common with those held by others. Wyatt-Smith et al. (2010) concluded that this tension is not necessarily a sign that teacher judgement is flawed or biased, but that assessment judgement involves a number of challenges.

Internal moderation procedures aim to ensure the consistency with which marking standards are applied within a school, both in terms of the reliability of teacher judgements and standards over time (Klenowski & Wyatt-Smith, 2013). In the context of non-examination elements of GCSEs and A levels, without consistency of marking within a school in terms of establishing an appropriate rank order, external moderation procedures would be difficult to implement appropriately. Aiding teacher development and improving the accuracy of future marking are likely to be additional aims of internal moderation. This study sought to improve our understanding of internal moderation practice in the context of GCSEs and A levels.

Method

This research involved the use of three complementary methods: semi-structured interviews; mock internal moderation sessions; and a questionnaire survey.

The interviews and moderation sessions were conducted with GCSE teachers with experience of internal moderation of coursework for English/English Literature, Geography, or Information and Communications Technology (ICT). These subject areas were selected to represent a variety of types of student work. The marking criteria for GCSE coursework in each of these subjects was *levels-based* with the mark range divided into a number of 'levels' or 'bands'. Each band related to a particular range of marks and had an associated description of the criteria that were expected to be met at that level.

Semi-structured interviews were conducted with 11 GCSE teachers. The participants were one English/English Literature teacher, five Geography teachers and five ICT teachers. The interview questions asked participants to describe how internal moderation is conducted and the thought processes involved, including how marking guidance is used and whether they feel the process works well.

Four mock internal moderation sessions were observed with some of the same participants: one session involving GCSE English/English Literature (one teacher); one session involving GCSE ICT (two teachers); and two sessions involving GCSE Geography (three teachers and two teachers respectively). At the English teacher's school, internal moderation of coursework was usually carried out by the head of department so, in order to mimic this, two of his colleagues also conducted some marking and these marks (along with some of his own marking) were moderated by the head of department as an individual activity. For Geography and ICT, the internal moderation was carried out as a group activity, mimicking usual practice in these school departments.

The sessions used student coursework provided by the researcher with each teacher marking four different students' projects before the internal moderation session. The students who prepared the coursework were unknown to the teachers, representing a departure from the usual situation where teachers mark work from their own students. However, during internal moderation teachers usually evaluate work from some students that they teach and some who are taught by colleagues, so it is not unrealistic to ask teachers to mark work from students they do not teach. Where mock internal moderation was an individual exercise (English) the time available allowed all 12 coursework folders to be considered. In all other cases (i.e., all those where the internal moderation exercise was conducted as a group activity), six coursework projects were considered in each case. (Note that this is more a reflection that the English moderation session was carried out after the school day when the participant had more time available, than an indication that individual moderation is faster or more efficient.)

All sessions were observed by the researcher and audio-recorded. For the individual session with the GCSE English participant, he was asked to 'think aloud' whilst conducting the task. He was instructed as follows, based on Ericsson and Simon (1993): 'I would like you to say out

loud everything that you would normally think to yourself silently whilst you are moderating. It may help if you imagine that you are in the room by yourself.' There is some debate around whether the 'think aloud' method can affect a participant's thinking whilst conducting a task (e.g., slowing down normal processes), however, it is generally felt to be a useful method providing more information than observation alone (for further discussion see Crisp, 2008; Ericsson & Simon, 1993; Green, 1998; Kobrin & Young, 2003; Nisbett & Wilson, 1977).

The questionnaire data reported in this article comes from a longer questionnaire that addressed teacher marking as well as internal moderation (Crisp, 2013) and which was completed by 378 secondary school teachers from a range of subject areas across the Arts, Sciences, Humanities, Technology, English, Business and Social Sciences. Only teachers with experience of internal moderation were asked to complete the questions relevant to the current study, thus the numbers of respondents for the relevant questions were lower, ranging from 261 to 282 (with a total of 288 answering at least 1 of the questions relating to internal moderation). Of the 288 responding teachers, 158 taught GCSE (but not A level), 54 taught A level (but not GCSE) and 68 taught both[1]. The relevant questions covered use of marking criteria in internal moderation, differences between internal moderation and marking, and any effects of social interactions and group dynamics on the process.

It should be noted that there are some limitations to this research. Firstly, the number of teachers involved in the interviews and internal moderation was fairly small. This was necessary due to the in-depth nature of analysis needed, but it is possible that variations in practice might have been seen if different teachers had participated. Secondly, the 'mock' nature of the internal moderation sessions could be criticised on the grounds of not being as authentic as asking teachers to evaluate the work of their own students. Work from students unknown to the teachers was used to avoid any risk of the research affecting the 'live' marking and internal moderation process for the schools' own students. The use of work from students not known to the teachers could mean that some specific issues around assessing their own students are missing from the current data. However, as mentioned earlier, during internal moderation teachers usually look at work from some students that they have not taught, as well as some from students that they have taught, so it is hoped that using students unknown to the teachers is not a significant weakness to the method.

Findings

Insights from the interviews

During interviews, teachers were asked about the process of internal moderation at their school. The English teacher described the individual approach to internal moderation at his school and how he collects up all marked coursework for the subject and then checks a sample of each teacher's marking. All other teachers interviewed described their use of *group moderation* with each teacher evaluating some examples of student work from other classes (e.g., a high-, middle-, and low-scoring example from each class might be selected for consideration and marked by the other teachers). This marking might be conducted before or at an internal moderation meeting where marks would then be compared between teachers, discrepancies discussed, justifications given and agreements reached. The internal moderation could result in one or more individual teachers returning to their marking for all coursework projects and adjusting their marks to bring them into line with the marking standards being applied by their colleagues. Most teachers felt that internal moderation worked well, and several quoted as evidence of this that their marking standards are usually similar (with only small mark differences found if any) and that their marks had rarely been adjusted by the external moderation process.

Teachers were also asked about any differences in how they evaluate work and in the use of criteria in internal moderation compared to marking. Mostly, the evaluation process was thought to be very similar between these two contexts. Some participants commented that during internal moderation each individual coursework project was considered more quickly, particularly if the second marking was conducted in the internal moderation meeting rather than in advance of it. In terms of use of the marking criteria, this was generally felt to be similar but a few teachers suggested that they made less direct use of the detail of the marking criteria when evaluating during internal moderation, as a broader view is taken. Some teachers mentioned that, when marking as part of internal moderation, they tended to be slightly harsher on students that they did not know because they had not seen the work progressing, and that they tended to defend the marks they had given to their own students. Nonetheless, the internal moderation process was thought to address any possible biases towards or against known or unknown students through discussion and refinement of marking.

Insights from the mock moderation sessions

The teacher participant who conducted the *mock moderation* as an individual activity considered each coursework folder in turn and usually orientated himself to the topic when starting to read. This was often followed by noting the mark(s) originally given to the work and what this may suggest (e.g., "a band 3 essay, this will probably not be as good as the previous piece"). Reading during internal moderation appeared to involve some skimming with any annotations (ticks, comments, etc.) somewhat guiding this process. An absence of teacher annotations, or only brief annotation, was sometimes commented on by the moderating teacher. Agreement with teacher marks or annotations was sometimes noted. In addition, the participant sometimes noted that work had been over- or under-valued, which then led to adjusting marks.

In the group moderation meetings, the teachers compared and discussed the marks, considering each coursework project in turn. For each project, they began by each stating the total mark that they had given. If the total marks were close together then little discussion was required but the grade that was likely to be equivalent to that mark might be noted. For a project with slightly larger differences between the total marks given by different teachers, there was a much lengthier discussion. One tendency was for the teacher who was furthest from the others to immediately consider their own marking to have been too lenient or too harsh.

At one school, the internal moderation meeting began by comparing the teachers' rank orders of total marks for the coursework projects. Any significant differences in rank orders were noted. The discussions around this process involved each teacher stating the mark they gave, comparing the mark to those proposed by their colleagues, noting similarities and differences and possible adjustments to marks.

1. A small number taught another qualification (e.g., BTEC Entry level) either as the only qualification they taught or alongside GCSE and/or A level.

There was a significant comparative element to the discussions in this school in terms of the teachers comparing the quality of one coursework project to another, often in terms of specific marking criteria. After identifying those coursework projects where there were discrepancies in the marks given by different teachers, the projects in question were discussed in more detail.

Discussion during the mock internal moderation meetings involved going back to evidence within the coursework projects, using the marking criteria (or a marking cover sheet attached to each project which lists the marking criteria), summarising the contents and features and quality of the work. Evaluations were usually stated at a fairly broad level (e.g., evaluation of data representation) but sometimes at a more specific level (e.g., evaluation of map use). Criteria with which there were discrepancies for a particular student were identified which led to discussion of different perspectives on that particular aspect of the student's work. Discussions sometimes focused on whether a particular part of a student's work constituted evidence towards a particular criterion. One teacher would show the other(s) the evidence of a particular criterion that they had accepted, and then the teachers would reach mutual agreement on whether to accept this as sufficient evidence. The more extreme-marking teacher might question their reasons for their mark and/or describe why they gave that mark and the other teachers would describe their rationales for the marks they gave. There were also discussions about the requirements of the marking criteria to clarify and confirm interpretations of these. This process led to agreement on the appropriate mark using the marking criteria. Usually marks were adjusted away from the more extreme mark and towards the consensus. The grade likely to relate to the mark was sometimes noted once the mark had been agreed or during discussions.

In two of the three mock group moderation meetings in this research, observing teachers' interactions suggested that the more senior teacher present tended to lead the direction of the discussion and appeared to be less likely to adjust the marks they gave, although they were not unwilling to listen and reconsider their initial mark. It is plausible that a more senior teacher has the most experience with marking and that their judgements are likely to be closest to the national standards. In which case, it would be appropriate that they have a stronger influence on the discussions and decisions. However, if their understanding of the marking criteria and expected standards is no stronger than that of their colleagues, then their more influential position could have an unhelpful effect on decision-making.

Insights from the questionnaire responses

As described earlier, the questionnaire data reported here comes from a longer questionnaire that addressed coursework marking as well as internal moderation (Crisp, 2013). Those teachers without involvement in internal moderation were asked to skip the relevant questions. Some 25 to 31 per cent of teachers omitted the closed questions in this section. This suggests that these teachers work in departments where they are the only teacher (perhaps due to small school size or limited uptake of the subject) or where one teacher, perhaps the head of department, conducts the internal moderation alone.

The questionnaire included three closed response questions on internal moderation with an open response question following each to elicit further detail. The closed response questions and the frequency of different responses to these are shown in Table 1.

Table 1: Closed response questions and frequencies of response

During internal moderation procedures, do you use the mark scheme criteria in exactly the same way as when marking? (N=282, omitted by 25.4% of whole sample)

Yes	No
96.8%	3.2%

How often do social interactions or group dynamics between teachers affect internal moderation procedures? (N=281, omitted by 25.7% of whole sample)

Never	Occasionally	Sometimes	Usually	Always
47.3%	27.8%	18.5%	4.6%	1.8%

Are there any other ways in which internal moderation judgements differ from marking judgements? (N=261, omitted by 31.0% of whole sample)

Yes	No
22.2%	77.8%

Firstly, teachers were asked whether marking criteria are used in exactly the same way in internal moderation as during marking. The majority of those responding felt this was the case. Respondents were asked to give examples if they felt there were differences. The responses are listed in Table 2. Comments included that internal moderation involves considering work more holistically, ranking work into order, using the criteria to justify decisions to others, and that teachers' annotations are used as well as the marking criteria during moderation.

Table 2: Reported examples of ways in which marking criteria are used differently in internal moderation compared to marking

- *Use the detailed breakdown, then look at how it is marked after.*
- *Rank the grades. Look again in coursework if think too low/too high.*
- *Also look at comments and cross-referencing.*
- *We need to compare decisions and justify them so I refer to it much more.*
- *Because we moderate our interpretations of what answers mean.*
- *Generally look at work as a whole.*
- *Use expertise of other teachers involved in marking/moderation.*
- *Sometimes we refer to teaching resources for the staff to help further.*
- *When marking I mark by question. When moderating I also mark overall.*
- *One member of staff is an examiner for an awarding body so she sometimes has additional information which can clarify the mark scheme.*
- *Don't focus on them.*
- *Using marking criteria as guidance.*
- *Take an overview; look at the annotations of the teacher to check where marks have been awarded.*

Secondly, teachers were asked about social interactions or group dynamics and how frequently these affect internal moderation procedures. Most respondents reported that these were infrequent influences on moderation. However, nearly a quarter reported that social interactions or group dynamics at least 'sometimes' affected procedures. Teachers were asked to provide an example, if possible. Fifty-seven responded with at least one point and their comments were analysed by grouping similar responses together (see Table 3). Some responses were

Table 3: Reported examples of social influences on internal moderation

Response	*Frequency*
Positive discussion to reach agreement/happy medium over differences in views/marks	6
Good relationship with other staff/positive working relationships	2
Constant ongoing dialogue with other staff during coursework teaching	2
Hierarchy in department	2
Level of experience or familiarity with qualification (e.g., inexperienced teachers led by more experienced, experienced teachers' marks get agreed more quickly)	5
Differences in experiences (e.g., different subject experiences)	4
Personality (e.g., persuasive, wilful, argumentative, emotional)	5
Collaborative working issues – need for give and take in team working	1
Taking offence/taking criticism badly	5
Personality clashes/personal differences	2
Risk of unprofessional behaviour	1
Taking over from another teacher who has not taught the group well	1
Differences in perceptions of student performance/differences in marks/differences in ideas about standards	10
Interpretation of the criteria (e.g., helps to hear how someone else interpreted the criteria)	4
Differences in thoroughness	1
Occasional bias against a pupil can be removed in internal moderation	2
Can consider the nature of the student group (e.g., if less able)	1
Social interactions affect time and support available	1
Practical issues regarding time (e.g., time-consuming perhaps because of arguments or getting off track; organising a time to suit everyone)	6
Total	**61**

positive, implying that working together was a useful and supportive part of the process. For example, six teachers commented that positive discussion was used to reach agreement over differences in views. Frequent ongoing dialogue with other teachers during coursework-related teaching was also mentioned as a positive feature. Several neutral comments about other teachers were made. These included that the level of experience of staff, differences in experiences such as different subject experiences and the hierarchy in a department could affect the internal moderation process (e.g., teachers with less experience may be led by teachers with more). Five teachers mentioned aspects of personality, such as persuasiveness or argumentativeness, as influences on internal moderation. Other comments included differences in perceptions of student performance or marks, and differences in interpretations of criteria. However, it was unclear from these comments how teachers felt social interactions in relation to these differences influenced the process. Several negative influences were mentioned, including issues about colleagues taking offence at criticism, and personality clashes. Student-focused comments included that the nature of a student can be taken into account through discussion, and that occasional bias against an individual student can be resolved. Two practical considerations were also noted: that social relationships affect the amount of time and support available to the teacher in relation to their marking; and that the process can be time-consuming due to arguments or getting 'off-track' during meetings.

The third closed question asked teachers if there were any other ways in which internal moderation judgements differ from marking judgements. Over three quarters of those who responded reported that there were not, suggesting that many teachers consider judgements in internal moderation similar to those in marking. Those that felt there were differences were asked to give an example, to which 38 teachers gave a response. Comments included: that different interpretations of the marking criteria influence the internal moderation process; that a view from a teacher who is less familiar with the student can aid objectivity; that internal moderation decisions involve discussion; and that one teacher may see qualities in the work that another did not identify.

Discussion

This study provides insights into the internal moderation processes used in schools to standardise marks before submission for external moderation. The mock internal moderation sessions showed that, as well as behaviours relating to the consideration of individual coursework projects (and thus common with marking), a number of additional behaviours occur, including noting and/or agreeing with the mark given or comments made, discussion of where evidence in the work meets particular marking criteria, discussion of requirements, and adjustment of marks. Interview comments suggested that internal moderation is felt to address any biases towards or against known or unknown students. In the questionnaire responses, teachers generally reported that internal moderation uses marking criteria in the same way as marking, though student work may be considered more holistically in the former. This may imply that the criteria provide not only the basis for judgements about marks in internal moderation processes, but also a common terminology that can be used by teachers in internal moderation discussions.

Given the levels-based nature of the mark schemes used to assess most non-examination work contributing to GCSEs and A levels, it is logical to expect that teachers look for evidence in the student work relating to particular skills, attempt to identify the most appropriate level by looking at the criteria described for each level, and then judge the mark to be awarded from the range relating to the level and how well the work has met the criteria. The current data would generally seem to be consistent with this, though perhaps there is insufficient data to claim this conclusively.

Previous research on grading meetings has shown that group dynamics influence the judgements of examining teams (Murphy et al., 1995). In the current context, teachers tended to report that social interactions and group dynamics were an infrequent influence on internal moderation. Whilst group dynamics were not felt to be a strong influence, there is clearly a social dimension to internal moderation. This is perhaps exemplified in the tendency for a teacher whose initial mark for a project was furthest from the other teachers' marks to immediately express that they were likely to be the one whose marking was out of line with national standards. Whilst this could be a logical

assumption in the circumstances, there is potentially a social element to this with confessing their own (possible) error before they are criticised by others acting as a device for 'saving face'. It would also seem to be a positive sign about the working relationships of the teams involved that participants were comfortable admitting a potential error to their colleagues. This relates to the notion of 'team psychological safety', which is defined as a shared belief amongst team members that the team is a safe context for interpersonal risk-taking (Edmondson, 1999). Team psychological safety facilitates behaviours such as admission and discussion of errors, and seeking information and feedback, and is associated with team learning.

As Adie, Lloyd and Beutel (2013) point out, the aim of a moderation processes is to provide 'a way to develop a shared understanding of standards of achievement and the qualities that will denote evidence of these standards' (p.971). Elements of this can be seen in the findings of the current study. Work by van der Schaaf, Baartman and Prins (2012) on moderation in a university context in the Netherlands analysed the quality of argumentation when tutors evaluated student portfolios. They found judgements to be of low-quality with many articulations not relating to relevant evidence. In contrast, the current study in the context of GCSE suggests considerable focus during internal moderation on relevant evidence in student work with frequent discussion of the location of relevant evidence and whether this evidence is sufficient to meet a particular criterion. This, along with reference to the marking criteria, discussion of requirements and the meaning of the marking criteria, is consistent with Cook and Brown's (1999) notion of tacit knowledge being made explicit and helping to refine and create new ways of knowing (a notion previously applied to consensus moderation by Klenowski and Wyatt-Smith, 2013). The new 'ways of knowing' created by involvement in an internal moderation meeting (and perhaps also to some extent from feedback on internal moderation in cases where it is conducted individually by a senior member of a department) should inform the remaining and future marking of each individual teacher in terms of understanding marking standards and how aspects of student work provide evidence of elements of the marking criteria.

Previous research has suggested that internal moderation is a useful professional development experience for teachers (e.g., Harlen, 2005; Smaill, 2012; Klenowski & Wyatt-Smith, 2010). The omission rate on the internal moderation sections of the questionnaire suggests that around a quarter of teachers who mark non-examined work may not get this experience, presumably due to being the only teacher in the school for a particular subject, or because one teacher conducts the internal moderation alone. This is an interesting finding in itself as, either through circumstance or design, these teachers are missing out on a potentially useful professional development experience.

Some of the potential challenges to teacher assessment may be at least partly mitigated by internal moderation. Purported challenges include that written criteria are subject to interpretation (e.g., Sadler, 1998), that teachers use tacit knowledge as well as the written criteria, and that they may assume that their own tacit knowledge is the same as that of others (Klenowski & Wyatt-Smith, 2013). Discussion of the meaning of criteria and of examples of how this is evidenced in student work would seem likely to reduce these problems.

Another outstanding question is whether both individual and group internal moderation approaches are equally effective. Group moderation would seem to have the benefit of discussion, of jointly refining understandings, and greater potential for continuing professional development. However, if one experienced teacher has a good 'feel' for the standards expected and a good understanding of the criteria, it could be easier and/or more efficient to obtain a coherent rank order for all students taking a particular subject at a particular school through one teacher working alone to moderate the work. This might provide weaker development for the other teachers but, arguably, achieving accurate results for the current cohort of students is a more immediate aim of internal moderation than providing professional development that may aid future practice. Further research could usefully explore the relative success of group and individual approaches to internal moderation in terms of whether a school's marks provide an appropriate rank order, whether a school's marks are adjusted, and whether individual teachers become more aligned with national standards over time through their experiences or feedback from moderation.

The evidence gathered in this research does not suggest any significant problems with the nature of the internal moderation processes used in schools in relation to non-examined GCSE and A level work. Attention is paid to relevant evidence in student work, moderation is reported to be infrequently influenced by group dynamics, the process is thought to act to remove any potential personal bias, and teachers tend to report that the process works well.

References

Adie, L., Lloyd, M., & Beutel, D. (2013). Identifying discourses of moderation in higher education. *Assessment & Evaluation in Higher Education, 38*(8), 968–977.

Allal, L., & Mottier Lopez, L. (2014). Teachers' professional judgment in the context of collaborative assessment practice. In C. Wyatt-Smith, V. Klenowski & P. Colbert (Eds.), *Designing assessment for quality learning* (pp.151–165). Dordrecht, The Netherlands: Springer.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*(4), 451–469.

Cobb, P., Gravemeijer, K., Yackel, E., McClain, K., & Whitenack, J. (1997). Mathematizing and symbolizing: The emergence of chains of signification in one first-grade classroom. In D. Kirshner & J. A. Whitson (Eds.), *Situated cognition: Social, semiotic, and psychological perspectives* (pp.151–233). Mahwah, NJ: Lawrence Erlbaum.

Cook, S. D. N., & Brown, J. S. (1999). Bridging epistemologies: The generative dance between organizational knowledge and organizational knowing. *Organization Science, 10*(4), 381–400.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation, 13*(5), 401–434.

Crisp, V. (2008). The validity of using verbal protocol analysis to investigate the processes involved in examination marking. *Research in Education, 79*(1), 1–12.

Crisp, V. (2013). Criteria, comparison and past experiences: How do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice, 20*(1), 127–144.

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

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly, 44*(2), 350–383.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* London: MIT Press.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge, UK: Cambridge University Press.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245–270.

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education, 28*(1), 91–105.

Klenowski, V., & Wyatt-Smith, C. M. (2010). Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform. *Assessment Matters, 2*, 107–31.

Klenowski, V., & Wyatt-Smith, C. M. (2013). *Assessment for education: Standards, judgement and moderation*. London: Sage.

Kobrin, J. L., & Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. *Applied Measurement in Education, 16*(2), 115–140.

Murphy, R., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J., & Gower, R. (1995). *The dynamics of GCSE awarding: Report to the School Curriculum and Assessment Authority*. Nottingham: University of Nottingham.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*(1), 77–84.

Sadler, D. R. (2013). Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy & Practice, 20*(1), 5–19.

Smaill, E. (2013). Moderating New Zealand's National Standards: teacher learning and assessment outcomes. *Assessment in Education: Principles, Policy & Practice, 20*(3), 250–265.

van der Schaaf, M., Baartman, L., & Prins, F. (2012). Exploring the role of assessment criteria during teachers' collaborative judgement processes of students' portfolios. *Assessment and Evaluation in Higher Education, 37*(7), 847–860.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. J. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, Policy & Practice, 17*(1), 59–75.

Which students benefit from retaking Mathematics and English GCSEs post-16?

Carmen Vidal Rodeiro Research Division

Introduction

Following the Wolf Report (Wolf, 2011), the UK Government legislated that from September 2013 all young people who did not achieve a grade C in Mathematics and English General Certificate of Secondary Education (GCSEs) had to continue studying these subjects post-16. Therefore, since 2014, students failing this requirement have continued to work towards achieving these qualifications or an approved interim qualification as a 'stepping stone' towards a GCSE. For some students, reaching the GCSE standard may potentially have required progressive stepping stones, for example, through Functional Skills qualifications, or through Foundation and Higher Free Standing Mathematics Qualifications.

According to a report published by the Policy Exchange in summer 2014 (Porter, 2015), 27% of the cohort who took GCSE English did not achieve a grade C or above (just over 125,000 students) and 31% of the cohort who took GCSE Mathematics did not achieve a grade C or above (just below 180,000 students). These students, who should have retaken English and Mathematics post-16, could also have been studying a variety of different courses. Some could have gone on to study academic courses, such as General Certificate of Education Advanced Subsidiary/Advanced levels (GCE AS/A levels), some could have been following alternative courses at different levels, such as BTECs, Cambridge Nationals, Cambridge Technicals, or vocationally related qualifications, and some might not have taken any other qualification.

Changes to the funding policy for 16–19 students in state-funded schools and colleges (for details, see https://www.gov.uk/guidance/16-to-19-funding-how-it-works) and the reform of post-16 accountability measures (DfE, 2017) are likely to have had an impact on enrolments in these centres and on entries for all types of qualifications in Key Stage 5 (KS5), but in particular for GCSEs in English and Mathematics. The 2015/16 academic year was the first in which it became a condition of colleges' funding that students who had previously achieved a grade D in English or Mathematics should retake the qualification. As a result, the overall number of entries among students aged 17 and over increased (Ofqual, 2016; 2017).

Recently, educational bodies across the sector, for example, The Office for Standards in Education, Children's Services and Skills (Ofsted), (Burke, 2016; Exley, 2016); the Association of Employment and Learning Providers (Martin, 2017); the Association of Colleges (Exley & Belgutay, 2017); the National Association of Head Teachers (NHAT, 2017); and the Learning and Work Institute (Belgutay, 2017) have been calling for a change in the resit policy. Their main reasons for requesting a review of the policy include:

- concerns over the lack of resources across the education system due to the increasing number of students required to retake the qualifications (e.g., insufficient funding; pressure on staff; logistical issues). This is a particular challenge for further education (FE) colleges, where the majority of the students retaking English and Mathematics GCSEs are enrolled;
- the huge numbers of learners aged 17 and older who failed to improve their grades after resitting GCSEs in English and/or Mathematics. In fact, the 2015/16 Ofsted Annual Report (Ofsted, 2016) stated that many students were still not getting at least a grade C by the age of 19;

- having to retake English and/or Mathematics GCSEs again and again until a grade C is achieved can be demotivating for many students and attendance to the lessons can become quite low; and
- for many students, an alternative qualification may be a more appropriate means of improving their English and Mathematics skills and ensuring that they are ready for work or further study. High-quality alternative curricula and qualifications (e.g., Functional Skills) for students aged 16–18 for whom GCSEs are not appropriate have been proposed by some of the educational bodies mentioned.

However, in April 2017, the Education and Skills Funding Agency (ESFA) confirmed that the condition of funding for post-16 institutions for 2017–18 would make resits compulsory for students who obtained a grade 3 or D in either English or Mathematics[1] (ESFA, 2017). Furthermore, the funding regulations stated that all 16 to 18-year-old students with a near pass (previously grade D, now grade 3) in these subjects must continue studying and then resit the GCSE, rather than take an alternative stepping stone qualification. For those students receiving grades lower than a D (and now a grade 3), the option of studying an alternative qualification is available.

The aim of this research was to contribute to the discussion on the English and Mathematics GCSEs resit policy by investigating the uptake of GCSEs in English and Mathematics in post-16 schools and colleges in England, and the types of students who are more likely to improve their grades as a result of resitting the qualifications. In particular, the following research questions were addressed:

1. How many KS5 students take GCSEs in English and/or Mathematics?
2. What grades did students have in their first GCSE attempt in these subjects?
3. Was the GCSE English and/or Mathematics grade obtained in the resit better than in the first attempt?
4. What types of students were more likely to improve their GCSE English and/or Mathematics grade if they resat the qualification in KS5?
5. Does taking GCSE English and/or GCSE Mathematics in KS5 have an effect on students' performance in Level 3 (A level and equivalent) qualifications?

Data and methodology

The KS5 extract from the 2016 National Pupil Database (NPD) was used in this research. The NPD is a database held by the Department for Education (DfE), consisting of results for all students in all qualifications/subjects in schools and colleges in England, as well as student characteristics such as age and gender. Data from the school census, which is primarily available for students from state-maintained schools, provided information on student characteristics such as ethnicity, special education needs, or level of deprivation.

The analyses carried out focused on 538,707 students who were 17 years-old at the start of the academic year 2015/16 and for whom there were records of qualifications, at any level, taken in 2015 or 2016 (when they were expected to be in Year 12 and 13, in the sixth form).

1. Note that June 2017 saw both the first cohort of students sit the reformed GCSEs (graded 9–1) and the final cohort take resits under the legacy version of the qualifications (graded A*-G).

It is important to note now that this research did not follow up students who did not achieve grade A*-C in GCSE English and/or Mathematics at the end of Key Stage 4 (KS4) and investigate their GCSE uptake in KS5. The Policy Exchange (Porter, 2015), Education Datalab (Allen, 2016) and the Department for Education (DfE, 2016) have produced reports looking at GCSE resits in English and/or Mathematics from that point of view.

The statistical methods used to answer the research questions varied from simple descriptive statistics to more robust and sophisticated analyses using propensity scores or regression techniques. In particular, statistical modelling was used to investigate:

- the types of students that were more likely to improve their GCSE English and/or Mathematics grades; and
- the effect of resitting GCSE English and/or Mathematics on students' performance in Level 3 qualifications.

For clarity, we explain the methods and describe their application in each specific context, together with their results, later in the article.

Results

Uptake of GCSE English and Mathematics by KS5 students

GCSE English

There were 72,995 students who sat GCSE English (English or English Language) during KS5. Of those, 8,382 (11.5%) did not have a record in the NPD of having sat the qualification during KS4 (sessions prior to November 2014). Note that the students considered in this research were in Year 13 in the academic year 2015/16 and that they were counted as taking GCSE in English whilst in KS5 if they sat the examination in the November 2014 session, or in any 2015 or 2016 session (November or June).

Some of the students considered in this research sat the GCSE only once during their KS5 years, but others had multiple attempts. Table 1, showing the distribution of the number of attempts in KS5, indicates that almost 70% of the students only sat the GCSE English once. However, a quarter of the students did so twice, and 7% three or four times.

Table 1: Distribution of the number of GCSE English attempts in KS5

No. attempts	*No. candidates*	*% candidates*
1	49,780	68.20
2	18,121	24.82
3	4,125	5.65
4	969	1.33

Figure 1 shows the grade achieved in the first attempt by the students who sat GCSE English in KS5 (note that this first attempt might have been at secondary schools during KS4 – Years 10 and 11). As expected, the majority of the candidates did not achieve grades A*-C in their first attempt at GCSE (only 9.4% of them did so overall). The group of candidates who had not sat the GCSE during KS4 achieved better grades; for example, 3% obtained a grade A* in their first attempt and 52% achieved grades A*-C. This contrasts with the percentages for the group that was resitting: only 3.8% and 0.6% achieved grades A*-C or A*-B respectively in their first attempt.

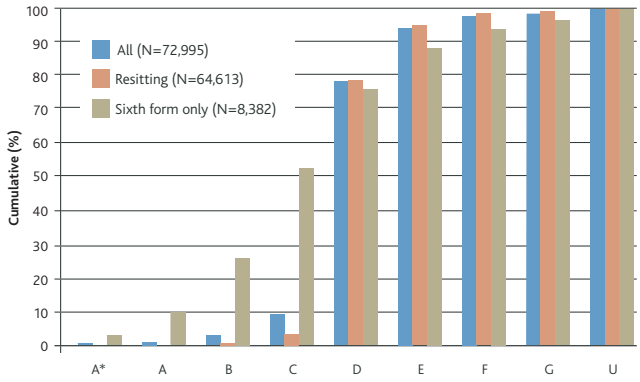

Figure 1: Grade distribution in GCSE English at first attempt

The percentage of candidates, amongst those who had resat the qualification in KS5, who improved their grade was calculated. If a candidate had sat the GCSE in English more than once in school, the best grade was considered as a baseline to calculate the improvement in KS5. More than half of the students (53%) did not improve their grade in GCSE English when they resat the qualification in KS5. Table 2 shows the changes by grade.

Table 2 shows that around 35% of the students with a grade C in GCSE English by the end of KS4 achieved the same grade during KS5, 28% improved their grade and achieved a grade B, and 17% performed worse and achieved a grade D. Overall, 65% of these pupils failed to improve their grade. Similarly, only 46% of the candidates with grade D in GCSE English by the end of KS4 (note that these candidates needed to continue studying English, as they did not achieved grades A*-C) improved their grade.

GCSE Mathematics

There were 67,759 students who sat GCSE Mathematics during KS5 (slightly lower than the number of students sitting GCSE English). Of those, 9,615 (14.2%) did not have a record in the NPD of having sat the qualification during KS4 (sessions prior to November 2014). As for GCSE English, the students considered in this research were in Year 13 in the academic year 2015/16, and they were counted as taking GCSE in Mathematics whilst in KS5 if they sat the examination in the November 2014 session or in any 2015 or 2016 session (November or June).

Table 3, showing the distribution of the number of GCSE Mathematics attempts in KS5, indicates that over 60% of the students only sat the qualification once. However, a quarter of the students did so twice, 9% three times, and just over 4% resat the qualification four times.

Table 3: Distribution of the number of GCSE Mathematics attempts in KS5

No. attempts	*No. candidates*	*% candidates*
1	42,579	62.84
2	16,605	24.51
3	5,828	8.60
4	2,747	4.05

Figure 2 shows the grade achieved in the first attempt by the students who sat GCSE Mathematics in KS5 (note that this first attempt might have been at secondary schools during KS4 – Years 10 and 11). As expected, the majority of the candidates did not achieve grade A*-C in their first attempt at GCSE (only 14% of them did so overall – this percentage is higher than in English though). The group of candidates who had not sat the GCSE during KS4 achieved better grades; for example, 8% obtained a grade A* in their first attempt and 60% achieved grade A*-C. This contrasts with the percentages for the group that was resitting: only 6.3% and 1.0% achieved grade A*-C or A*-B respectively in their first attempt.

The percentage of candidates, amongst those who had resat the qualification in KS5, who improved their grade was also calculated for GCSE Mathematics. As before, if a candidate had sat the GCSE in Mathematics more than once in school, the best grade was considered as a baseline to calculate the improvement in KS5. Almost 60% of the students did not improve their grade in GCSE Mathematics when they resat the qualification in KS5. Table 4 shows the changes by grade.

Table 4 shows that around 51% of the students with a grade C in GCSE Mathematics by the end of KS4 achieved the same grade during KS5, 31% improved their grade and achieved a grade B, and 13% performed worse and achieved a grade D. Overall, 66% of these students failed to improve their grade. Similarly, just over 40% of the

Table 2: Changes in GCSE English grade (best grade in KS4 vs. best grade in KS5)

Best grade in KS4	*Best grade in KS5*									*No. candidates*
	*A**	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*U*	
A*	16.67	33.33	0.00	16.67	16.67	16.67	0.00	0.00	0.00	6
A	35.00	31.67	6.67	15.00	5.00	5.00	0.00	0.00	1.67	60
B	6.19	31.86	36.58	12.98	5.31	0.88	1.18	2.06	2.95	339
C	1.48	6.17	27.54	34.67	17.20	4.17	1.56	4.04	3.17	2,302
D	0.03	0.34	3.95	41.80	27.35	13.19	4.42	3.53	5.39	52,158
E	0.04	0.24	2.08	20.29	29.90	25.45	9.93	4.86	7.21	7,879
F	0.07	0.74	1.19	7.67	18.39	30.68	20.63	8.27	12.36	1,343
G	0.00	0.31	3.09	16.36	14.81	18.21	16.98	13.89	16.36	324
U	0.50	1.49	2.48	14.36	21.29	17.82	12.87	6.44	22.77	202

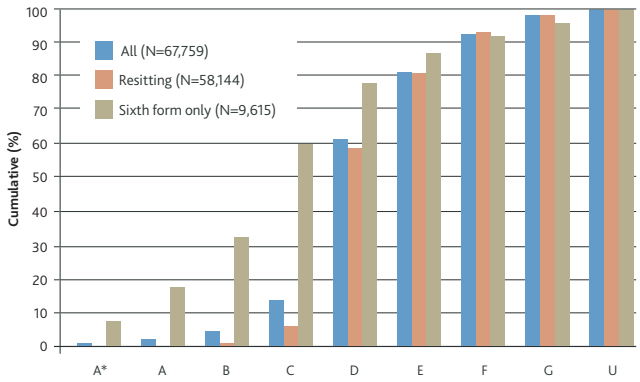

Figure 2: Grade distribution in GCSE Mathematics at first attempt

candidates with grade D in GCSE Mathematics by the end of KS4 (note that these candidates needed to continue studying Mathematics, as they did not achieved grades A*-C) improved their grade.

Which students were more likely to improve their GCSE grades in English and Mathematics?

An investigation into the types of students who were more likely to improve their GCSE English/Mathematics grades as a result of resitting during their KS5 years was carried out in this research. In particular, the following candidates' characteristics were looked at: gender, overall attainment at Level 2 (measured by the average KS4 points per entry[2]), type of centre attended, number of attempts in GCSE English/Mathematics during KS5, resitting GCSE Mathematics/English or not, and size of their Level 3 portfolio of qualifications.

Multilevel logistic regression modelling was used as an analytic framework to identify and control for the range of factors already mentioned. Logistic regression is a type of regression analysis that is used when the dependent variable or outcome is a dichotomous variable (i.e., it takes only two values, which usually represent the occurrence or non-occurrence of some event) and the independent variables are continuous, categorical, or both. It is used to predict the probability that the event of interest will occur as a function of the independent variables (see, e.g., Hosmer & Lemeshow, 2000). A multilevel model was proposed due to the hierarchical or clustered structure of the data (students grouped within centres). If we failed to recognise this hierarchical structure, then the standard errors of the regression coefficients would be underestimated, leading to an overstatement of the statistical significance. Detailed discussions of the implementation and outcomes of the multilevel logistic regression can be found in Goldstein (2011).

For the purpose of the analyses presented in this article, the dependent variable for the model was the improvement (or not) of the grade in GCSE English/Mathematics.

The models in this research take the following form:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 IV1_{ij} + \beta_2 IV2_{ij} + \beta_3 IV3_{ij} + \cdots + \beta_k IVk_{ij} + u_j$$

where p_{ij} is the probability of student i in centre j improving their GCSE grade by the end of KS5, $IV1$ to IVk are the independent variables, β_0 to β_k are the regression coefficients or fixed effects and u_j is a random variable at centre level which followed a normal distribution with mean zero.

A positive regression coefficient for an independent variable means that the variable increases the probability of the outcome, while a negative regression coefficient means that the variable decreases the probability of the outcome. The size of the coefficient gives an indication of the size of the effect that the variable is having on the probability of the outcome. In particular, a large regression coefficient means that the variable strongly influences the probability of the outcome; while a near-zero regression coefficient means that the variable has little influence on the probability of the outcome. However, it is important to keep in mind the scale of the independent variables when interpreting the regression coefficients (e.g., the variable *percentage of Level 3 qualifications* has a range between 0 and 100, whilst the variable *number of attempts in GCSE English* ranges from 1 to 4).

The results of the regression model for English are presented in Table 5 and the results for Mathematics are presented in Table 6. All the variables were statistically significant predictors of GCSE English and GCSE Mathematics grade improvement. In other words, each of the candidate characteristics displayed a statistically significant association (either positive or negative) with improving the GCSE grade in KS5. A discussion of these associations follows.

2. Here, per entry means per GCSE or equivalent entry.

Table 4: Changes in GCSE Mathematics grade (best grade in KS4 vs. best grade in KS5)

Best grade in KS4	*Best grade in KS5*									*No. candidates*
	*A**	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*U*	
A*	27.27	0.00	9.09	18.18	18.18	0.00	18.18	0.00	9.09	11
A	52.75	37.36	4.40	2.20	2.20	0.00	0.00	1.10	0.00	91
B	4.62	35.64	46.37	10.89	1.82	0.33	0.00	0.00	0.33	606
C	0.28	1.96	31.37	50.90	13.35	1.62	0.11	0.11	0.28	3,513
D	0.00	0.01	0.29	39.72	33.19	16.62	5.09	2.25	2.82	38,168
E	0.00	0.01	0.15	14.04	29.66	32.09	13.57	5.04	5.45	8,440
F	0.00	0.00	0.02	3.76	14.50	28.48	28.40	15.28	9.56	4,091
G	0.00	0.05	0.05	1.26	5.13	12.98	19.13	30.94	30.46	2,065
U	0.00	0.09	0.09	1.21	2.07	5.44	8.63	20.79	61.69	1,159

Gender

Gender was a significant predictor of GCSE grade improvement, once the other individual and centre characteristics were accounted for. In particular, female students were more likely to improve their grades in GCSE English than male students. Conversely, male students were more likely to improve their grades in GCSE Mathematics than female students.

Average KS4 points per entry

Prior performance (e.g., in GCSE and equivalent qualifications) was positively associated to GCSE grade improvement in both English and Mathematics, with students of high prior attainment more likely than students of low attainment to achieve an improvement.

Centre type[3]

In English and Mathematics, against the baseline of comprehensive schools, candidates in FE colleges were significantly less likely to improve their GCSE grade, once the other candidate characteristics were controlled for. Conversely, candidates in sixth form colleges and schools in the 'Other' category were more likely to improve their grade.

Table 5: Characteristics of candidates improving their GCSE English grade – regression model results

Variables		*Estimate*	*Standard Error*	p-*value*
Intercept		-1.196	0.095	<.0001
Gender	Female [Male]	0.121	0.020	<.0001
Average KS4 points per entry		0.059	0.002	<.0001
Centre type	Sixth form college	0.238	0.101	0.018
	Academy (comprehensive)	-0.042	0.060	0.486
	Academy (modern)	0.152	0.157	0.334
	Academy (selective)	-1.168	0.202	<.0001
	FE college	-0.960	0.072	<.0001
	Grammar	0.045	0.659	0.945
	Independent	0.028	0.109	0.798
	Other	0.491	0.161	0.002
	Secondary modern [Comprehensive]	0.250	0.224	0.264
No. of attempts in GCSE English		-0.370	0.012	<.0001
Resitting GCSE Maths	No [Yes]	-0.077	0.021	0.000
Percentage of Level 3 qualifications		0.014	0.000	<.0001

There were some contrasting results for English and Mathematics: Against the baseline of comprehensive schools, candidates in selective academies were less likely to improve their GCSE English grade (no significant effect in Mathematics) than candidates in comprehensive schools. Similarly, candidates in independent schools were more likely to improve their GCSE Mathematics grade (no significant effect in English) than candidates in comprehensive schools.

3. Note that 55% and 47% of the students retaking GCSE English and GCSE Mathematics, respectively, were in FE colleges; around 20% of students in each subject were in comprehensive academies; between 9% and 12% were in sixth form colleges or comprehensive schools; and just below 5% were in independent schools.

Table 6: Characteristics of candidates improving their GCSE Mathematics grade – regression model results

Variables		*Estimate*	*Standard Error*	p-*value*
Intercept		0.328	0.089	<.0000
Gender	Female [Male]	-0.270	0.020	<.0001
Average KS4 points per entry		0.026	0.002	<.0001
Centre type	Sixth form college	0.463	0.081	<.0001
	Academy (comprehensive)	-0.007	0.048	0.883
	Academy (modern)	0.029	0.124	0.818
	Academy (selective)	-0.087	0.241	0.717
	FE college	-0.518	0.059	<.0001
	Grammar	-0.441	0.524	0.401
	Independent	0.571	0.093	<.0001
	Other	0.421	0.126	0.001
	Secondary modern [Comprehensive]	0.092	0.172	0.593
No. of attempts in GCSE Maths		-0.341	0.010	<.0001
Resitting GCSE English	No [Yes]	-0.080	0.022	0.000
Percentage of Level 3 qualifications		0.008	0.000	<.0001

Candidates in non-selective academies (comprehensive or secondary modern), secondary modern schools or grammar schools were not significantly more or less likely to improve their GCSE grade in either subject than candidates in comprehensive schools.

Number of attempts in GCSE English/Mathematics

The probability of improving the grade in GCSE English or in GCSE Mathematics decreased with an increasing number of resits in the subject. Figure 3 shows that, for example, the probability of improving the grade for students with one resit attempt was around 0.72 in English and 0.76 in Mathematics, for those with two attempts decreased to 0.63 in English and 0.69 in Mathematics and, for those with three attempts to 0.55 in English and 0.61 in Mathematics (note that this is for a female student, in a comprehensive school, not resitting both English and Mathematics, with average KS4 prior attainment, and with 40 per cent of their qualifications taken at Level 3). However, it should be noted that the students who resat English and/or Mathematics several times might have been those who struggled the most with these subjects and, therefore, their chances of improving the grade were low.

Resitting GCSE Mathematics/English

Attempting a resit in GCSE Mathematics as well as resitting GCSE English (or the other way around) was significantly associated with a higher probability of improving the grade. The effect was, however, fairly small (see Tables 5 and 6).

Percentage of Level 3 qualifications taken alongside

The volume of Level 3 qualifications taken by students resitting a GCSE in English or Mathematics was positively associated with GCSE grade improvement. In particular, Figure 4 shows that students with higher percentages of Level 3 qualifications were more likely than students with lower percentages (or no Level 3 qualifications) to improve their GCSE English and Mathematics grades. As before, we should note that the

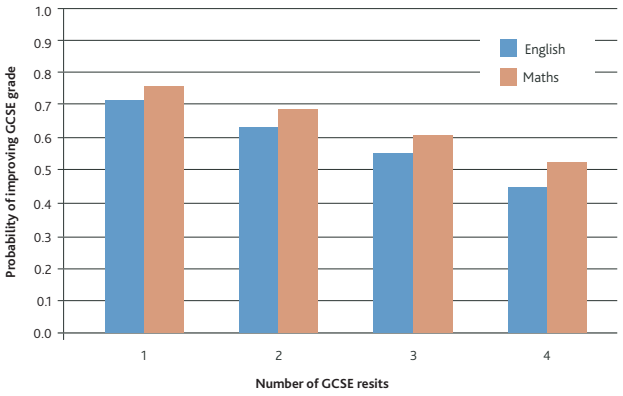

Figure 3: Probability of improving the GCSE grade, by number of resits

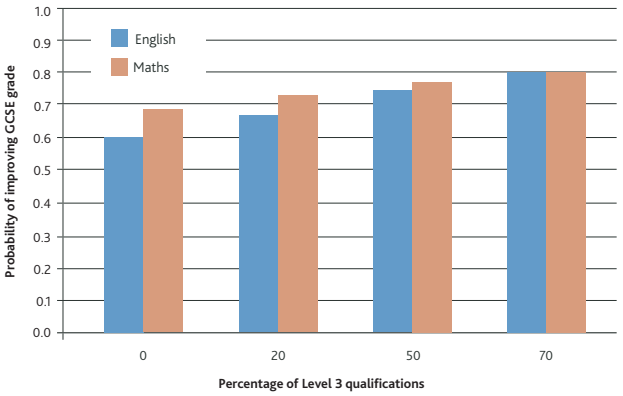

Figure 4: Probability of improving the GCSE grade, by the percentage of Level 3 qualifications

probabilities shown in Figure 4 are for a female student, in a comprehensive school, not resitting both English and Mathematics, with average KS4 prior attainment, and one resitting attempt in the subject. This could be the result of students taking a higher percentage of Level 3 qualifications (e.g., more AS and A levels) perhaps being more academically motivated than those with lower percentages of qualifications at Level 3.

Effect of resitting GCSE English and/or Mathematics on students' performance at Level 3

This section of the article investigates the effect of resitting GCSE English and/or Mathematics on students' performance at Level 3, which was measured by the total GCE A level and equivalent points score.

A total of 334,655 students (aged 17 at the start of the 2015/16 academic year) were considered for this investigation. These students took, at least, one qualification at Level 3. Just over 13 per cent of them (45,589 students) took GCSE English and/or Mathematics alongside[4].

In a first step, descriptive analyses were carried out to look at the uptake and performance of Level 3 qualifications for two different groups of students: no resits, resitting GCSE English and/or Mathematics.

Table 7 and Table 8 that follow suggest that uptake and overall performance in Level 3 varies by whether the student resits or not.

Table 7 shows that the total number of Level 3 entries was lower for candidates with resits. In particular, the average number of A level subjects attempted by candidates without resits was two, whilst for those with resits was below one (in fact, looking at those figures, students resitting GCSE English and/or Mathematics did not seem very likely to study AS or A level subjects in KS5).

Table 8, which shows the total GCE A level and equivalent points score for the same groups of students, indicates that there was a substantial difference in the performance at Level 3. In particular, the difference between students with no resits and those with at least one was just over 40 points. This is equivalent, for example, to an A level at grade B or two AS levels at grades A and C[5].

Table 7: Uptake of Level 3 qualifications, by resitting behaviour

Level 3 entries	*Resitting*	*Mean*	*SD*	*Min*	*Max*
Total entries at Level 3[6]	No	3.41	0.71	0.17	9.00
	GCSE English and/or Maths	2.61	0.95	0.17	9.00
No. of A levels	No	2.11	1.35	0	7
	GCSE English and/or Maths	0.62	1.08	0	6
No. of AS levels	No	0.91	0.89	0	8
	GCSE English and/or Maths	0.40	0.81	0	9

Table 8: Performance at Level 3 (total GCE A level and equivalent points score), by resitting behaviour

Resitting	*Mean*	*SD*	*Min*	*Max*
No	117.78	51.83	0.00	510.50
GCSE English and/or Maths	75.48	44.07	0.00	305.00

However, the above descriptive analyses do not account for possible differences in the two groups of students (no resits, resitting English and/or Mathematics) and it is necessary to disentangle the effect of the resits from other confounding factors that are likely to affect, in particular, the students' performance.

Therefore, in the following analyses, background characteristics of the different groups of students are accounted for. In order to do so, propensity scores were used to control for imbalances in the characteristics of the students with the different resitting behaviours (e.g., total number of entries at Level 3, number of A levels, number of AS levels, prior attainment at Level 2, gender, ethnicity, special needs, first language, free school meals eligibility, level of deprivation, and type of centre). Overall, performance at Level 3 is then compared for comparable groups of students resitting and not resitting GCSE English and/or Mathematics alongside Level 3 qualifications.

Previous studies carried out at Cambridge Assessment (e.g., Gill, 2014) have used either nearest neighbour methods or inverse probability weighting to match groups. The main practical difficulty of these methods is that the propensity score must be estimated. Researchers have found that a misspecification of the propensity score model can result in bias of the estimated effects (e.g., Kang & Schafer, 2007; Smith & Todd, 2005). As a consequence, the above strategies do not often achieve the goal of balancing the characteristics of the two groups under consideration. However, recent research by Imai and Ratkovic (2014) suggests that this issue can be addressed by adjusting the way in which the propensity score is produced so that it is deliberately designed to

4. In particular, 25,671 students took GCSE English and 27,272 took GCSE Mathematics.
5. For details on the performance point scores for each qualification see DfE (2017).
6. Note that 'Total entries at Level 3' in Table 7 refers to the total number of GCE A level and equivalent entries. There are Level 3 qualifications that are 'smaller' that an A level and, therefore, the total number of entries at Level 3 can be smaller than one.

achieve balance between the groups even if the underlying model (i.e., the model that captures the relationship between the background characteristics and the group a candidate is assigned to) is not correctly specified. Although this method is relatively robust to model misspecification, its successful application requires identifying a complete set of confounders, which is not always possible.

The covariate balancing propensity score (CBPS) methodology has been implemented in the R package CBPS (Fong, Ratkovic, & Imai, 2014), which has been used in this research. Statistical significance of the differences between the groups of students with the different resitting patterns was assessed using the R package 'survey' (Lumley, 2015). This package allows us to calculate the standard errors of the estimates whilst accounting for the multilevel structure of the data. In particular, a two-level multilevel structure was considered, with students clustered within centres.

Results of the estimates of the average performance at Level 3 for both groups of students (no resits, resitting English and/or Mathematics GCSEs) after the covariate balancing propensity score method was applied to the data in this research are given in Table 9. The difference, together with its standard error, is also reported.

Table 9: Performance at Level 3 (total GCE A level and equivalent points score), by resitting behaviour - propensity score estimates

Resitting GCSE English and/or Maths	*Level 3 points score*
No resits	82.18
At least one resit	75.48
Difference/Standard Error	-6.704/0.4974

Table 9 shows that when only 'comparable' candidates are considered in the analyses there is a statistically significant effect, although small, of resitting GCSE English and/or Mathematics on performance at Level 3. This indicates that the differences observed before the propensity score procedure was carried out (Table 8) were largely due to the different composition of the two groups of students.

In particular, the analyses carried out after the propensity score procedure show that the difference in the performance at Level 3 between candidates with different resitting behaviours was just below seven points, which means that candidates resitting English and/or Mathematics in KS5 obtained on average seven points less than similar candidates not resitting the GCSEs. Although smaller than before (Table 8), this difference is still of practical importance (e.g., the number of points is equivalent to a grade E at AS level and it is just a bit short of a grade E at A level) and, therefore, statistically significant.

An assumption for the propensity score estimates to hold is based on the effectiveness of reducing the covariate imbalance between the two groups of students under consideration. In this research, for each background covariate, the absolute values of the mean differences before and after matching were inspected and showed a good match. Note that the propensity score analysis only controls for cohort characteristics that were put into the analysis. There would have been other confounding factors, such as student motivation, that could bias the results but data was not available for them.

Summary and conclusions

Good grades (A*-C) in GCSEs in English and Mathematics are considered the benchmark to which all young people should attain. They are necessary to progress to AS/A level and university, apprenticeships and employment. Without them, students' choices could be reduced. Students who do not get the grades at age 16 can 'remedy' that in KS5.

In this research, there were 72,995 students who sat GCSE English and 67,759 students who sat GCSE Mathematics during their KS5 years. Some of these students only sat the GCSE qualification once during KS5 but others did so multiple times. For example, around 25 per cent of the students in both subjects sat the qualifications twice.

The majority of the students taking English and Mathematics GCSEs during their KS5 years had not achieved a good grade (A*-C) by the end of KS4. Furthermore, the data showed that 53% of the students taking GCSE English and 60% of those taking GCSE Mathematics did not improve their grade, despite one or more attempts. In fact, many of them obtained lower grades than the first time they took the exams. The shadow education secretary has recently said that a shortage of Mathematics and English teachers in schools and FE colleges may lay behind the failure of many students to improve their grades (Griffiths, 2016). Additionally, Impetus (2017) reported that issues with funding might mean that schools are not dedicating enough time to prepare for the resits and therefore are not giving students the chance to achieve a good grade. NHAT (2017), however, reports that forcing young people to resit the qualifications when so many still fail to improve their grades can be demotivating and disheartening, resulting in further disengagement with the subject and little likelihood of improving their previous performance. Nonetheless, and despite the fact that some students will not improve their GCSE English or Mathematics grades in KS5, even after multiple resit opportunities, there are other students who really value the chance to achieve the grade they need to progress to FE or employment.

In order to investigate which students have better chances to improve their GCSE English or Mathematics grades when resitting the qualifications in KS5, multilevel logistic regression analyses were carried out. The outcomes of the analyses showed that female students were more likely to improve their GCSE English grades than males, whilst the opposite was true for GCSE Mathematics. Students of high prior attainment were more likely than students of low prior attainment to achieve an improvement. This last finding supports research from Impetus (2016) that shows that students from disadvantaged backgrounds, who usually have lower prior attainment than students from more affluent backgrounds, are more likely than middle-class or more wealthy students to leave education at age 19 without achieving a good grade in English and/or Mathematics.

Porter (2015) reported that FE colleges had much higher numbers of students who decided to retake English or Mathematics at GCSE. This could be because students with low achievement in these qualifications might be disengaged from school and keen to move to college, or because schools and sixth form colleges have higher entrance criteria for entering post-16 education, and therefore students with lower grades at GCSE have to move to an alternative type of centre. Another explanation could be that FE colleges usually offer a wider range of qualifications, including at Level 2 and below, than other types of centres. Our research showed that against the baseline of comprehensive schools, students in FE colleges were significantly less likely to improve their GCSE grades.

And, conversely, students in independent schools were more likely to get better grades in their resits than in their first attempt.

The regression analyses also showed that the probability of improving the grade in English or Mathematics decreased with the number of resitting attempts. However, the students with more resits might be those who struggle the most with these subjects and, therefore, their chances of improving the grade are low. Resitting both English and Mathematics was, however, significantly associated with a higher probability of improvement.

The students retaking English and/or Mathematics in KS5 could also be studying a variety of different courses at different levels. This research showed that, students with higher percentages of Level 3 qualifications were more likely than students with lower percentages (or no Level 3 qualifications) to improve their grades. This may be because students who are trying to achieve a higher level qualification are more motivated to get a good grade in their GCSEs than those who are not taking any Level 3 qualification at the same time.

The fact that students take GCSE English and Mathematics in KS5 has an impact on the number of and performance in AS/A levels and other Level 3 qualifications. As expected, this research showed that the total number of Level 3 entries, and in particular the number of AS/A level qualifications, was lower for candidates with resits than for those without resits. There was also a difference in the performance at Level 3 between the group of students who resat English and/or Mathematics and the group of students who did not, even after taking into account students' background characteristics using propensity score matching techniques. Specifically, the difference between students with no resits and those with at least one was just under seven points. This difference, which is statistically significant and of practical importance, is equivalent, for example, to a grade E at AS level.

Although the policy of improving literacy and numeracy levels amongst school children and ensuring that all young people gain 'good' qualifications in English and Mathematics by the age of 19 seems to be a good idea, its implementation has perhaps not had the intended impact in practice. In fact, Ofsted, DfE advisers and other educational bodies have recently questioned the GCSE resits policy in English and Mathematics (e.g., Ofsted 2016; Belgutay, 2017; Martin, 2017; NHAT, 2017; Offord, 2017; Ward, 2017) for a variety of reasons, as discussed in the introductory section of this article. Firstly, schools and colleges might not have the delivery capacity to offer English and/or Mathematics to KS5 students. Secondly, and as shown in this research, the GCSE resits improvement rates continue to be low. Thirdly, there might be more fitting solutions or alternative pathways to enable students' English and Mathematics skills to develop further (e.g., high-quality Functional Skills qualifications or other qualifications relevant to the world of work). The outcomes of this research could add one more reason to consider whether compulsory resitting of English and Mathematics GCSEs for all students with a grade D is the right policy: the fact that the retakes might be hindering the KS5 prospects of some students.

References

Allen, R. (2016). *Repeat after 'E': the treadmill of post-16 GCSE maths and English retakes*. London: Education Datalab. Retrieved from https://educationdatalab.org.uk/2016/08/repeat-after-e-the-treadmill-of-post-16-gcse-Mathematics-and-english-retakes/

Belgutay, J. (2017). English and maths resits not scrapped for 2017–18. *Times Education Supplement* (2017, April 10). Retrieved from https://www.tes.com/news/further-education/breaking-news/english-and-Mathematics-gcse-resits-not-scrapped-2017-18

Burke, J. (2016). Ofsted annual report says GCSE English and maths resit policy is failing. *FE Week* (2016, December 1). Retrieved from https://feweek.co.uk/2016/12/01/ofsted-annual-report-says-gcse-english-and-math-resit-policy-is-failing

DfE. (2016). *Level 1 and 2 attainment in English and Mathematics by students aged 16–18: academic year 2014/15* (SFR 15/2016). London: Department for Education.

DfE. (2017). *16–19 Accountability Measures: Technical guide for measures in 2016 and 2017*. London: Department for Education.

ESFA. (2017). *Funding guidance for young people 2017 to 2018*. London: Education and Skills Funding Agency.

Exley, S. (2016). Ofsted questions English and maths GCSE resits policy. *Times Education Supplement*. (2016, October 24). Retrieved from https://www.tes.com/news/further-education/breaking-news/ofsted-questions-english-and-Mathematics-gcse-resits-policy

Exley, S. & Belgutay, J. (2017). 'Crippling' GCSE English and maths resits set to rise again. *Times Education Supplement*. (2017, July 14). Retrieved from https://www.tes.com/news/further-education/breaking-news/crippling-gcse-english-and-Mathematics-resits-set-rise-again

Fong, C., Ratkovic, M. & Imai, K. (2014). *R package for Covariate Balancing Propensity Score*. Retrieved from https://cran.r-project.org/src/contrib/Archive/CBPS/

Gill, T. (2014). An investigation of the effect of early entry on overall GCSE performance, using a propensity score matching method. *Research Matters: A Cambridge Assessment publication, 18*, 28–36.

Goldstein, H. (2001). *Multilevel Statistical Models*. New York: John Wiley & Sons.

Griffiths, S. (2016). Half of resit pupils get a lower grade. *The Sunday Times*, p.16. (2016, January 10).

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley & Sons.

Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. Journal of the *Royal Statistical Society Series B, 76*(1), 243–263.

Impetus. (2016). *The road most travelled?* The 16–19 journey through education and training. London: Impetus, The Private Equity Foundation.

Impetus. (2017). *Life after school: confronting the crisis*. London: Impetus, The Private Equity Foundation.

Kang, J.D. & Schafer, J.L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussions). *Statistical Science, 22*(4), 523–539.

Lumley, T. (2015). *R package 'survey': analysis of complex survey samples*. Retrieved from http://cran.r-project.org/web/packages/survey/survey.pdf

Martin, W. (2017). AELP: 2017 should be last year for compulsory GCSE resits. *Times Education Supplement*. (2017, August 23). Retrieved from https://www.tes.com/news/further-education/breaking-news/aelp-2017-should-be-last-year-compulsory-gcse-resits.

NHAT. (2017). Government continues to require students to resit GCSE maths and English. *National Association of Head Teachers* (2017, April 12). Retrieved 23 January 2018 from http://www.naht.org.uk/welcome/news-and-media/key-topics/assessment/government-requires-students-to-resit-gcse-Mathematics-and-english

Offord, P. (2017). DfE blasted for flip-flopping on GCSE maths and English resits. *FE Week*. (2017, April 24). Retrieved from https://feweek.co.uk/2017/04/24/dfe-blasted-for-flip-flopping-on-gcse-Mathematics-and-english-resits

Ofsted. (2016). *The Annual Report of Her Majesty's Chief Inspector of Education, Children's Services and Skills 2015/16*. London: Her Majesty's Stationery Office.

Ofqual. (2016). *Summer exam entries: GCSEs, Level 1/2 Certificates, AS and A levels in England. Provisional figures April 2016*. Coventry: Office of Qualifications and Examinations Regulation.

Ofqual. (2017). *Provisional summer 2017 exam entries: GCSEs, AS and A levels*. Coventry: Office of Qualifications and Examinations Regulation.

Porter, N. (2015). *Crossing the line. Improving success rates among students retaking English and Mathematics GCSEs. A Policy Exchange Policy Bite*. London: Policy Exchange.

Smith, J.A. & Todd, P.E. (2005). Does matching overcome LaLonde's critique of non-experimental estimators? *Journal of Econometrics, 125*(1–2), 305–353.

Ward, H. (2017). DfE adviser calls for maths GCSE resits U-turn. *Times Education Supplement*. (2017, May 26). Retrieved from https://www.tes.com/news/school-news/breaking-news/exclusive-dfe-adviser-calls-Mathematics-gcse-resits-u-turn.

Wolf, A. (2011). *Review of Vocational Education – The Wolf Report* (DfE-00031-2011). London: Department for Education.

How many students will achieve straight grade 9s in reformed GCSEs?

Tom Benton Research Division

Introduction

"It's difficult to make predictions, especially about the future" (Danish proverb)

As General Certificate of Secondary Education (GCSE) qualifications are reformed in England, the grading scale is changing from students being awarded grades A*-G to being awarded grades 9–1, with grade 9 representing the highest grade and also relating to a level of achievement above that of the existing grade A*. This process began in practice in summer 2017 when Mathematics, English Language, and English Literature GCSEs were awarded on the new grading scale. The majority of subjects with large entries will be switching to the new grading scale as part of awarding in summer 2018 and the remainder will be switching in summer 2019[1].

This article attempts to predict the number of pupils who will achieve a perfect set of grade 9s in whichever reformed GCSEs they choose to take. This question sprang to prominence in the media in April 2017 when Tim Leunig, the then chief scientific advisor of the Department for Education (DfE), tweeted that he expected only two pupils to achieve grade 9 in all of their GCSEs. This led to contact between the TES and Cambridge Assessment and, subsequently, to the author giving his own alternative view that 'hundreds' of pupils will achieve grade 9 in every GCSE that they take[2]. For the remainder of this article we will refer to this accomplishment as achieving 'straight' grade 9s.

This article gives more details of how such a prediction might be made. As well as the evident interest in this question externally, it may be of substantive importance as it relates to the extent to which reformed GCSEs, and grade 9 in particular, will be able to discriminate between the very highest performing students.

Since making the original forecast of 'hundreds' of pupils to achieve straight grade 9s in April 2017, more information about attainment in reformed GCSEs has been published by both The Office of Qualifications and Examination Regulation (Ofqual)[3] and the DfE[4]. Naturally, this article makes use of this later information but the rationale is the same as for the earlier predictions. Note that, at the time of writing, the latest national pupil level data available to the author dates from summer 2016.

One method of making the prediction would be to retrospectively set the grade 9 boundary in all existing GCSEs using the formula used to define how many should achieve grade 9 in each subject (see Benton, 2016). It would then be a simple task to just count how many pupils actually attained notional grade 9s in all of the GCSEs they had entered. However, it was not possible to access the raw marks achieved by pupils on a national level, and the techniques employed in this article are entirely based upon data regarding the grades achieved by pupils.

Some simple methods of estimation

To begin with, we consider a very simple way to estimate the number of pupils who will achieve straight grade 9s to illustrate how it might be possible to reach a prediction of around two pupils. The first step is to consider the number of students who achieved straight grade A* in all of their GCSEs historically. Based on Gill (2017), who provides numbers based on students taking at least 5 GCSEs in June 2015, this value might be taken to be 3,300. Next, using an early proposal for the definition of grade 9 (Ofqual, 2014, p.20), we might assume that in every GCSE, around half of those awarded grade A* would be awarded grade 9. Thus, we might assume that, amongst those achieving straight grade A*s, half of these would fail to achieve grade 9 in the first GCSE we consider. This leaves just 1,650 candidates. Applying the same idea to the second GCSE again reduces the number by half to 825 pupils. If we continue with this process of halving the values until we reach 10 GCSEs, then we end up with a prediction of just 3 pupils to achieve straight grade 9s.

However, there are a number of flaws in the above calculation. Firstly, in each subject, the percentage of candidates who will be awarded grade 9 as a percentage of those who would have been awarded grade A* is a little higher than 50 per cent. It varies between subjects, as the percentage who will be awarded grade 9 is tied to the percentage historically awarded grade A or above rather than A* (see Benton, 2016).

1. https://www.gov.uk/government/publications/get-the-facts-gcse-and-a-level-reform/get-the-facts-as-and-a-level-reform
2. https://www.tes.com/news/school-news/breaking-news/exclusive-major-exam-board-predicts-hundreds-will-get-straight-grade
3. https://www.gov.uk/government/news/guide-to-gcse-results-for-england-2017
4. https://www.gov.uk/government/statistics/revised-gcse-and-equivalent-results-in-england-2016-to-2017

However, on average it is slightly over 60 per cent. Secondly, the logic ignores the possibility that the more grade 9s you have already achieved, the more likely you are to get the next one. For example, although we might only expect 60 per cent of those who have achieved the equivalent of grade A* or above in an individual subject to be awarded grade 9, the percentage of candidates who will be awarded grade 9 out of those who have achieved the equivalent of grade A* in this subject *and achieved grade 9 in all of their other subjects* should be somewhat higher. Taking account of this fact is crucial if we are to make an accurate prediction. Finally, the calculations in the preceding paragraph assume that all of the students we are interested in took 10 GCSEs. In reality, the number of GCSEs taken will vary between candidates.

One way to account for the correlations between achievement in different subjects, and thus the fact that those getting grade 9 in some will be more likely to get grade 9 in others, is to assume that all candidates have an underlying level of general ability that influences their achievement in all of the GCSEs that they take. This idea has been prevalent in psychometrics for more than 100 years (see Spearman, 1904) and has previously been used to help analyse possible differences in difficulty between subjects (Coe, 2008). The idea is that each person has an, unmeasured, level of ability from somewhere on the normal distribution that influences their achievement in any assessment. In theory this ranges from minus infinity for people with zero chance of answering any question to plus infinity for anyone who is all-knowing. However, for practical purposes, nearly all people would be within the range -4 to +4 and we focus on working on the probability of people in this range getting straight grade 9s in all their GCSEs. Although there are some weaknesses in this approach that will be described later, the calculations are simple enough to be performed using no software more complicated than Microsoft® Excel, and will also serve to illustrate some of the difficulties involved with predicting how many pupils will get grade 9 in all their GCSEs. The details of the calculation steps are shown in Table 1. Where applicable, the Excel formulae that were used to complete calculations are shown.

To begin with, we specify the percentage of candidates we would expect to achieve grade 9 in any subject if the given subject was taken by every eligible student in the country. This percentage is set to be 3.1 to match the average percentage awarded grade 9 across the three reformed GCSEs awarded in summer 2017[5]. Next, we find the equivalent cut point in the normal distribution (1.87). This means that, if nationally available raw marks from each GCSE were transformed to a scale with a standard normal distribution with a mean of 0 and a standard deviation of 1, then the grade 9 boundary would be located at this point.

5. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/639824/GCSE_results_2017_infographic.pdf.

Table 1: Calculation steps based on assuming a single underlying latent trait

			Percentage of candidates who will get grade 9 or above in each subject nationally *(pc9)*		3.10%
			Normalised cut-off for grade 9 *(cutoff)* (=NORMINV(*pc9*,0,1))		1.87
			Correlation between general and specific abilities *(correl)*		0.75
A. Normalised general ability	*B. Weight*	*C. Expected specific normalised score (=A x* correl*)*	*D. Standard deviation of score given general ability (=SQRT(1-*correl*^2))*	*E. Probability of getting grade 9 in any one subject (1-NORMDIST(*cutoff*,C,D,TRUE))*	*F. Probability of getting 10 x grade 9s (=E^10)*
4	0.01%	-3	0.66	0.00%	0.00%
-3.5	0.04%	-2.63	0.66	0.00%	0.00%
-3	0.22%	-2.25	0.66	0.00%	0.00%
-2.5	0.88%	-1.88	0.66	0.00%	0.00%
-2	2.70%	-1.5	0.66	0.00%	0.00%
-1.5	6.48%	-1.13	0.66	0.00%	0.00%
-1	12.10%	-0.75	0.66	0.00%	0.00%
-0.5	17.60%	-0.38	0.66	0.04%	0.00%
0	19.95%	0	0.66	0.24%	0.00%
0.5	17.60%	0.38	0.66	1.21%	0.00%
1	12.10%	0.75	0.66	4.57%	0.00%
1.5	6.48%	1.13	0.66	13.12%	0.00%
2	2.70%	1.5	0.66	28.99%	0.00%
2.5	0.88%	1.88	0.66	50.52%	0.11%
3	0.22%	2.25	0.66	71.91%	3.70%
3.5	0.04%	2.63	0.66	87.43%	26.10%
4	0.01%	3	0.66	95.67%	64.26%
Total expected to get straight grade 9s					0.0248%
Expected number to get straight grade 9s out of 500,000					124

The next part of the calculation requires us to specify the expected correlation between each student's underlying general ability and their normalised scores in each individual subject. This figure is chosen as 0.75 as previous research has shown that for large scale GCSEs a typical correlation between subject-grade and mean GCSE grade is 0.75 (Benton & Sutch, 2013, p.7, Table 2). As such, it provides a reasonable idea of the link between general ability and specific ability in particular subjects.

To complete calculations, we assume that underlying general ability follows a normal distribution nationally. Then, for each possible level of underlying ability (column A in Table 1) we can calculate:

- the proportion of candidates we expect to have this level of ability (column B);
- the expected normalised score of candidates with this ability in each individual subject (column C) and how much uncertainty there is in scores in specific subjects given general ability (column D);
- using C and D, the probability of the candidate getting grade 9 in an individual subject given their level of general ability (column E); and
- the probability that this achievement of grade 9 will be repeated across 10 different GCSEs (column F). This is calculated as just column E to the power of 10, the assumption being that *given candidates' underlying ability* achievement in separate GCSE subjects is independent.

Note that, although in reality we expect general ability to form a continuum, this is approximated by just 17 points on this scale between -4 and +4. This method of approximation, known as quadrature, is commonly used within psychometrics and the weights shown in column B are just set to be proportional to the density of the standard normal distribution, but also to sum to 100 per cent.

By taking a weighted average (weights in column B) of the values in column F we can estimate that less than 0.03 per cent of candidates (that is, less than 3 in 10,000) would be expected to achieve straight grade 9s across 10 GCSEs. If we imagine a GCSE cohort of 500,000 candidates this would mean that just over 100 of them would achieve straight grade 9s.

There are a number of flaws in these calculations, but before discussing these it is worth noticing the values in column F. For example, it is interesting to note in these calculations that even a candidate with an underlying general ability of 2.5, which would be enough to place them in the top 1 per cent of performers, still has a vanishingly small chance of achieving straight grade 9s across 10 subjects. In fact, nearly all of the candidates predicted to achieve straight grade 9s come from the final three rows of ability – that is candidates at or above the 99.9th percentile of general ability. It is very rare that calculations in education need to focus upon such extreme values. As such, the predicted numbers of candidates to achieve straight grade 9s are very sensitive to the assumptions underlying the model.

Table 2 begins to show some of this sensitivity. Moving from top to bottom allows us to see the impact of the assumed number of GCSEs on calculations. As can be seen, the more GCSEs are taken by candidates, the fewer the number of candidates we expect to achieve straight grade 9s. After all, it is easier to get straight 9s in 8 subjects than in 10. In the first row, the number expected to get straight grade 9s across 3 subjects is included as this allows direct comparison with the 2,050 candidates known to have achieved straight grade 9s in all of Mathematics, English Language and English Literature GCSEs in 2017[6].

Moving from left to right shows the impact of the assumed correlation between individual GCSEs and general ability. If this level of correlation were to drop in reformed GCSEs from its historical level to a substantially lower value of 0.7, then our prediction of the number of straight grade 9 candidates would decrease by nearly two-thirds. In contrast, if it were to increase to 0.8, then the number of straight grade 9 candidates would nearly double. In fact, analysis of data collated from Ofqual's web analytics page[7] suggests that the correlation between English Language and Mathematics grades in reformed GCSEs is close to 0.7 which would imply a correlation of both with underlying general ability of around 0.8 (the square root of 0.7). With this in mind, it is of no surprise that it is this value that gives the closest match to the known actual number of straight grade 9s for the three reformed GCSEs in 2017 (i.e., the published number of 2,050). However, this can not necessarily be taken to imply that this value is the most appropriate one for predictions into the future.

Table 2: Predicted number of candidates who will achieve straight grade 9s from a cohort size of 500,000 depending upon correlations between general ability and individual subjects and the number of GCSEs taken

No. of GCSEs being taken by each candidate	*Expected correlation between general ability and scores in individual GCSEs*		
	0.7	*0.75*	*0.8*
3	1073	1560	2260
...	...	...	...
8	82	196	440
9	60	154	365
10	45	124	309

Although interesting, these calculations still have some weaknesses. For example, they assume that all candidates take the same set of subjects, thus ignoring the impact of subject choice on candidates' likely achievement (see Benton & Bramley, 2017). They also assume a common correlation between all subjects and general ability rather than noticing that some subjects (e.g., English Language and English Literature) are more strongly correlated to each other than to others. Finally, they again assume that all candidates take the same number of GCSEs. For these and other reasons, and in order to improve accuracy, we move from predictions based almost entirely on theory to predictions built directly from pupil level data.

Empirical estimates based on data from 2016

The data set

The data for analysis was taken from the National Pupil Database (NPD), which is held by the DfE and consists of results for all students in all subjects in schools and colleges in England. The analysis focussed upon the GCSE results of all candidates in Year 11 in the academic year 2015/16. This was the most recent set of national data available to the

6. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/639824/GCSE_results_2017_infographic.pdf.
7. https://analytics.ofqual.gov.uk/apps/2017/GCSE/9to1/

author at the time of writing. International GCSEs were counted as GCSEs in the analysis as they were a common alternative qualification that is judged to be equivalent (Gill, 2017). However, the NPD only included the results of international GCSEs which had been accredited for use in state schools. There were also several non-accredited international GCSEs which were taken by some students attending independent schools and which, therefore, were not in the NPD. As such they could not be counted in the analysis in this article.

For this particular cohort of students, early entry was fairly common in some subjects such as English Language and Mathematics. To ensure that the cohort of students with results in Mathematics, English Language, and English Literature used in our analysis was of roughly the same size as the cohort taking the three reformed subjects in 2017, it was necessary to include early entries. However, if there were students with multiple entries, rather than simply take each student's highest grade in each subject, results within GCSEs were taken in preference to results in international GCSEs (as this article is really concerned with GCSEs), and results from June 2016 were taken in preference to any earlier results. Entries to the (now discontinued) combined GCSE English Language and Literature were counted as if they were English Language entries.

In order to restrict the analysis to a manageable number of GCSE subjects, our analysis first looked at the most common subject choices amongst students who achieved straight grade A* grade across at least five different GCSEs. Table 3 shows all GCSE subjects taken by at least 40 straight grade A* candidates. Subsequent columns show the percentage of straight grade A* candidates taking each subject compared to the percentage of all candidates taking each subject (of those who took at least five GCSEs). There are some very large differences in subject popularity between these two groups. For example, whilst more than half of all students chose to study Combined Science (Science [Core] and Additional Science), only just over one in twenty straight grade A* candidates chose these subjects, whilst the popularity of the separate sciences (Biology, Chemistry, Physics) was much higher amongst the straight grade A* candidates. Both Ancient and Modern Languages were far more popular amongst the group of candidates who achieved straight grade A*s than amongst GCSE candidates as a whole. Indeed, it is notable than GCSE Chinese, taken by less than 0.5% of candidates nationally, was taken by almost 7% of straight grade A* candidates. Similar comments, though to a slightly lesser extent, could be made about both GCSE Russian and GCSE Italian. The biggest differences of all can be seen for GCSE Latin and GCSE Classical Greek taken by around a third and a tenth of straight grade A* candidates respectively compared to around 1% and 0.2% of candidates nationally. To illustrate this further, the final column of Table 3 shows that almost 40% of *all candidates* who entered GCSE Classical Greek actually achieved straight grade A*s across at least 5 GCSE entries.

Analysis was restricted to candidates taking at least one of the subjects in Table 3 and at least three GCSEs in total. Having made these choices, some descriptive information about the data set used for analysis is shown in Table 4. In particular, for those candidates with results in all of Mathematics, English Language, and English Literature, some comparisons are made to data published by Ofqual after summer 2017. After making this comparison, there was some concern over the fact that the percentages of candidates achieving grade A (equivalent to grade 7 in 2017) and above in these subjects (and in Mathematics in particular) were larger in the analysis data set used in this article than in Ofqual's published data from summer 2017. The reasons for these differences are not known. Potentially they may relate to decisions within high-performing independent schools to continue to use unreformed international GCSEs rather than switch to using reformed GCSEs at this stage. However, regardless of the reasons for the differences, to improve the comparability of the two data sets, 8,000 candidates who had taken all three subjects of interest and had achieved grade A in Mathematics were randomly selected for removal from the data set. As shown by the final column of Table 4, although this step reduced the overall number of candidates used in analysis, it ensured that the number of high-performing candidates in each subject was closer to that within Ofqual's published data.

Although we cannot assume this completely removed the differences in the characteristics of candidates in the two data sets, it should help to ensure we make valid comparisons. It should be noted that if these candidates were added back in to analysis, the later predictions of the number of candidates to achieve straight grade 9s would increase.

Table 3: The most popular GCSE subjects taken by candidates achieving straight grade A*s across at least five GCSEs in 2016

GCSE subject	*No. of straight grade A* candidates*	*% of straight grade A* candidates taking subject*	*% of all candidates taking subject*	*% of all subject entrants that get straight grade A*s*
English Literature	3,421	88.6%	93.9%	0.7%
Biology	2,386	61.8%	26.3%	1.7%
History	2,328	60.3%	45.6%	0.9%
Physics	2,225	57.6%	25.9%	1.6%
English Language	2,215	57.4%	92.6%	0.4%
Chemistry	2,164	56.1%	25.9%	1.5%
French	2,151	55.7%	26.0%	1.5%
Geography	2,029	52.6%	42.2%	0.9%
Mathematics	1,905	49.4%	95.7%	0.4%
Religious Studies	1,651	42.8%	48.0%	0.6%
Latin	1,325	34.3%	1.2%	20.3%
Spanish	1,161	30.1%	16.6%	1.3%
German	807	20.9%	9.3%	1.6%
Music	566	14.7%	7.5%	1.4%
Classical Greek	363	9.4%	0.2%	39.9%
Art & Design (Fine Art)	352	9.1%	8.8%	0.7%
Art & Design	325	8.4%	14.2%	0.4%
Computing	291	7.5%	11.3%	0.5%
Chinese	268	6.9%	0.6%	8.4%
Drama & Theatre Studies	251	6.5%	12.1%	0.4%
PE/Sports Studies	230	6.0%	20.5%	0.2%
Science (Core)	226	5.9%	69.6%	0.1%
Additional Science	221	5.7%	63.2%	0.1%
D&T: Resistant Materials	177	4.6%	8.4%	0.4%
Statistics	162	4.2%	8.8%	0.3%
Italian	162	4.2%	0.8%	3.9%
ICT	161	4.2%	13.6%	0.2%
Business Studies	134	3.5%	13.4%	0.2%
Russian	127	3.3%	0.3%	6.9%
D&T: Product Design	85	2.2%	6.6%	0.2%
Classical Civilisation	84	2.2%	0.6%	2.5%
Economics	81	2.1%	1.7%	0.9%
D&T: Textiles Technology	66	1.7%	3.9%	0.3%
Astronomy	60	1.6%	0.3%	3.2%
D&T: Food Technology	60	1.6%	6.0%	0.2%
D&T: Graphic Products	47	1.2%	4.7%	0.2%

Table 4: A comparison of the analysis data to figures published by Ofqual for results from summer 2017

Statistic	*Ofqual data (2017)*	*Initial analysis data set (2016)*	*Final analysis data set (2016)*
Total number of pupils	-	574,879	566,879
No. of pupils with grades in all of Maths, English Language, and English Literature (3-subject candidates)	508,950	506,226	498,226
No. of 3-subject candidates achieving grade A/7 or above in Maths	102,950	111,061	103,061
No. of 3-subject candidates achieving grade A/7 or above in English Language	84,750	89,311	85,292
No. of 3-subject candidates achieving grade A/7 or above in English Literature	96,050	98,121	93,883

Having collated the data for analysis, three different methods were applied to attempt to make predictions of how many candidates would achieve straight grade 9s. In each case, predictions of how many of the candidates who had taken all of Mathematics, English Language, and English Literature we would expect to achieve straight grade 9s *in these subjects only* were also produced so that these could be compared to the known results from summer 2017. Finally, as it may be relevant to performance tables, predictions were also made for how many candidates we would expect to achieve a perfect score in their *Attainment 8*[8] measure used for accountability.

Prediction using uni-dimensional item response theory method

The first set of predictions were made using a very common method with psychometrics – that of item response theory (IRT). As with the last set of theoretical calculations shown earlier, the method assumes that all of the relationships between achievements in different subjects can be explained by a single underlying latent trait (or general ability). However, the form of IRT model we used (the graded response model) allows for the fact that some subjects may be more strongly related to this underlying general ability than others (perhaps such as Art) which may consist of more specific skills. In addition, it is not assumed that, if all students took all subjects, then the grade distributions would be the same in every one of them.

Specifically, we define Y_{ij} as the grade achieved by the ith student in the jth GCSE subject being analysed. Then the probability that they achieve grade k or above is calculated as:

$$P(Y_{ij} \geq k) = \frac{\exp(\alpha_j\theta_i - d_{jk})}{1 + \exp(\alpha_j\theta_i - d_{jk})}$$

In this notation θ_i is the ability of student i. Abilities are defined to follow a normal distribution nationally with a mean of 0 and a standard deviation of 1. The α_j parameter defines the strength of the relationship between general ability and the grades achieved within subject j. In this notation, the d_{jk} (or difficulty) parameters specify the log-odds of an average ability student achieving below grade k. The values of d_{jk} increase for higher grades as, obviously, the chances of getting, say, grade A or above are lower than the chances of getting grade B or above.

The α_j and d_{jk} parameters for each subject were estimated using the national data set of GCSE results from 2016 described earlier and the R software package mirt (Chalmers, 2012). Although it is possible to also produce direct estimates of each individual candidate's ability (θ_i), such estimates are unlikely to properly reflect the full national distribution of candidate ability. For this reason, the method of plausible values is used instead (Wu, 2005). This method is commonly used in the analysis of large-scale international surveys such as the Programme for International Student Assessment (PISA). Rather than trying to get the most accurate estimate of ability for each individual student, these values are chosen for each student to be within the likely range of their true values given their GCSE grades in each subject, but also so that they are likely to have the correct distribution across the population as a whole (Marsman, Maris, Bechger, & Glas, 2016).

The combination of item parameters and (plausible) ability estimates allow us to simulate the likely grades of any students in any combination of GCSE subjects. However, so far, because grade 9 had not been defined in 2016, such simulations only go up as far as grade A*. In order to go beyond this, one more step is required – the imputation of *plausible marks*.

One way to view the graded response IRT model, as previously defined, is that each student's achievement in any particular GCSE subject follows a logistic distribution centred around their scaled ability $\alpha_j\theta_i$. Specifically, we might define M_{ij} as some monotonic transformation of the marks that candidate i achieves in subject j. Then, we can say that:

$$M_{ij} = \alpha_j\theta_i + \varepsilon_{ij}$$

where ε_{ij} is the candidate's overachievement in subject j relative to their general ability and the values of ε_{ij} in any two subjects are independent. We also define the ε_{ij} to follow a logistic distribution. Finally, we use the values of d_{jk} to be the grade boundaries on this mark scale so that a candidate's grade is greater than or equal to k if $M_{ij} \geq d_{jk}$.

Using this formulation, it is possible to simulate plausible marks for all candidates in all of the subjects that they have taken. Simulated plausible grades can be created by using the d_{jk} parameters as grade boundaries. Note that, since we are working from simulated marks, each candidate's simulated grade may not match their actual grade in each subject. Next, we identify a grade 9 boundary on this newly defined mark scale. In order to do this, we first calculate the percentage of candidates that achieved grade A or above in the subject. Now, according to the definition of grade 9, the percentage of candidates who achieve grade 9 is derived from the percentage achieving grade A or above using the tailored formula specified in Benton (2016). Once this percentage is calculated, we simply identify the location of a grade 9 boundary on the simulated mark scale so that the percentage of candidates achieving grade 9 matches the intended number.

With simulated marks and grade boundaries in place it is then possible to calculate exactly how many students achieved straight grade 9s given their simulated marks and this provides the basis of our main prediction.

However, some initial checks on these predictions led to some

8. Attainment 8 is calculated by looking at each candidate's grade in Mathematics, their best grade from English Language and English Literature, the best three grades they achieve in any EBacc subjects (Science subjects, Computer Science, History, Geography, and Languages), and their best three grades from GCSEs not already used within previous categories. For further details see https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/676184/Secondary_accountability_measures_January_2018.pdf.

concerns over their accuracy. For example, when used to simulate grade A*s (i.e., before the grade 9 boundary had been defined) the method predicted that more than 6,300 students would achieve grade A* in all of Mathematics, English Language, and English Literature, when in reality fewer than 5,900 did so. Of slightly greater concern was the fact that the model predicted that more than 2,800 candidates would achieve grade 9 in these same 3 subjects when statistics published by Ofqual have revealed that only 2,050 candidates achieved this in summer 2017.

Attempts to correct these overestimates by fitting more complex, multi-dimensional IRT models were unsuccessful. For this reason, an alternative method that was not reliant on IRT was attempted.

Prediction using plausible marks from logistic regression and actual grades

Rather than relying on a particular IRT model, an alternative was to attempt to derive plausible marks using logistic regression. The first step was to model each candidate's probability of achieving a grade A* in each GCSE subject as a function of their mean GCSE grade in all of their other GCSEs. The formula for logistic regression defines the probability of achieving a grade A* as:

$$P(Y_{ij} \geq A^*) = \frac{\exp(\beta\ MeanGCSE_i\ \alpha_j)}{1 + \exp(\beta_j MeanGCSE_i - \alpha_j)}$$

As can be seen, this formula is very similar to the one for the graded response model outlined earlier. However, rather than relying on a latent variable (θ) that is assumed to follow a normal distribution, probabilities are modelled based on an observed variable (mean GCSE grade). The β_j parameters defined the strength of the relationship between grades in a particular GCSE and the mean grade in other GCSEs. Only a single intercept parameter (α_j) is defined as this model just focusses upon grade A* as this is the most informative existing grade for the research question being studied.

The fitted logistic regression model can be used to produce plausible marks in each GCSE as before using the equation:

$$M_{ij} = \beta_j MeanGCSE_i + \varepsilon_{ij}$$

where ε_{ij} is the candidate's overachievement in subject j relative to mean GCSE grade in other subjects and is simulated from a logistic distribution. However, to improve upon this method, we make a further amendment so that for each candidate their simulated plausible marks will be consistent with the actual grade they achieved. This method is illustrated in Figure 1. The top two panels show the possible score distributions from standard simulation for two candidates who, in reality, both achieved grade A* in Mathematics. The candidate on the left had an average grade of A in their other GCSEs, whereas the candidate on the right achieved straight grade A*s. The red dotted line shows the grade A* boundary on the scale of plausible marks. Notice that both candidates have a high chance of being simulated a plausible mark below this boundary even though in reality both achieved grade A*. To address this we can instead use conditional simulation as illustrated in the bottom two panels of Figure 1. In this method, each candidate's plausible mark is selected from the truncated part of the distribution above the grade A* boundary. Note that the mean of this truncated distribution for the candidate who had achieved straight grade A*s in all of their other GCSEs remains (slightly) higher than the mean for the candidate who had only averaged at grade A. In this way, the simulation ensures that even within candidates who have achieved a grade A* in a given subject, we expect the highest marks to occur amongst students with high attainment elsewhere.

Note that, as shown in Figure 2, the method of simulation we have described makes almost no difference to the overall distribution of

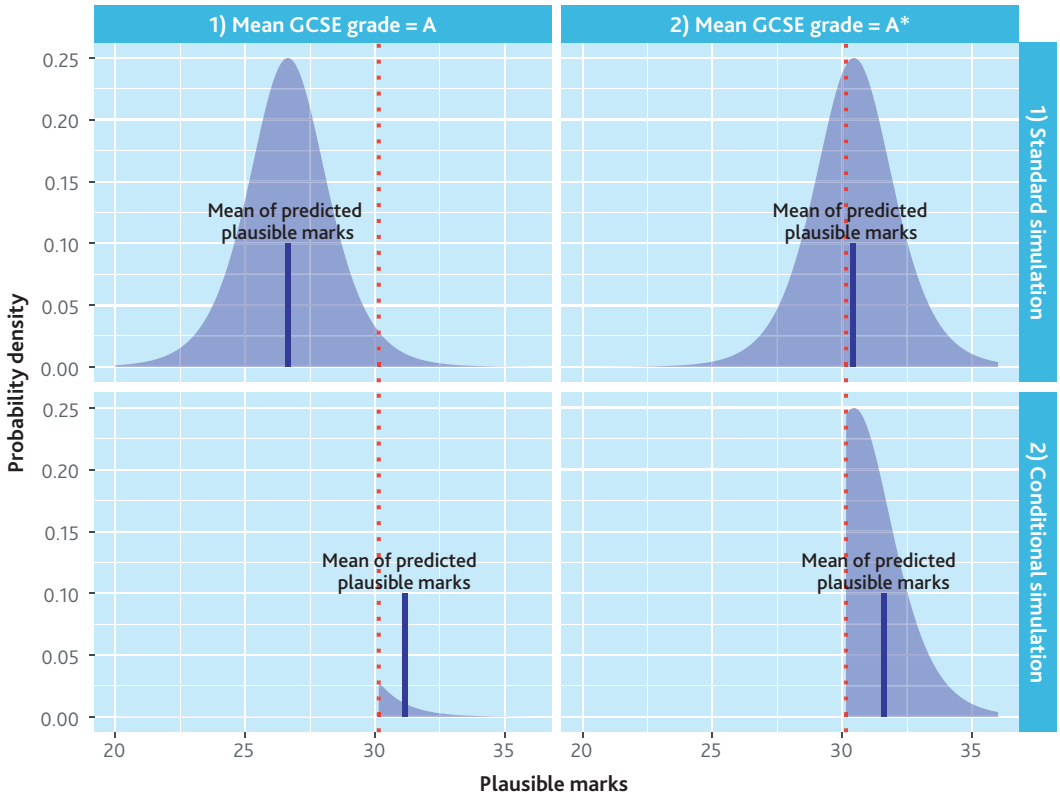

Figure 1: Illustration of different methods of simulating plausible marks for two candidates who had achieved grade A* in Mathematics

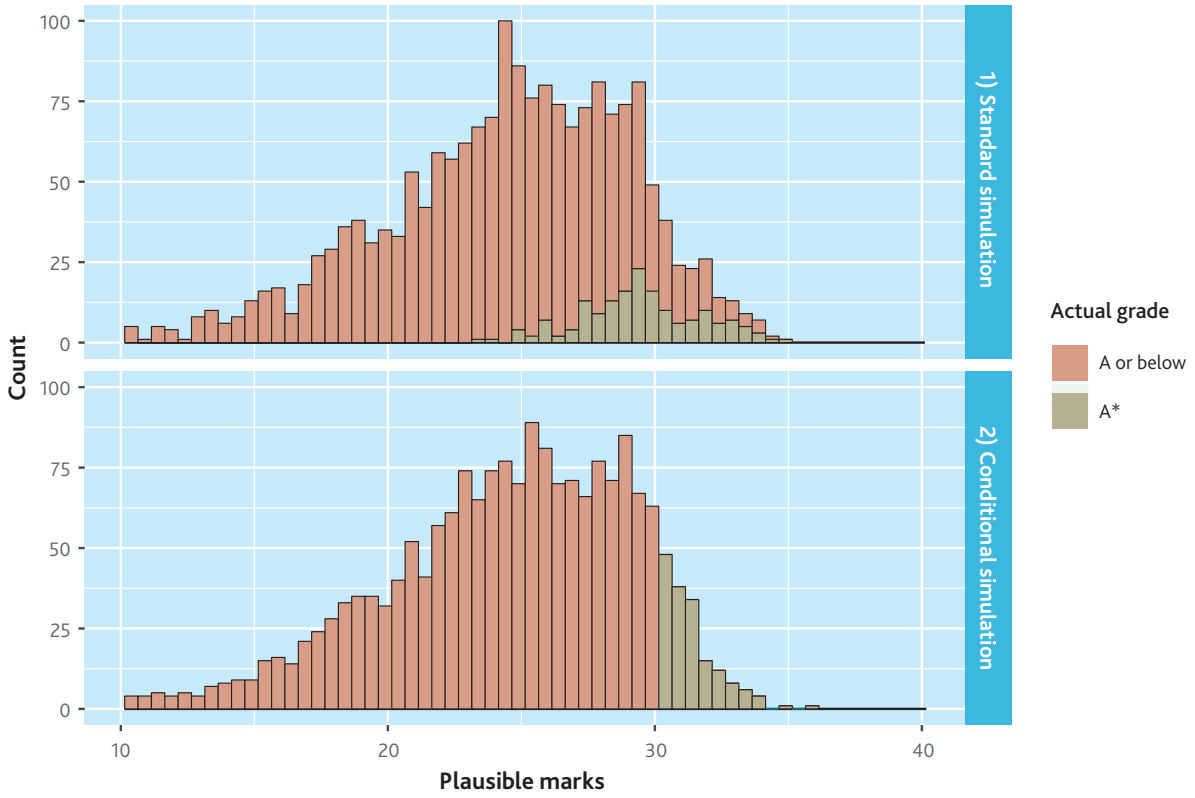

Figure 2: Overall distributions of plausible marks in Mathematics based on standard simulation and simulation conditional upon the actual grade candidates achieved

plausible marks. The overall distribution of plausible marks remains the same but we have ensured that, at least up to grade A*, the combinations of grades achieved by candidates will be consistent with reality. This means that, for example, candidates simulated to achieve straight grade 9s will always come from amongst those who actually achieved straight grade A*s. In this way we ensure that, at least up to grade A*, relationships between subjects are preserved. As such, the simulation is purely being used to ensure that, amongst candidates with a grade A* in any subject, it is recognised that the highest marks are likely to be achieved by candidates who have achieved good grades elsewhere. However, it should be remembered that measured achievement in other GCSEs is still capped at grade A*, and so the method is still likely to underestimate the true level of correlation between the marks of candidates within the highest grades in different subjects. This means that the predictions from this method are likely to be a lower bound to what should be expected in future rather than an exact prediction.

By design, the number of candidates predicted to get straight grade A*s via this method matched the actual number achieving this within the data set. More importantly, the predicted number to achieve straight grade 9s in Mathematics, English Language, and English Literature was very close at 1,829 to the actual number (2,050) reported by Ofqual for summer 2017. Furthermore, the fact that the prediction was marginally lower than the true number fits with the expectation that this method should provide a lower bound.

Prediction using a hybrid method

As we have mentioned, the weakness of the method based on logistic regression was that it only accounted known abilities up to grade A*. As such, it could not simulate the likely effect of very high ability students being likely to achieve well above the grade A* boundary across many subjects. In an attempt to address this, a second version of simulation based on logistic regression was run. However, rather than using the mean GCSE grade as the predictor of plausible marks in each subject, the simulated plausible values of general ability derived from the original IRT analysis were used. Unlike the mean GCSE, these simulated ability values were not capped and so, when used to further simulate plausible marks, could allow for higher correlations between different subjects at the top end of achievement. As we described for the previous method, plausible marks were simulated conditionally upon the actual grade that students achieved in each subject.

Again, by design, with the hybrid method the number of candidates predicted to get straight grade A*s matched the actual number achieving this within the data set. Also, as expected given that the cap on general ability had been lifted, the predicted number of students to achieve straight grade 9s in Mathematics, English Language, and English Literature rose considerably to 2,607 which is higher than the officially published number of 2,050. This may potentially indicate continuing weaknesses in the method. However, as shown earlier in Table 4, we already know that the characteristics of the 2017 cohort analysed by Ofqual differ a little from the 2016 data used in this analysis. Although some allowance has been made for this by removing a number of grade A and above Mathematics candidates, it is possible that this has not fully accounted for the differences between the data sets. For example, at the time of writing, the full extent to which high-performing independent schools have or have not transferred their entries from existing international GCSEs to reformed GCSEs is not known. If many of these centres had not switched over, it may explain the fact that some of our predictions are higher than currently published results. In theory, if more of such centres were to move to taking reformed GCSEs in future, it could substantially increase the number of candidates achieving straight grade 9s.

Summary of results

Predictions from each of the three methods are summarised in Table 5. This table includes comparisons to actual results in terms of straight grade A*s in 2016, published results from Ofqual, and also to statistics published by the DfE regarding the number of candidates achieving grade 9 in Mathematics and then grade 9 in either English Language or English Literature in summer 2017. However, it should be noted that this last published statistic is limited to pupils in state schools and so will be lower than the full number. With this in mind, the results confirm the suggestion that the predictions based on logistic regression using the mean GCSE will be too low and that the predictions using either IRT or the hybrid method will be too high. With these caveats in mind, the results suggest that:

- between 1,000 and 2,000 candidates will achieve straight grade 9s in at least 3 GCSEs;
- if we restrict ourselves to those taking at least 5 GCSEs, between 600 and 1,500 candidates will achieve straight grade 9s;
- of those taking at least 8 GCSEs, between 200 and 900 candidates will achieve straight 9s;
- of those taking at least 10 GCSEs, between 100 and 600 candidates will achieve straight 9s; and
- thinking about the Attainment 8 accountability measure, we should expect more than 2,000 candidates to achieve a perfect score and that the number may be as high as 4,000.

Given the difficulty of attaining grade 9, the size of this last prediction is a particular surprise. The cause for the increase is that, although achieving a perfect score in Attainment 8 requires students to achieve at least eight grade 9s, it does allow for them to achieve below grade 9 in at least some GCSEs. A similar increase can be seen historically in statistics published in Gill (2017) which show that in June 2015, although only 3,300 achieved straight grade A*s, more than 8,500 achieved grade A*s in 8 or more subjects. On a similar theme, although Ofqual statistics show that only 2,050 pupils achieved straight grade 9s in Mathematics, English Language, and English Literature, the DfE's statistics show that, in state schools alone, more than 6,000 pupils achieved perfect scores across both the Mathematics and English 'pillars' of Attainment 8.

Final predictions

Overall, the analysis in this article confirms the initial prediction made in April 2017 that 'hundreds' of pupils will achieve straight grade 9s. If we restrict to candidates taking at least 8 GCSEs then the prediction is that between 200 and 900 of them will achieve straight grade 9s. If we take a purely literal definition of straight grade 9s, and include all candidates regardless of how few GCSEs they have taken, then the number is likely to be greater than 1,000. This article also provides a new prediction that at least 2,000 pupils will achieve perfect Attainment 8 scores in their GCSEs, and that this number may be as high as 4,000.

As might be expected, there are a number of caveats to these predictions. Firstly, it should be noted that these predictions are based upon GCSE and international GCSE entry patterns from June 2016. If GCSE reform leads to major changes in the popularity of different subjects and, in particular, to the numbers of GCSEs taken by different students, then this may have a noticeable impact upon the actual results. In addition, the extent to which high-attaining independent schools, which have historically entered their students for international

Table 5: Predictions and comparisons to (some) known results from 2017

Statistic (No. to achieve …)	*Predictions*				*Actual results*		
	No. of candidates relevant to prediction (base N)	*Method*			*No.*	*… out of (No. of relevant candidates)*	*Source*
		Pure IRT	*Logistic regression-based plausible marks*	*Hybrid*			
… straight grade A*s in Maths, English Language, and English Literature	498,226	6,382	5,891	5,891	5,891	498,226	2016 analysis data set
… straight grade 9s in at least 3 GCSEs	565,431	2,045	1,077	2,054	-	-	-
… straight grade 9s in at least 5 GCSEs	535,216	1,563	620	1,516	-	-	-
… straight grade 9s in at least 8 GCSEs	382,278	894	216	817	-	-	-
… straight grade 9s in at least 10 GCSEs	131,876	619	110	508	-	-	-
… straight grade 9s in Maths, English Language, and English Literature	498,226	2,816	1,829	2,607	2,050	508,950	Ofqual analytics 2017[a]
… grade 9 in both the Maths and English pillars of Attainment 8	537,207	8,388	6,396	7,250	6,129[b]	527,859	DfE statistics 2017[c]
… a perfect Attainment 8 score	566,879	4,247	2,598	3,797	-	-	-

a. https://analytics.ofqual.gov.uk/apps/2017/GCSE/9to1/
b. The published figure from the DfE is restricted to state funded schools only.
c. See Characteristics national tables at https://www.gov.uk/government/statistics/revised-gcse-and-equivalent-results-in-england-2016-to-2017

GCSEs, either have (or will) switch their entries to reformed GCSEs is not known. Whether such schools contribute to the national GCSE results will make a noticeable difference.

Although many GCSE subjects will have been reformed by summer 2018, the final test of these predictions will not be until after summer 2019. In particular, our analysis has shown that some minor Modern Languages (Chinese, Russian, and Italian) are very popular amongst candidates who have achieved straight grade A*s historically and so, only when results for the reformed versions of these subjects are available (summer 2019), will we know the accuracy of our predictions.

Regardless of whether the predictions are right or wrong, one thing is clear: Achieving grade 9 in any GCSE subject is hard. Congratulations to all those students who achieve it in any subject at all.

References

Benton, T. (2016). *A possible formula to determine the percentage of candidates who should receive the new GCSE grade 9 in each subject*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/298710-a-possible-formula-to-determine-the-percentage-of-candidates-who-should-receive-the-new-gcse-grade-9-in-each-subject.pdf.

Benton, T. & Bramley, T. (2017). *Some thoughts on the 'Comparative Progression Analysis' method for investigating inter-subject comparability*. Cambridge Assessment Research Report. Cambridge UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/416591-some-thoughts-on-the-comparative-progression-analysis-method-for-investigating-inter-subject-comparability.pdf/

Benton, T. & Sutch, T. (2013). *Exploring the value of GCSE prediction matrices based upon attainment at Key Stage 2*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/181034-exploring-the-value-of-gcse-prediction-matrices-based-upon-attainment-at-key-stage-2.pdf.

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. Retrieved from http://www.jstatsoft.org/v48/i06/.

Coe, R. (2008). Comparability of GCSE examinations in different subjects: an application of the Rasch model. *Oxford Review of Education, 34*(5), 609–636.

Gill, T. (2017). *Candidates awarded A* and A grades at GCSE in 2015*. Statistics Report Series No. 111. Cambridge, UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/352252-candidates-awarded-a-and-a-grades-at-gcse-in-2015.pdf.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from Plausible Values? *Psychometrika, 81*(2), 274–289.

Ofqual. (2014). *Consultation on Setting the Grade Standards of new GCSEs in England*. Ofqual, Ofqual/14/5401, Coventry. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/630155/2014-04-03-consultation-on-setting-the-grade-standards-of-new-gcses-in-england.pdf.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology, 15*(2), 201–292. Retrieved from http://www.jstor.org/stable/i261624

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2–3), 114–128.

How do you solve a problem like transition? A qualitative evaluation of additional support classes at three university Biology departments

Simon Child Research Division, **Sanjana Mehta** Cambridge Assessment English, **Frances Wilson** OCR, **Irenka Suto** Research Division, and **Sally Brown** Cambridge Assessment Network

(The study was completed when the second and third authors were based in the Research Division, and the fifth author was based at OCR)

Introduction

Some university Biology departments have introduced additional support classes for students who struggle with the transition from school or college to higher education (HE). In this study, classes at three contrasting British universities were investigated. The structure and content of the classes were compared, and reasons for introducing the classes were explored. Data collection comprised linked observation and interview methods from three stakeholder perspectives: lecturer, undergraduate, and teacher. This article discusses the transitional challenges identified by the different stakeholders in relation to the recently completed reforms to General Certificate of Education (GCE) Advanced level (A level) qualifications in the Sciences.

Background

Many students experience difficulties in making the transition from school or college to university (Lowe & Cook, 2003; Pampaka, Williams, & Hutcheson, 2012) and lecturers have frequently expressed dissatisfaction with the skills and knowledge that new undergraduates possess when they first enter university following their A levels (Mehta, Suto, & Brown, 2012). In the Biosciences, report-writing and mathematical abilities have been identified as weak (H. Jones, 2011; Suto, 2012). Skills in practical Science, including the use of equipment and data analysis, have also raised concern (J. Wilson, 2008). Poor retention of basic biological concepts has been attributed to a reliance on surface learning during A level (pre-reform) and other pre-university

courses (H. Jones et al., 2015). It has been argued that elements of the assessment model of the pre-reform A levels, such as the modular approach and the assessment of practical Science, contributed to the issues identified (H. Jones, 2011). This prompted a period of qualifications reform that began in 2012, with new Science A levels being introduced for first teaching from September 2015.

Further transitional challenges relate to changes in culture, novel subject content (Conley, 2010) and unfamiliar pedagogical approaches. For example, difficulties can result from reduced contact time (McDonald & Robinson, 2014) and an increased focus on independent learning (Zimmerman, 2008). Linked to this, changes in assessment practice, including reduced formative feedback, can be problematic (Beaumont, O'Doherty, & Shannon, 2011; Race, 2009). There may be a long-term impact on students' learning approaches, engagement, and subsequent degree success (B.D. Jones, 2009; McDonald & Robinson, 2014).

As a potential remedy for the issues identified with students' transitions to university prior to the introduction of reformed A levels, some universities introduced additional support classes for first-year undergraduates. There were two main approaches: (i) 'bolt-on' study skills courses, offered as stand-alone modules (Wingate, 2006); and (ii) 'built-in' integrated modules which embed the development of transferable skills with knowledge building within a subject area (Mehta et al., 2012). In recent years there have been examples of pre-university courses offered by university departments that have aimed to ease the subsequent transition for students (H. Jones, Gaskell, Prendergast & Bavage, 2017). Suto (2012) reported that almost 60 per cent of Biology lecturers claimed their institutions offered additional support classes. Most classes focused on writing, numeracy, independent learning, and other interdisciplinary skills. However, some also covered subject-specific content knowledge, and content in related subjects such as Chemistry. McDonald and Robinson (2014) found that additional support benefits first-year undergraduates in Biology, improving both engagement and examination results.

To date, additional support classes have been analysed primarily within individual institutions (Jansen & Suhre, 2010; Knox, 2005; Laing, Robinson, & Johnston, 2005). As universities vary considerably in size, student and lecturer characteristics, and course structure, these evaluations may lack generalisability. Furthermore, classes supporting new Biology undergraduates have typically been evaluated via retention statistics (e.g., McDonald & Robinson, 2014) or student questionnaires. A key limitation of such approaches is their failure to capture the perspectives of other stakeholders. Arguably, a broader and more holistic approach is needed.

In this article, 'built-in' additional support classes at three British universities with contrasting affiliations and student intakes were investigated. Linked observation and interview methods were used to obtain multiple-stakeholder perspectives. That is, in addition to undergraduates, lecturers responsible for class delivery, and teachers with an in-depth understanding of pre-reform A level Biology (the main pre-university curriculum at the time of data collection) and its underpinning pedagogy, were invited to evaluate the classes. Unusually, the A level teachers visited the universities in person, observing the additional support classes and discussing them subsequently with the lecturers who delivered them. As the teachers could draw upon their specialist knowledge in these discussions, qualitative data with high integrity could be generated, comprising well-informed perspectives on the classes.

Analysis across universities offered the potential to triangulate transitional issues and common approaches underpinning the three courses. Four research questions were addressed:

1. How are the classes structured?
2. What is taught?
3. Why were the classes introduced?
4. How effective are the classes?

The research was conducted in the context of reforms to A level Biology (The Office of Qualifications and Examinations Regulation [Ofqual], 2015). Engaging the HE community in the redevelopment of A level qualifications was regarded as an imperative of the reform agenda, as demonstrated by the publication of the Smith Report (Smith, 2013) commissioned by Ofqual, and the founding of the A-Level Content Advisory Board (ALCAB). The Smith Report (2013) and ALCAB consulted with the HE community so that it could "… play a more active, substantial and ongoing role in A levels" (Smith, 2013, p.3). The findings we describe in this article were intended to inform those with responsibility for ensuring that future pre-university curricula would better meet the requirements of HE, thereby reducing the need for additional support classes in the future.

Materials and methods

Three Biology lecturers, each at different universities offering additional support classes (see Table 1), and three A level teachers, each from different schools, took part. All were recruited from a database of previous research participants who had stated a willingness to participate again. They were selected to cover a diversity of institutions, engendering breadth in the data generated. Each lecturer recruited two second-year undergraduates who had taken A level Biology prior to university and who had experienced the entire support provided by their respective courses during their first undergraduate year. Subsequently, each undergraduate's individual consent was obtained by the researchers. All participants' characteristics are shown in Table 2.

Table 1: University affiliations and Biology undergraduate course details

	University A	*University B*	*University C*
Affiliation	Russell Group	Former 1994 Group	No affiliation
Typical A level grade entry requirements	AAB	AAB	BBB
Undergraduate courses offered	Biomedical Sciences; Biology; Biochemical, Molecular & Macro Biology; Sports & Exercise Sciences	Biological Sciences; Biochemistry; Biomedicine; Ecology; Molecular Biology & Genetics	Biomedical Science; Biological Sciences; Ecology; Pharmaceutical & Chemical Sciences

Data was collected over three weeks. At each university, four types of data were obtained. Firstly, the lecturer was interviewed. Secondly, a paired interview was conducted with the two undergraduates. All interviews were semi-structured in the same way and were designed to elicit structural and content information, as well as views on the additional support classes and the transition between school and university.

Table 2: Study participants

	University A	*University B*	*University C*
LECTURERS			
Title	Lecturer of Human Physiology	Senior Lecturer of Biology	Principal Lecturer & Deputy Head of School
No. of years' experience	10	13	17
Teaching responsibilities	Teaches first-, second-, and third-year undergraduates	Teaches first-year undergraduates	Teaches first-, second-, and third-year undergraduates
UNDERGRADUATE 1			
Degree course	Biological Sciences	Biology	Biomedical Sciences
Length of course (years)	3	4	3
Subjects taken at A level	Biology, Mathematics, Sociology, and Physical Education (AS)[a]	Biology, Chemistry, and Mathematics	Biology, Chemistry, Mathematics, and Psychology (AS)[a]
Year of A level completion	2011	2011	2010
UNDERGRADUATE 2			
Degree course	Biological Sciences	Biology BSc (Honours)	*N/A (Only one student was interviewed at University C due to the withdrawal of the second student at late notice)*
Length of course (years)	3	3	
Subjects taken at A level	Biology, Chemistry, Mathematics, and Drama & Theatre Studies (AS)[a]	Biology, Geography, and Law	
Year of A level completion	2011	2011	
A LEVEL TEACHER			
School type	State comprehensive	Independent	State comprehensive
No. of years' experience	20+	12	4

a. Advanced Subsidiary level.

Questions on the latter focussed on:

- the usefulness of the knowledge and skills learned at A level Biology for university study;
- the effectiveness of skills developed in additional support classes; and
- future changes to A level Biology.

Thirdly, the A level teacher observed an additional support class. To structure observations and aid note-taking, a form was provided. This prompted the teacher on: (a) positive or negative aspects of the class, and (b) similarities and differences to A level, in terms of (i) content, (ii) use of facilities and resources, and (iii) teaching technique. The form provided space for further reflective comments on: content, the depth of knowledge exhibited by students, pedagogy, the transition to university, and changes in personal perceptions of A levels as a result of the observations. Additionally, each teacher received the teaching materials provided to students across the entire course of additional support classes, to enable a more complete view of its aims and content.

Fourthly, a researcher-facilitated discussion took place between the lecturer and the A level teacher. The researcher used pre-prepared prompts to stimulate the transfer of views and opinions. The prompts related to class aims, communication between schools and universities, and other potential improvements to the transition process.

Data analysis

Audio recordings of the interviews and facilitated discussions were transcribed. Following a preliminary reading of the transcripts, an initial framework was formulated. This enabled the data to be segmented, and then coded by source and theme, using MAXQDA (software for qualitative and mixed methods data analysis, VERBI, 2013). In an iterative process involving two researchers, content addressing each research question was identified and coded further using a refined framework of more nuanced codes. The coding was conducted by two researchers separately and was compared subsequently to confirm reliability. Using a linked framework containing some overlapping codes, comments from the A level teachers' observation forms were coded by hand. All data relating to each research question was then collated, and analysed qualitatively, to make comparisons and discern shared and conflicting perspectives.

Results

Class structure and content

Addressing the first two research questions, Table 3 provides an overview of how the courses were structured and what was taught.

Despite their contrasting student cohorts, similarities across the universities were found. At all three, classes took place regularly throughout the academic year, and were presented by a range of teaching staff. Coverage of particular topics was timed strategically to coincide with and therefore facilitate students' study and assignments in related areas. In all three observed classes, presenters made links with parts of the traditional undergraduate curriculum, thereby supporting the development of genuinely transferable skills. There was also considerable overlap in content. Courses at all three universities covered report-writing and data presentation, and two out of the three focused on each of: independent learning, literature searches, and data analysis and interpretation. This indicates shared concerns about the transferable skills of new undergraduates, including those with high A level grades. There were also differences in content among the courses. Classes at University A had a greater focus on assessment in the wider undergraduate course and how to get the best from it. University B included content on employability and employment options after graduation. University C had a greater focus on practical skills and scientific method. A description of the observed classes follows.

University A

The class covered scientific report-writing. The 45-minute formal element comprised 3 short presentations. Firstly, a librarian presented

Table 3: Structure and content of additional support courses

		University A	*University B*	*University C*
Year of introduction		2009	2000	2004
Format		Fortnightly sessions – 45-minute formal presentation then small group discussions with peer support mentors. Portfolio of internet resources and 'top tips' for each topic compiled by the course director	Weekly alternating 1-hour lectures and seminars	Weekly alternating lectures and laboratory work, lasting 2–3 hours
Duration (years)		1	1	1
Attendance		Optional	Compulsory	Compulsory
Assessment		No assessment	One coursework essay; a poster presentation; synoptic exam which includes an essay on 1 of 10 options	Examination (50%) on the principles of laboratory techniques. Coursework (50%) comprised learning exercises with self-assessments of study skills, scientific communication and laboratory skills. Students also completed practical assessment and report-writing tasks
No. of students in the observed class		64	100	150
Class teachers		Individual topics delivered by members of the department and the wider university. Four second-year undergraduate students employed as peer support mentors	Individual topics delivered by different lecturers from the department. Additional contributions from non-academic staff from across the university	Individual topics delivered by different lecturers from the department
Content	*Independent learning, including self-monitoring and goal-setting*	Yes	—	Yes
	Teamwork	—	Yes	—
	Literature searches	—	Yes	Yes
	Laboratory skills	—	—	Yes
	Scientific method and measurement	—	—	Yes
	Data analysis and interpretation	—	Yes	Yes
	Data presentation	Yes	Yes	Yes
	Report-writing, including structuring arguments	Yes	Yes	Yes
	Referencing skills	Yes	—	—
	Presentation skills	—	Yes	—
	Understanding marking and assessment criteria	Yes	—	—
	Using feedback	Yes	—	—
	Employability and employment options	—	Yes	—

information on referencing styles and systems. The topic was introduced with a short quiz to engage students. The librarian then explained the importance of correct referencing, differences between references and citations, and referencing software. Secondly, a lecturer presented the topic of scientific report-writing, explaining the need for different sections (e.g., introduction, materials and methods). Students were shown examples of well- and poorly-drawn figures, and the formality of the language needed in reports was emphasised. A third lecturer explained the qualitative criteria used to assess practical reports. The criteria corresponded to different grade bands, with a description of the standards expected at each band in relation to different sections of the report.

University B

The class was delivered primarily as a lecture, and covered data analysis and presentation. Experimental data was used to describe frequency

histograms. Mean values were calculated, and the interpretation of frequency histograms was then explained. Further explanation of error bars, standard deviation, standard error and confidence intervals was also provided. The lecturer also explained when to use different types of graphs and encouraged students to look at graphs in published papers. Towards the end, the students were asked to complete a task about labelling graphs and writing legends.

University C

The class comprised two parts. The first part focussed on safe laboratory practices and writing laboratory reports. A quiz was used to explore scientific attitudes and the behaviours and discipline needed for laboratory work. The students reviewed laboratory photographs and discussed their observations concerning fire hazards, contamination, obstructions, and radiation. A presentation on laboratory reports included points on writing long and short reports with an explanation of the standard structure used. The lecturer also covered the presentation of graphs and tables. Students were shown examples of well- and poorly-constructed graphs, and graphs from published scientific papers. The style of language used in reports was explained.

In the second part, feedback was given on a scientific calculations exercise that all students were expected to have completed in advance. The lecturer adopted a guided problem-solving teaching technique, solving equations on the board and encouraging questions from students related to intermediate steps. The lecturer linked these calculations to students' practical work.

Reasons for introducing the classes

When interviewed, the lecturers were asked why their departments had introduced additional support classes. All three offered reasons which related ultimately to improving report-writing, practical skills, and mathematical or statistical skills, which are known contributors to transitional problems (H. Jones, 2011; Suto, 2012; J. Wilson, 2008). Additionally, and in line with needs identified by Beaumont, O'Doherty, & Shannon (2011), the classes at University B had been introduced to reorient students towards unfamiliar types of assessment. Where made during other stages of data collection, comments on these themes from the students and A level teachers concurred with those of the lecturers.

Report-writing, practical and statistical skills

According to the lecturer from University B, the classes were introduced to understand new undergraduates' levels of preparedness in Mathematics and other key areas, in order to plan further learning. The classes were also intended to highlight to students the importance of certain skills by teaching them in a biological context. All three lecturers indicated that a challenge for classes was to contextualise students' understanding of Mathematics and Statistics in relation to Biology in a manner that students "don't realise often that they are doing Maths". One lecturer believed that Statistics at A level was taught "in a very dry format" and, as a result, students fail to make links between Statistics and Biology.

The lecturer from University C explained that classes were introduced in response to perceived weaknesses of new undergraduates in the practical elements of data collection and analysis, including health and safety in the laboratory, the use of specific equipment such as centrifuges, and statistical analysis. This explanation was supported by the focus of the observed class at University C on practical work, and by the interviewed student at University C, who commented:

> *I had never seen them [gills and pipettes] in my life. But, as soon as we came here, it's like a weekly thing now. You are constantly working with gills and pipettes. It's just simple things like that… You do a lab, at least once a week, sometimes two to three… If you had a bit of a better foundation [at A level], you would be more confident when you are in the laboratory… And like laboratory books… you have to maintain your laboratory book and keep it up-to-date… All the basic skills that you could have easily picked up at Biology A level just by having a book... We didn't have that.* (Undergraduate, University C)

Similarly, a student at University B attributed difficulties in scientific design, practical work and report-writing, to inexperience at A level:

> *[At A level] you would never plan an experiment, or do an experiment and write it up, so you wouldn't gain the writing skills. You get to university and they say, "write a scientific report", and you have never written one before in your life.* (Undergraduate 1, University B)

Concurring, another student commented that they had "missed out on constructing an argument and writing scientifically" in A level Biology. The lecturer at University A also linked under-preparedness in report-writing to her perception of the main pre-university curricula:

> *… Some of these students, if they do Science A levels, haven't written a full sentence since GCSEs[1]. In fact, one student told me today he managed to do GCSEs without writing many sentences! Those [skills] are so fundamental and what they don't understand is that biologists are judged by how they communicate through their writing… There is a real lack of understanding of what it is to study Biology.*
> (Lecturer, University A)

This perception was corroborated by an A level teacher, who observed the class on scientific report-writing and referencing skills at University A. In her reflective comments, she stated that this content was not covered at A level Biology as there is no requirement for report-writing. Subsequently, they explained to the lecturer that although the presentation of graphs and bar charts, and data analysis, are covered at A level Biology, this is only in a piecemeal way:

> *Under the present A level specification only very few skills are covered… and these tend to be a 'bitty' and not in the context of a complete investigation …*
> (A level teacher during discussion with Lecturer, University A)

The A level teacher who observed the class on data analysis and presentation at University B indicated that A level students are taught graph drawing, including the plotting of error bars and the calculation of standard deviation, but were not taught how to write detailed legends, which formed an important part of the class she observed. She also suggested that A level Biology students were not given sufficient practice in evaluating results. Moreover, the A level teacher who observed the class on scientific calculations at University C thought the calculations were more difficult compared to those in A level Biology, because they required prior knowledge of moles and molarity.

Differences in assessment approaches

The lecturer at University A explained that the additional support classes had originally been designed to support new undergraduates with BTEC

1. General Certificate of Secondary Education, usually taken at age 16.

qualifications, as part of an opening access agenda. It was felt that the assessments experienced by these students differed substantially from university assessments, and the classes were needed to reorient students. Level 3 Technical qualifications (such as BTECs and Cambridge Technicals) were perceived to differ in terms of their assessment model because they were criterion-referenced (based on meeting of specified learning outcomes) and internally assessed (Wolf, 2011)[2]. Level 3 Technical qualifications also contain a variety of optional units. This can mean that students arrive at university with a range of assessment 'routes' through this qualification type. After a pilot year, however, the initiative was opened up to all undergraduates in the department on a non-compulsory basis and had steadily increased in popularity.

The A level teachers and the undergraduates also perceived important differences in assessment styles between A level and undergraduate courses, and in their washback on learning. According to the teachers, their students tended to be driven completely by exam preparation and were unable to appreciate that learning at A level links to the next stage. Concurring with H. Jones et al. (2015), the undergraduates indicated that this resulted in surface learning at A level. For example:

> *[At university] it just becomes more obvious that you have to go away and do your own reading, and you have to figure out how to learn, whereas before, you could get away with just reading through and, like, just 'blagging' it.* (Undergraduate 2, University A)

Class efficacy

Within the broad theme of class efficacy, the interview and observation data supported three main strands of evaluation. That is, the participants reviewed the classes in terms of: (i) the pedagogical approaches used; (ii) the transferable skills developed; and (iii) the overlap and gaps with A level curricula.

Pedagogy

All three teachers indicated that teaching approaches used in the classes differed markedly from those at A level. The observed classes generally comprised formal presentations with limited interaction between the presenter and students, relative to that in a school classroom. Each teacher noted that each class covered multiple topics, which were delivered rapidly and were therefore challenging for students to adapt to. One teacher wrote:

> *From an Ofsted [schools inspection authority] point of view, schools are now expected to include many teaching and learning styles within the hour, whereas, at university, to sit and listen, and make notes for half an hour, is considered the norm. Therefore, there will be some students who will find it difficult to sit and listen and make notes.* (A level teacher observing class at University B)

The teachers most valued the parts of the classes that comprised more interactive teaching, such as quizzes and group discussions. However, they felt there was greater scope for further questioning and interaction, which would help to build students' confidence to ask questions in a large lecture room. During discussions with the lecturers there was some recognition, however, that the lecturers had made the classes less interactive due to time constraints.

Three further aspects of pedagogy were valued highly. First, the teacher visiting University A thought the presentation of objectives at the outset of the class was an effective teaching strategy, as it facilitated understanding of the session's relevance. Moreover, she suggested that A level classes should also start with such clear objectives. Secondly, the teacher visiting University C regarded the lecturer's approach of making explicit links between the calculation questions and students' practical work requirements a positive and useful teaching strategy.

Thirdly, the teacher visiting University A appreciated the use of second-year undergraduates as mentors to first years. She reported that similar initiatives were being implemented in many schools, where final-year A level students acted as mentors to younger students. The undergraduates at University A also appreciated the peer support mentors in their first year, and the amount of contact time provided with them.

Transferable skills

The undergraduates shared many positive experiences of the support they had received through the classes. All felt they had improved specific skills which they could apply to other modules in the first year, then continue to use in their second year. These included: essay and report-writing, reading journals, study skills, statistics, and data presentation. In their paired interview, the undergraduates at University A commented:

> *Yes, it is more of a development step, rather than being like the be all and the end all.* (Undergraduate 1, University A)

> *Sometimes you just need a few pointers to then develop your own way of doing things.* (Undergraduate 2, University A)

> *The scientific reports we have got this year [in the second year] are obviously the same sort of thing. The questions are different and the way they are laid out is different, but obviously the basic skills that you have got about explaining your results and things like that, you can carry through.* (Undergraduate 1, University A)

These two undergraduates felt that the largely informal, formative nature of assessment in classes at University A aided the development of these skills.

At University B, the undergraduates believed the classes had facilitated their understanding of the demands and expectations of university study as well as developing their skills. For example, in relation to support with scientific report-writing, one commented:

> *We went through it [report-writing] in a lot of depth, started slowly and built up from there… We started off with small tasks in groups and they gave us a lot of support for the first report we wrote… where to start off, what subjects to go into, what to read up on. At the start, it was even where to read up.* (Undergraduate 1, University B)

The teacher who visited University A reported that students who study A level Biology are not taught referencing skills or the formulation of hypotheses, and do not develop sufficient experience in using statistical methods. She therefore considered the additional support classes essential for developing these transferable skills. The A level teacher who observed the class at University B (on data analysis and presentation) concurred.

2. External assessment has since been introduced for new versions of Level 3 Technical qualifications.

Overlap and gaps with A level content

When evaluating the subject content of additional support classes and the wider Biology curricula, the participants held mixed views on the extent of overlap and gaps between A level and the first year of university. This may have reflected differences in the curricula of different institutions and examination boards. Overlap, where it arose, was not generally viewed as problematic. For example, several students identified considerable overlap in topics such as metabolism and genetics but were not dissatisfied. They explained that in general, the Biology content at university was more advanced and detailed compared to at A level, but the difference was manageable:

> *I think at A level there are snippets of each bit, whereas when you get up here to university, they explain it in more context and link it all together.* (Undergraduate 1, University B)

Another student thought that although the content taught at A level was appropriate, students were misled into believing they had covered topics comprehensively. She implied it would be better for A level students to be taught that their curriculum was part of a bigger picture:

> *It wasn't just that they [at A level] didn't go into nearly as much detail, which is what you would expect, but because they oversimplified it. You kind of thought this was the whole picture. So, when you come here [to university] and they told you that there is this, this, this, this, this, it didn't really help knowing the previous knowledge because it had loads of gaps in it and it didn't really make sense as a whole anymore.* (Undergraduate 2, University A)

The lecturers felt they had a good understanding of the subject knowledge students were likely to have upon entering university. Areas of perceived gaps in knowledge included Physiology, Cell Biology, and Evolution. The lecturers endeavoured to sculpt classes accordingly. For example, the lecturer at University B explained that her additional support course included a combination of several basic concepts in relation to some topics, and sophisticated concepts in relation to other topics:

> *For these topics [Plant Biology]… we can't teach detail. We have to teach them basic concepts because they don't have them there, whereas, with the molecular and genetic stuff, it is obvious that they're coming in with much better knowledge now. Whatever they are doing in schools takes two or three years to feed in.* (Lecturer, University B)

Similarly, when comparing the content of her classes with A level content, the lecturer at University A explained:

> *I am not saying we repeat A level content. We do it in a different way. We do it related to what they need to know as an undergraduate in that particular programme.* (Lecturer, University A)

Discussion

This research identifies important similarities across contrasting universities in how they address the transitional gaps between school or college and HE. Additional support classes were introduced to target a particular subset of skills related to scientific investigation that university lecturers had prioritised. Areas of perceived weakness included the component elements that contribute to an effective research report including initial data collection (practical skills), analysis, and the conventions of academic writing.

Although this article describes findings from only a small sample of universities, the method afforded the opportunity for comparisons to be made across the transitional divide between A level and undergraduate study. The second-year undergraduates interviewed had reflected effectively on their experiences when beginning university, whilst the A level teachers and lecturers were able to discuss areas of overlap with respect to their pedagogical approaches and content coverage. This innovative approach meant that new insights and the triangulation of views were possible.

The research contributed to the evidence base that determined what was required to improve the transition to university for first-year undergraduates. The reform agenda in England and Wales was underpinned by a 'design down' method (Baber, Castro, & Bragg, 2010; Conley, 2010; Smith, 2013), based on the principle that the needs of higher levels of education dictate the format, structure, and content of assignments at the lower level. An important outcome of the research for qualifications reform in scientific subjects was a renewed consideration of how students could obtain a more well-rounded experience of practical Science that more closely resembled university study, whilst simultaneously meeting the assessment obligations underpinning the delivery of large-scale general qualifications (Abrahams & Reiss, 2015).

The pre-reform assessment model at A level assessed practical Science through externally set but internally marked controlled assessments. The issues that were identified with practical skills informed the development of a new 'endorsement' assessment model for practical Science in the reformed A level Science qualifications. This reformed approach to practical Science was piloted in late 2014 (Inter-Board Working Group, 2014), before becoming part of the specification for all A level Science qualifications from September 2015. It comprises observations of a student's practical skills conducted by the teacher (called the *practical endorsement*), and a written examination element (Evans & Wade, 2015; Wade & Abrahams, 2015). For the practical endorsement, students receive either a pass or a fail grade which is based on Common Practical Assessment Criteria (CPAC) that the teacher applies in their observations of an individual student's practical activities. The practical activities targeted are defined by the specification, for example, OCR's A levels in Science subjects define 12 practical 'groups', with each containing 3 potential practical activities. It is intended that schools choose a minimum of 12 activities that cover the required range of skills and techniques contained in the specification (Evans & Wade, 2015). Students also maintain a record of their activities in a log book. To supplement the practical endorsement, a minimum of 15 per cent of the written examination marks must be related to the 12 practical activities covered as part of the course.

It is argued that this approach to practical Science assessment rewards both procedural skills (e.g., a student's ability in using materials and equipment) and process skills (e.g., conceptual understanding, making predictions and communication) through assessment (Abrahams & Reiss, 2015). It was intended that the new approach to practical Science assessment would encourage a broader range of practical activity in schools through the practical endorsement, whilst also assessing aspects of understanding through the written examination component (Evans & Wade, 2015). Others were critical of the practical endorsement approach for potentially devaluing practical Science,

because the pass/fail grading does not contribute to the overall grade for the A level (Biology Education Research Group, 2014). An initial evaluation of the practical endorsement assessment model from the perspective of the teachers who delivered the course has been conducted by Cadwallader and Clinkemaillie (2017) at Ofqual. The teachers interviewed in the study stated that the practical endorsement approach had increased the amount of practical work undertaken by students. The teachers explained that the new arrangements required students to take a more 'hands on' approach and there was an element of repetition of practical tasks that improved students' skills with equipment and procedures. This finding suggests that one of the issues we have raised in this article, that first-year undergraduates were not well prepared in using laboratory techniques, is effectively targeted in the reformed A level Biology.

The reformed A levels also have an increased emphasis on mathematical understanding. One of the issues raised in our research was that statistical methods and presentation were only studied in a 'piecemeal' way in the pre-reform A level Biology. In the reformed A level, however, mathematical content is intended to be covered within full practical investigations and embedded throughout the syllabus content. For example, OCR (2015) has mapped mathematical techniques and understanding that will be demonstrated across different sections of the syllabus content for A level Biology. Cadwallader and Clinkemaillie (2017) found that whilst students were covering more mathematical content in the reformed A levels, this introduced difficulties in finding sufficient time to complete scientific investigations. It is not clear how teachers are reconciling this tension in their pedagogical approaches and how this might affect the skills that students acquire before university study in Science subjects.

Cadwallader and Clinkemaillie (2017) also reported that teachers perceived that the reforms will improve the transition to university. The first cohort of students that were taught the reformed qualifications are, at the time of writing, in their first year of undergraduate study. It remains an open question whether the transitional challenges observed in our study have been resolved, and to what extent any observed improvements are due to the reformed qualifications. It is also important to acknowledge, that even if the reforms have achieved closer alignment between the knowledge and skills acquired at A level and those required for first-year undergraduate study, that there are also other transitional challenges that students must negotiate. Amongst other things, students have to embed themselves in university culture and adapt to a greater range of assessment methods (Beaumont et al., 2011; F. Wilson, Child & Suto, 2013). In our study, the pedagogical methods in the university classes were noted to be markedly different to how A level teachers would approach teaching similar (but more advanced) content. The students themselves noted that understanding the expectations of university study was an important outcome of the additional support classes. Students' emerging awareness of academic conventions related to report-writing, statistics and practical work can be applied in their first summative assessment attempts (Conley, 2010), which for Biology courses typically take place at the end of the first semester (Child, F. Wilson, & Suto, 2013; F. Wilson et al., 2013). Assessment of learning in the additional support modules is typically simultaneous with first assessment attempts in other modules (Child et al., 2013; F. Wilson et al., 2013). This suggests that there is mutual application of knowledge and understanding from additional support classes to the course as a whole. In the subsequent semesters, students are also able to use the feedback they receive to guide their later assessment attempts (H. Jones, 2013).

Finally, A levels are not just designed for students applying to university. One focus of the practical endorsement approach is improving students' abilities in using technical equipment. This might have implications for students who are intending to move into employment or onto Further Education in vocational areas. The focus of this article was on one specific section of the overall cohort: students who attend university to study a Science-based subject. It is a question for future research to understand the impact of reforms of general qualifications for students moving onto other educational or employment destinations.

Acknowledgements

We are grateful to Sylvia Green, formerly of the Research Division, and Gill Elliott, Research Division, for their helpful advice on an earlier draft of this article. We would also like to thank Magda Werno, formerly of the Research Division, for her administrative assistance, and Fiona Beedle, Assessment Research and Development, for her help with the literature search for this research. Finally, we are especially grateful to the university lecturers, A level teachers, and university students for their engagement with the study.

References

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

Baber, L. D., Castro, E. L., & Bragg, D. D. (2010). *Measuring success: David Conley's college readiness framework and the Illinois College and Career Readiness Act*. Champaign, IL: Office of Community College Research and Leadership.

Beaumont, C., O'Doherty, M., & Shannon, L. (2011). Reconceptualising assessment feedback: A key to improving student learning? *Studies in Higher Education, 36*(6), 671–687.

Biology Education Research Group. (2014). How important is the assessment of practical work? An opinion piece on the new Biology A-level from BERG. *Journal of Biological Education, 48*(4), 176–178.

Cadwallader, S., & Clinckemaille, L. (2017). *The impact of qualification reform on A level science practical work. Paper 1: Teacher perspectives after one year*. Retrieved 23 January 2018 from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/633574/Research_report_on_A_level_science_practicals_27.07.2017.pdf

Child, S. F. J., Wilson, F., & Suto, I. (2013). *"I've never done one of these before." A comparison of the assessment 'diet' at A level and the first year of university*. Paper presented at the 4th Assessment in Higher Education conference, Birmingham, UK.

Conley, D. T. (2010). *College and career ready: Helping all students succeed beyond high school*. San Francisco: Jossey-Bass.

Evans, S., & Wade, N. (2015). Endorsing the practical endorsement? OCR's approach to practical assessment in science A-levels. *School Science Review, 96*(357), 59–68.

Inter-board Working Group for A Level Science Practicals. (2014). *Summary of Cross-Board Trialling of the A Level Science Practical Endorsement*. Retrieved 12 March 2018 from: http://www.ocr.org.uk/Images/202103-summary-of-cross-board-trialling-of-the-a-level-science-practical-endorsement.pdf

Jansen, E. P. W. A., & Suhre, C. J. M. (2010). The effect of secondary school study skills preparation on first-year university achievement. *Educational Studies, 36*(5), 569–580.

Jones, B. D. (2009). Motivating students to engage in learning: The MUSIC model of academic motivation. *International Journal of Teaching and Learning in Higher Education, 21*(2), 272–285.

Jones, H. (2011). Are our students prepared for university? *Bioscience Education, 18*(2), 1–12.

Jones, H., Gaskell, E. H., Prendergast, J. R., & Bavage, A. D. (2017). Unexpected benefits of pre-university skills training for A-level students. *Educational Studies, 43*(1), 67–70.

Jones, H., Hoppitt, L., James. H., Prendergast, J. Rutherford, S., Yeoman, K., & Young, M. (2013). Exploring students' initial reactions to feedback they receive on coursework. *Bioscience Education, 20*(1), 3–21.

Jones, H., Black, B., Green, J., Langton, P., Rutherford, S., Scott, J., & Brown, S. (2015). Indications of knowledge retention in the transition to higher education. *Journal of Biological Education,49*(3), 261–273.

Knox, H. (2005). Making the transition from further to higher education: The impact of a preparatory module on retention, progression and performance. *Journal of Further and Higher Education, 29*(2), 103–110.

Laing, C., Robinson, A., & Johnston, V. (2005). Managing the transition into higher education. *Active Learning in Higher Education, 6*(3), 243–255.

Lowe, H., & Cook, A. (2003). Mind the gap: Are students prepared for higher education? *Journal of Further and Higher Education, 27*(1), 53–76.

OCR. (2015). *A Level Specification: Biology A.* Retrieved from: http://www.ocr.org.uk/Images/171736-specification-accredited-a-level-gce-Biology-a-h420.pdf

McDonald, J. E., & Robinson, R. L. (2014). Enhancing first year undergraduate student engagement via the School of Biological Sciences tutorials module. *Bioscience Education, 22*(1), 54–69.

Mehta, S., Suto, I., & Brown, S. (2012). *How effective are curricula for 16 to 19 year olds as preparation for university? A qualitative investigation of lecturers' views*. Retrieved from http://www.cambridgeassessment.org.uk/Images/116015-cambridge-assessment-he-research-qualitative-investigation-executive-summary.pdf

Ofqual. (2015). *Get the facts: GCSE and A level reform*. Retrieved 25 January, 2018 from https://www.gov.uk/government/publications/get-the-facts-gcse-and-a-level-reform.

Pampaka, M., Williams, J., & Hutcheson, G. (2012). Measuring students' transition into university and its association with learning outcomes. *British Educational Research Journal, 38*(6), 1041–1071.

Race, P. (2009). Assessment. In The Higher Education Academy (Ed.), *Assessment: Resources, references and tools for assessment in the biosciences* (pp.1–4). Leeds: The Higher Education Academy UK Centre for Bioscience.

Smith, M. E. (2013). *Independent chair's report on the review of current GCE 'specification content' within subject criteria: A report to Ofqual*. Retrieved 12 February 2018, from http://ofqual.gov.uk/qualifications-and-assessments/qualification-reform/a-level-reform/

Suto, I. (2012). *How well prepared are new undergraduates for university study? An investigation of lecturers' perceptions and experiences*. Paper presented at the annual conference of the Society for Research into Higher Education, Newport, Wales, UK.

Wade, N., & Abrahams, I. (2015). *Validity issues in the reform of a practical science assessment: An English case study*. Paper presented at the conference of the International Association for Educational Assessment, University of Kansas, USA.

Wilson, J. et al. (2008). *1st year practicals: Their role in developing future bioscientists*. Retrieved 22 January 2018 from https://synergy.st-andrews.ac.uk/vannesmithlab/files/2015/08/Adams_et_al08CentreBioReport.pdf

Wilson, F., Child, S. F. J., & Suto, I. (2013). *Aspiring to bridge the gap between A level and HE: A study of assessment and additional support lessons*. Paper presented at the conference of the Society for Research in Higher Education, Newport, Wales.

Wingate, U. (2006). Doing away with 'study skills'. *Teaching in Higher Education, 11*(4), 457–469.

Wolf, A. (2011). *Review of Vocational Education – The Wolf Report*. (DfE-00031-2011). London: Department for Education.

VERBI. (2013). MAXQDA [Computer software]. Retrieved from http://www.maxqda.com/products

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183.

Research News

Karen Barden Research Division

Conference presentations

British Educational Research Association (BERA)

The University of Sussex, Brighton, hosted the BERA Annual Conference in September 2017. This provided an opportunity to develop new ideas, and to build new relationships within the research education community. Several researchers from the Research Division of Cambridge Assessment attended the conference and the following papers were presented:

Ellie Darlington: *What is a non-specialist teacher?*

Gill Elliott: *Aspects of Writing: challenges and benefits of longitudinal research.*

Martin Johnson: *What is effective feedback in a professional learning context? A study of how examination markers feed back to each other on their marking performance.*

Carmen Vidal Rodeiro and Joanna Williamson: *Education and employment destinations of students in England: the value of 14-19 qualification.*

Nicky Rushton: *Spelling errors in 16-year-olds' writing.*

Sylvia Vitello and Cara Crawford: *Foundation or Higher tier? Effects of moving from a modular to linear system of GCSE assessment.*

International Association for Educational Assessment (IAEA)

The 43rd Annual Conference of the IAEA took place in Batumi, Georgia, in October 2017 with the theme of *Assessment as a Social Lever.* Stuart Shaw, Cambridge Assessment International Education, presented a paper co-authored with Research Division colleagues Carmen Vidal Rodeiro and Cara Crawford on *Predicting the success of the Cambridge Advanced International Certificate of Education (AICE) Diploma in the United States.*

Stuart Shaw also presented papers on *The construction of a validity portfolio for general educational qualifications: a suggested approach to large-scale validation; An Exploration of the Nature and Assessment of Student Reflection*, co-authored with his Cambridge Assessment International Education colleague, Martina Kuvalja; and a post-conference workshop entitled *Issues around how best to provide evidence for assessment validity, reliability and fairness: the practice and challenge of validation.*

Association for Educational Assessment-Europe (AEA-Europe)

Held in November 2017, the 18th AEA-Europe Annual Conference took place in Prague, Czech Republic, under the theme of *Assessment Cultures in a Globalised World*. Several researchers from Cambridge Assessment attended the conference and the following papers were presented:

Tom Benton, Research Division: *Pooling the totality of our data resources to maintain standards in the face of changing cohorts.*

Tom Bramley and Tom Benton, Research Division: *Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests.*

Lucy Chambers, Filio Constantinou, Nadir Zanini and Nicole Klir, Research Division: *Alternative uses of examination data: the case of English language writing.*

Gill Elliott and Nicky Rushton, Research Division: *Popular perceptions about the comparability of assessments in England. A tension between academia and the mainstream broadcast and print media?*

Sarah Hughes, Cambridge Assessment International Education: *Developing a culture of research-informed practice by encouraging research uptake in an assessment organisation.*

Martin Johnson and Nicky Rushton, Research Division: *A culture of question writing: How do question writers compose examination questions in an examination paper?*

Martina Kuvalja, Stuart Shaw and Sarah Hughes, Cambridge Assessment International Education: *Cambridge progression: Teachers' perspectives.*

Tim Oates, Assessment Research and Development: *Should there be a single assessment culture in a globalised world?*

Stuart Shaw and Martina Kuvalja, Cambridge Assessment International Education: *An exploration of the nature and assessment of student reflection.*

Sylvia Vitello and Tom Bramley, Research Division: *The effect of adaptivity on the reliability coefficient in Comparative Judgement.*

Frances Wilson, Neil Wade, OCR, Stuart Shaw, Sarah Hughes and Sarah Mattey, Cambridge Assessment International Education: *Evaluating written assessments of practical work – a taxonomy.*

The following poster was also presented:

Gill Elliott, Nicky Rushton and Jo Ireland: *Is the General Certificate of Secondary Education (GCSE) in England incongruous in the light of other jurisdictions' approaches to assessment?*

MAXQDA International Conference

The MAXQDA International Conference took place in March 2018 in Berlin, Germany, bringing together international researchers who work and teach with MAXQDA. The event centred on questions on how to optimize the use of MAXQDA in the various methodological and thematical settings of qualitative and mixed methods research. Irenka Suto, Research Division, gave a poster presentation on *How do you solve a problem like the transition to university? The use of MAXQDA in a qualitative evaluation of additional support classes for undergraduate biologists.*

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

The Cambridge Approach to Improving Education

Tim Oates, CBE, Group Director of Assessment Research and Development at Cambridge Assessment, led the research into and wrote *A Cambridge Approach to Improving Education. Using international insights to manage complexity*[1] in which he sets out his findings and guiding principles for policymakers. It is part of a wide-ranging study of educational improvement across a range of jurisdictions and follows on from *The Cambridge Approach to Textbooks published in 2016*[2].

The launch in September 2017 was hosted by the UK think tank Policy Exchange. It included presentations by experts in the field including Dr John Jerrim, Reader in Educational and Social Statistics at the UCL Institute of Education, and John Blake, Head of Education and Social Reform, Policy Exchange.

The presentations were accompanied by discussion and debate with the attending education experts. This covered areas including whether smaller jurisdictions do better in international comparisons, how to explain the recent success of London, and how governments should respond to the challenges created by the digital revolution.

Further details and related materials, including a free download of the document, can be found on our website: http://www.cambridgeassessment.org.uk/news/cambridge-approach-to-improving-education-launched/ A recording of the event is also available on the Policy Exchange website at https://policyexchange.org.uk/pxevents/a-cambridge-approach-to-improving-education/10.1080/0305764X.2017.1337723

Questioning Questions

In November 2017, more than 100 people attended Cambridge Assessment's flagship autumn event, *Questioning Questions*. The audience heard from education experts including Daisy Christodoulou, No More Marking's Director of Education, and Professor Bill Lucas, Director of the Centre for Real-World Learning and Professor of Learning at the University of Winchester, in the debate on how assessment can be used to drive effective learning.

Further details and related materials, including videos of all of the conference presentations, can be found on our website: http://www.cambridgeassessment.org.uk/questioning-questions/

1. Oates, T. (2017). *A Cambridge Approach to Improving Education. Using international insights to manage complexity*. Cambridge, UK: Cambridge Assessment. Available online at: http://www.cambridgeassessment.org.uk/Images/cambridge-approach-to-improving-education.pdf
2. Oates, T. (2016). *The Cambridge Approach to Textbooks. Principles for designing high-quality textbooks and resource materials*. Cambridge, UK: Cambridge Assessment. Available online at: http://www.cambridgeassessment.org.uk/Images/299335-the-cambridge-approach-to-textbooks.pdf

Publications

The following articles have been published since *Research Matters Issue 24:*

Benton, T. and Bramley, T. (2017). *Some thoughts on the 'Comparative Progression Analysis' method for investigating inter-subject comparability*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at http://www.cambridgeassessment.org.uk/Images/416591-some-thoughts-on-the-comparative-progression-analysis-method-for-investigating-inter-subject-comparability.pdf

Bramley, T. and Vitello, S. (2018). The effect of adaptivity on the reliability coefficient in comparative judgement. *Assessment in Education: Principles, Policy & Practice*. Advance online publication available at http://doi.org/10.1080/0969594X.2017.1418734

Child, S. F. J. and Shaw, S. D. (2018). Towards an operational framework for establishing and assessing collaborative interactions. *Research Papers in Education*. Advance online publication available at http://doi.org/10.1080/02671522.2018.1424928

Crawford, C. and Benton, T. (2017). *Volatility happens: Understanding variation in schools' GCSE results*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at http://www.cambridgeassessment.org.uk/Images/372751-volatility-happens-understanding-variation-in-schools-gcseresults.pdf

Crisp, V. (2017). Exploring the relationship between validity and comparability in assessment. *London Review of Education, 15*(3), 523–535. Available online at https://doi.org/10.18546/LRE.15.3.13

Crisp, V., Johnson, M., and Constantinou, F. (2018). A question of quality: Conceptualisations of quality in the context of educational test questions. *Research in Education*. Advance online publication available at https://doi.org/10.1177/0034523717752203

Darlington, E. (2017). *Other jurisdictions' use of technology in Mathematics curricula*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at http://www.cambridgeassessment.org.uk/Images/426821-other-jurisdictions-use-of-technology-in-mathematics-curricula.pdf

Gill, T. (2017). *The impact of the introduction of Progress 8 on the uptake and provision of qualifications in English schools*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at http://www.cambridgeassessment.org.uk/Images/421442-the-impact-of-the-introduction-of-progress-8-on-the-uptake-and-provision-of-qualifications-in-english-schools.pdf

Johnson, S. (2017). *Design challenges for national assessment in this accountability era: A background paper commissioned by Cambridge Assessment*. Cambridge, UK: Cambridge Assessment. Available online at http://www.cambridgeassessment.org.uk/Images/428588-design-challenges-for-national-assessment-in-this-accountability-era.pdf

Shaw. S. D. and Imam, H. C. (2017). Towards a Scale of Academic Language Proficiency. Learning and Assessment: Making the Connections. *Proceedings of the Association of Language Testers in Europe (ALTE) 6th International Conference*. Cambridge English Language Assessment/ALTE/cliQ, 224–235. Available online at https://alte.wildapricot.org/resources/Documents/ALTE%202017%20Proceedings%20FINAL.pdf

Vidal Rodeiro, C.L., Crawford, C., and Shaw, S. (2017). From "AICE"-ing the Test to Earning the Degree: Enrollment and Graduation Patterns among Students with the Cambridge Advanced International

Certificate of Education (AICE) Diploma. *College and University: Educating the Modern Higher Education Administration Professional, 92*(4), 12–23. Available online at http://www4.aacrao.org/C&U/9204/12/

Vitello, S. and Crawford, C. (2018). Which tier? Effects of linear assessment and student characteristics on GCSE entry decisions. *British Educational Research Journal, 44*(1), 94–118. Available online at https://onlinelibrary.wiley.com/doi/epdf/10.1002/berj.3320

Zanini, N. and Williamson, J. (2017). *Learning aims: A preliminary exploration to monitor A/AS level reform*. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment. Available online at http://www.cambridgeassessment.org.uk/Images/360511-learning-aims-a-preliminary-exploration-to-monitor-a-aslevel-reform.pdf

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

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

Our Research Team

Cambridge Assessment is home to the largest research capability of its kind in Europe. You can now meet the people behind our leading-edge work, including some of the contributing authors to *Research Matters*, at http://www.cambridgeassessment.org.uk/our-research/our-research-team/

Statistics Reports and Data Bytes

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

The most recent additions to the series are:

- *Statistics Report Series No.114: Uptake of GCSE subjects 2016*
- *Statistics Report Series No.115: Provision of GCSE subjects 2016*
- *Statistics Report Series No.116: Uptake of GSE A level subjects 2016*
- *Statistics Report Series No.117: Provision of GSE A level subjects 2016*
- *Statistics Report Series No.118: Geographical variations in A level uptake in 2016*
- *Statistics Report Series No.119: Candidates awarded the A* grade at A level in 2016.*

Data Bytes is a series of data graphics from our Research Division, designed to bring the latest trends and research in educational assessment to a wide audience. Topics are often chosen to coincide with contemporary news or recent Cambridge Assessment research outputs.

Since *Research Matters* Issue 24, we have published the following *Data Bytes*, all of which can be found on our website at http://www.cambridgeassessment.org.uk/our-research/data-bytes/:

- December 2017:
  Geographical variations in A level subject uptake in 2016
  (Interactive and linked to *Statistics Report Series No.118*)
- January 2018:
  How much do I need to write to get top marks at GCSE?
  How much do I need to write to get top marks at A level?
- February 2018:
  Provision of reformed AS levels
- March 2018:
  Influence of KS2 National Curriculum levels on GCSE tier entry.

All details are correct at the time of going to print in May 2018

Contents / *Issue 25* / Spring 2018

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

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

1726807442

ISSN: 1755–6031