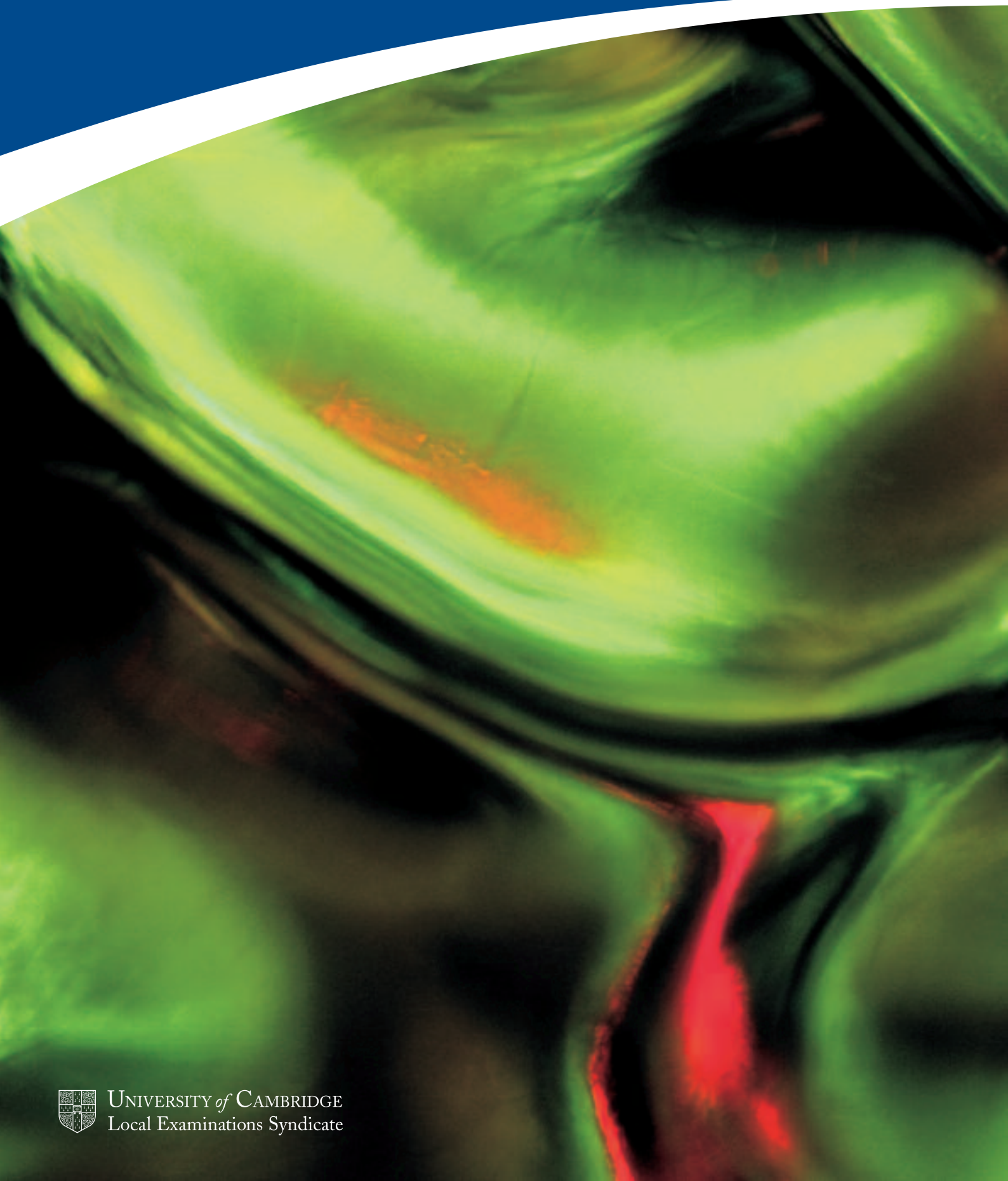
Issue 18 Summer 2014

Research Matters

CAMBRIDGE ASSESSMENT

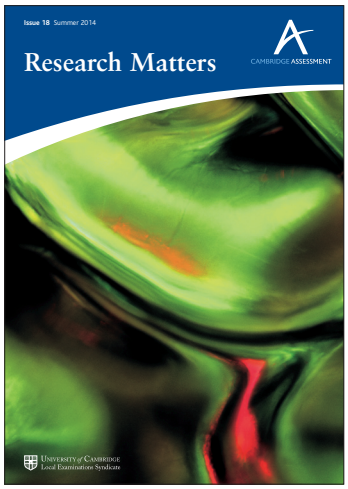


University of Cambridge
Local Examinations Syndicate

CAMBRIDGE ASSESSMENT

Citation

Articles in this publication should be cited as: Gill, T. & Suto, I (2014). Students' views and experiences of A level module re-sits. *Research Matters: A Cambridge Assessment Publication*, 18, 10–18.



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

The full issue and previous issues are available on our website:
www.cambridgeassessment.org.uk/ca/Our_Services/Research

Research Matters : 18

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

This edition of *Research Matters* engages with issues of choices and routes. Just how much diversity in qualifications is needed to maximise student engagement and to respond to societal and economic needs? Is too much choice dysfunctional? Writing in *The Harvard Crimson* in 1953, J. Anthony Lukas traced the near 100-years of oscillating expansion and rationalisation of the number of courses at the Ivy League university. Expansion occurred as the institution responded to students' and academics' interests. Rationalisation occurred when the number of courses swelled into unsustainable, sprawling incoherence. This mirrors precisely the oscillation in qualifications in England. Growth of innovative and well-evidenced qualifications was a feature of the 1990s in England, giving us Nuffield Science, SMP Maths and other important new qualifications. The final years of the Qualifications and Assessment Authority saw a dogged pursuit of root and branch rationalisation of qualifications – demonstrably pushed too far in the ill-fated Diploma development. The historical oscillation at both Harvard and in England shows that we have not yet been able to settle on a level of diversity which is appropriate to genuine needs. I consider a measure of innovation and choice in qualifications to be essential to a healthy education systems. But too much diversity gives rise to technical difficulties in comparability, bewildering choice for parents and pupils, poor signalling in the systems, and the tendency to 'closed doors' rather than 'tradable credit'. In a system which operates as a very imperfect market, with both power and information asymmetries, we cannot rely on market forces to weed out low quality, low value qualifications efficiently. Poorly-theorised rationalisation is equally dangerous to the interests of students, society and the economy. We should therefore turn towards evidence – of comparable standards, of need, of student engagement, and of progression and return. Research on these matters is fundamental to a healthy system, and not simply a 'nice to have'.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

Most of the articles in this issue relate to the issue of choice. Teachers and students must decide which qualifications to choose, when to take the assessment, and sometimes which topics to study. Such choice may have both intended and unintended consequences. Child, Darlington and Gill explore the choices of units and topics made by teachers and students in A level History, examining which topics are most commonly taught and why. They discuss the implications of these choices for the breadth and depth of historical knowledge gained by A level History students. The complexity resulting from the availability of multiple routes through a qualification creates a challenge for awarding bodies, because they must ensure the comparability of qualifications. Benton investigates a method for calculating the reliability of complex qualifications, finding that reliability remains high for A level Maths, despite a number of different options.

Two articles in this issue explore the perceived problems associated with choice relating to the timing of assessment. Gill and Suto investigate why students choose to re-sit modules at A level, and the likely impact that the reduction in opportunities for re-sitting will have on students' choices. At GCSE, some schools choose to enter students early for examinations, to "get them out of the way" and allow more time for further study in other subjects. Gill examines the impact of this decision on students' overall GCSE grades, and finds that the impact of early entry is limited, though the picture is somewhat complex.

In her article on progression from a vocational qualification, OCR Nationals, to further study or employment, Vidal Rodeiro shows that students who take OCR Nationals may progress to higher education, further study at school, or work based training, demonstrating that high quality vocational qualifications enable students to choose from a range of different progression paths. The debate surrounding the relative difficulty of A level subjects has been controversial for some time. The perception that it is more difficult to achieve high grades in some subjects may discourage students from choosing these subjects. Bramley explores several methods for examining this issue, focusing on a novel technique 'multidimensional scaling'. He concludes that while this technique has certain advantages, the results are difficult to interpret. Also focussing on comparability, Yim provides a discussion of a sophisticated method which could be used by awarding bodies to ensure comparability of qualifications across time. Looking to the future, Dhawan and Zanini's article on the use of big data opens up new possibilities for investigating how students and teachers may make decisions, leading to additional ways of understanding how and why they make the choices that they do.

Frances Wilson *Research Officer, Research Division; Guest Editor*

An analysis of the unit and topic choices made in an OCR A level History course

Simon Child, Ellie Darlington and **Tim Gill** Research Division

Introduction

In England and Wales, the primary qualification for 16–19 year-olds, the A level, is currently undergoing a period of reform. The reforms were initiated by the UK Coalition Government in 2010, with the publication of a White Paper – *The Importance of Teaching* (Department for Education, 2010). In the white paper, the Government outlined that qualifications should "match up to the best internationally in providing a good basis for [future] education and employment" (p.40), while also providing an effective accountability measure of schools and colleges in the future (Ofqual, 2013).

One of the A levels that has been identified as requiring reform is History. History is one of the most popular subjects at A level, ranking as the fifth most taken A level subject, and the sixth most taken AS level subject (Joint Council for Qualifications, 2013). According to recent research by Vidal Rodeiro and Sutch (2013), 17.1% of university applicants, and 13.5% of students overall, take A level History.

Interestingly, and problematically, there is no currently accepted body of knowledge that forms a prerequisite for the study of History at university (Hibbert, 2006). Indeed, this may explain that, while A level History is increasing in popularity (Joint Council for Qualifications, 2013), it is currently not included in the admissions criteria for undergraduate History for 9 out of the 23 Russell Group universities[1]. In the Smith (2013) review, there was little consensus reached on the fundamental History topics that should be taught at A level. Changes that were proposed in the review were limited to confirming that A level candidates should study "a range of topics from a chronological range of at least 200 years" and should "study the History of more than [one] country or state" (p.lxxxviii).

One of the issues in determining appropriate content for A level History is that the study of History can potentially serve a number of purposes. One of the key motivations for studying History is identity formation. As Harris (2013) noted:

> *Without an understanding of where we have come from, without knowledge of accepted values and practice, individuals would not know how to operate within society.* (p.408).

Harris (2013) argued that History operates for communities in much the same was as memory does for individuals, in that it facilitates more informed decision making. There is also the challenge of determining which historical topics to target, as each topic will have implications for individuals' identity formation. Students are likely to inhabit multiple identities stemming from their ethnic background, culture, language and religion (Department for Education and Skills, 2007), and it has been argued that this diversity should be acknowledged in History courses (Harris, 2013). However, political rhetoric related to History education often revolves around the creation of a sense of national identity and belonging (Harris, 2013). For example, Secretary of State for Education Michael Gove has said that, in the UK, History should focus on "our island story" (Gove, 2010). This movement towards Anglo-centrism has been criticised as potentially neglecting cultural and social History in favour of "chronological big stories" (Bowen, Bradley, Middleton, Mackillop, & Sheldon, 2012, p.126). Similarly, it has been suggested that there is too much focus in the National Curriculum and post-compulsory history qualifications on European History (Bowen *et al.*, 2012). Indeed, Tillbrook (2002) reported that, at one point, 83% of marks awarded by one examination board for A level History were for the study of only twenty years of German History.

Given the potential of History qualifications to instil knowledge on a wide variety of topics, it is perhaps unsurprising that schools are offered flexibility in the topics they cover. For example, in the current OCR A level (specification A), students can take one of 16 modular routes through the course, and a range of different topics within each module can be taught (see Table 1 for the current historical coverage of the OCR A level). Other examination boards offer fewer options in terms of unit choice, but a greater range of topic options within units. For example, the AQA AS level comprises two compulsory units. One unit has 14 topic options, while the second unit has 18 topic options. Similarly, the WJEC A2 level qualification comprises two compulsory units (one coursework and one examination). For the coursework unit, 9 topic options are offered, while 36 topic options are offered for the examination unit.

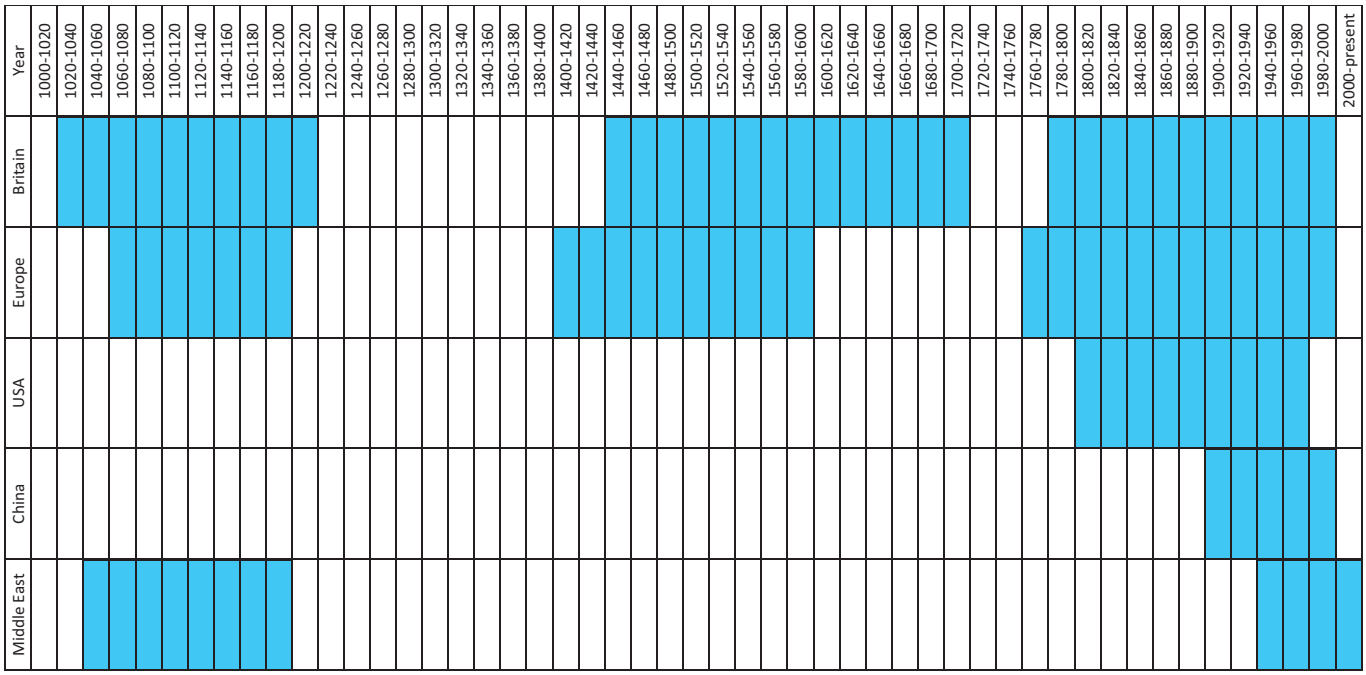
Aims of the current study

This study aimed to explore how schools that offer A level History use the options available to them, in terms of unit and topic choices. Previously it has not been possible to analyse data on the content choices that schools make in History qualifications. However, the movement to computer-based marking within part of the OCR A level, and the concomitant increase in the amount of detailed data that is automatically collected, provided an opportunity to examine the topic choices schools make, at the levels of the unit and the question.

Specifically, this study aimed to determine which units and topics were most commonly taught. It was intended that this data would help establish how optionality within A level History is used, and whether it meets the desired purpose of exposing students to a broad range of historical periods and topics. To investigate further how different schools may utilise the optionality available to them, comparisons were made between different school types (state vs independent), and schools with different levels of performance. It has been found that the uptake of A level History varies according to school type. Burn and Harris (2012) found that, in a sample of 403 centres, 31–40% of 'new' academies, 21–30% of grammar schools, 11–20% of comprehensive and independent schools, and less than 10% of 'old' academies offered

1. Based on the 'V100' standard BA History course (see ucas.com for further details).

Table 1: Current scope of OCR A level History (specification A)



AS level History. A second aim of the current study was to establish whether there were also differences qualitatively in the A level History content typically taught to students from different school types.

Method

There were two phases to this study. First, an analysis of candidates' topic choices for one AS level History unit (F961 in June 2013) was conducted. This unit was marked using Scoris, the new online marking platform for OCR examinations, which allowed data at question level to be captured and analysed. For this unit, schools have a choice of two unit options relating to broad historical periods: Option A, Medieval and Early Modern; or Option B, Modern. Within each unit option there is then a choice of six separate *topics* that may be taught. There is a separate exam paper for each unit option, with students required to answer any 2 questions from a choice of 18 (3 from each topic).

Secondly, a questionnaire was developed that asked heads of History departments about their schools' A level History unit and topic choices across the entire A level History course. This was with the intention of gathering data on the modules where online methods of marking were yet to be introduced. The method of data collection and analysis for both phases is provided below.

Database collation and analysis

The data for analysis of unit F961 was taken from a number of different sources. The information on the unit(s) offered[2] by schools and the topics and questions answered by students in the examinations was downloaded from OCR's internal databases. The unit option and topic choices were analysed by school type and by school attainment level. This data was merged with the National Centre Number database to get information on the school type. To obtain a measure of school attainment, the data was merged with the National Pupil Database to get data on the performance of all students within each school.

For the analysis by school type, schools were grouped into two categories: state (including comprehensives, academies, grammar schools, secondary modern schools and further education, tertiary and sixth form colleges) and independent schools. It was not possible to have a finer grouping of school type because of the low numbers of schools in some categories. Schools categorised as 'other' or 'unidentified' were excluded from the analysis. There were 240 centres categorised as state schools, totalling 5,676 students, and 123 centres categorised as independent schools, totalling 2,439 students.

For the analysis by school attainment level, centres were categorised into one of three groups (low, medium, and high attaining) by their mean A level score in June 2013 across all subjects and all examination boards. This was calculated by assigning a number to each A level grade (A*=6, A=5, etc.) and taking the mean of all A levels taken by all of the students at the school. There were 117 schools within each of the attainment categories. Low attaining schools had a mean A level score of 2.86, medium attaining schools had a mean A level score of 3.49, and high attaining schools had a mean A level score of 4.32.

A handful of centres were found to have ten or fewer A level results. With so few results the overall mean may not be very reliable as a measure of attainment so these centres were excluded from the analysis.

Questionnaire

Participants

Centres with candidates who took OCR A level History in June 2013 were identified using the internal database systems at Cambridge Assessment. Each centre was contacted by telephone, and asked to provide the full name and contact details for the head of the History department or equivalent. The heads of department were then emailed and invited to fill out the questionnaire, which they could access via a web link. For their time, they were offered the opportunity to enter into a prize draw.

Overall, 638 heads of department were contacted either to their direct email or to a general school email address. Ninety heads of department

2. By 'offered' we mean that at least one of the students in any one centre took an examination in that unit.

returned the questionnaire (a return rate of 14%). Overall, participants had a mean of 6.71 years of experience (*SD* = 6.21 years) as head of department at the centre where they were currently employed. The centres had spent a mean of 11.89 years teaching OCR A level History (*SD* = 6.25 years).

Eighty-five of the participants provided information about the type of school where they were teaching. Fifty-two of the centres were state schools, and 33 were independent schools. The percentage of schools in this sample that were independent (39%) is slightly higher than the overall percentage of independent schools that take OCR History (34%). However, we deemed that this sample was broadly representative of the total population of centres that offered OCR A level History in 2013.

Questionnaire development

The questionnaire was developed by members of the research team in collaboration with the OCR general qualifications reform team for History. The questionnaire comprised three sections. The first section asked participants for details of their centre and teaching experience. The second section asked about the *unit* options that centres offered to their students, and probed the reasons for these choices. The third section was similar to the second section, but asked participants about *topic* choices within units. Finally, participants were given the opportunity to add any further comments.

Piloting

Before the questionnaire was made live, a draft version was checked by the OCR subject team for History, to ensure that appropriate terminology and question response choices were included. The questionnaire was then sent to a pilot participant, who was a head of department for History. The pilot participant was asked to check the questionnaire for anything that they felt would not be understood by participants, and errors in spelling or grammar. They were also asked if there were responses that could be added to any of the questions. Once the recommended changes were made, the final version was sent to the main cohort.

Results

Analysis of candidates' unit and topic choices

Unit choice

For Unit F961, there is a choice of two options that schools can offer: F961A – Medieval and Early Modern 1035–1642; and F961B – Modern 1783–1994. Between the two school types, a similar proportion of schools offered option A (47.5% state, 46.3% independent), while the proportion of schools that offered option A was also similar between the three school attainment groups (44.4% high attaining, 47.9% medium attaining, 48.7% low attaining).

Option B was offered less often in independent schools (39.8%) compared to state schools (46.7%), and also in high attaining schools (39.3%) compared to medium attaining (46.2%) or low attaining (48.7%) schools.

Most centres offered only one of these units to their students but some schools (8.5%) offered both. This was more common in independent schools (13.8%) compared to state schools (5.8%). High attaining schools were also more likely to offer both options compared to state or lower attaining schools (16.2% high attaining, 6% medium attaining, and 2% low attaining).

Topic choice

For both unit options, in the exams students were required to answer 2 questions from a choice of 18. Each of the six topics had three questions each. Although students are allowed to mix questions from different topics, it was found that the vast majority (97.2% for option A and 98.6% for option B) answered questions from one topic only. To simplify the analysis, students who answered questions from more than one topic were removed from the data.

There were 30 centres which had some students answering questions from one topic and some from another topic, suggesting that more than one topic had been taught in the school. However, it was still the case that the vast majority of students in these schools did not mix topics in their exam papers. It is possible that these schools taught the topics to different classes.

It is therefore assumed that choice of topic is made at the school level, and students are usually taught one topic only. The following analysis looks at the choice of topic by school type and school attainment group.

Figure 1 presents the percentage of schools choosing each topic for the two options. Schools are counted twice if questions from more than one topic were answered by their students. Amongst schools, *Henry VIII to Mary I* was the most popular for option A (chosen by around 30%), followed by *Lancastrians, Yorkists and Tudors* and *England under Elizabeth I*. For option B, *From Pitt to Peel* was the most popular choice (26.6%), followed by *Domestic developments and Foreign & Imperial policies (1856–1914)*. Results were similar when the raw number of students answering questions from each topic was analysed.

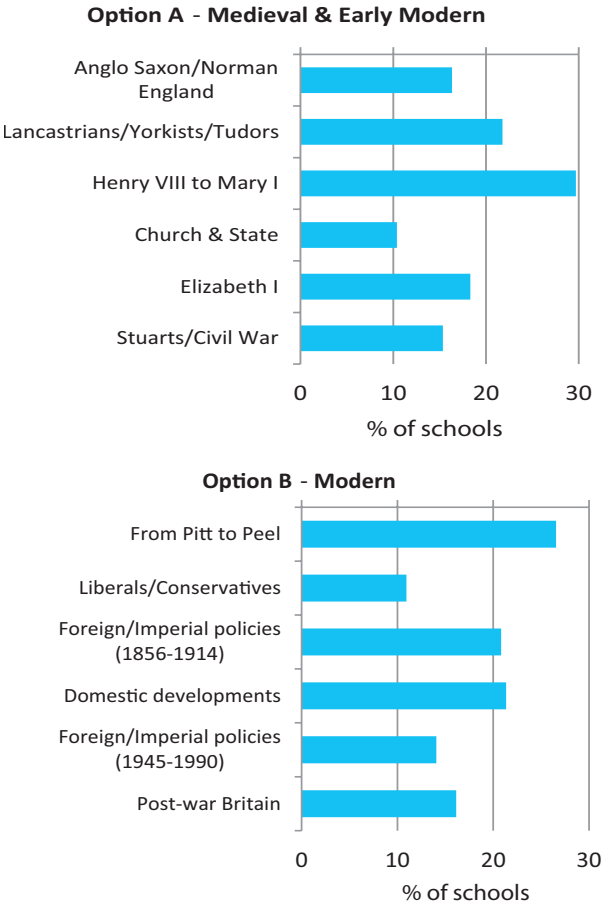


Figure 1: Percentage of schools choosing each topic

Topic choice by school type

The first stage of this analysis determined the percentage of schools choosing each topic by school type for options A and B (see Figures 2 and 3 respectively). The statistical significance of any differences between groups in topic choice was measured in two different ways. When comparing school types, an independent samples *Z*-test of differences in proportions was used. However, this method can only be used to compare two different groups, so for differences between school attainment groups a Chi-square frequency test was used.

There were no large differences in topic choice observed between school types, with both types most likely to choose *Henry VIII to Mary I*, followed by *Lancastrians, Yorkists and Tudors*. Independent schools were more likely to choose *Anglo Saxon/Norman England* and less likely to choose *England under Elizabeth I* compared to state schools. However, none of the differences in proportions choosing each topic between state and independent schools were statistically significant.

For option B there were some more substantial differences. State schools were most likely to choose *Domestic developments* or *From Pitt to Peel*, whereas independent schools were most likely to choose *Foreign & Imperial policies (1856–1914)* or *From Pitt to Peel*. Two of the differences between school types were to a statistically significant level. These were 31.8% of independent schools choosing *Foreign & Imperial policies (1856–1914)*, compared with 13.5% of state schools (the probability that this difference could have occurred by chance, p=.003) and 26.2% of state schools choosing *Domestic developments*, compared with 12.1% of independent schools (p=.024).

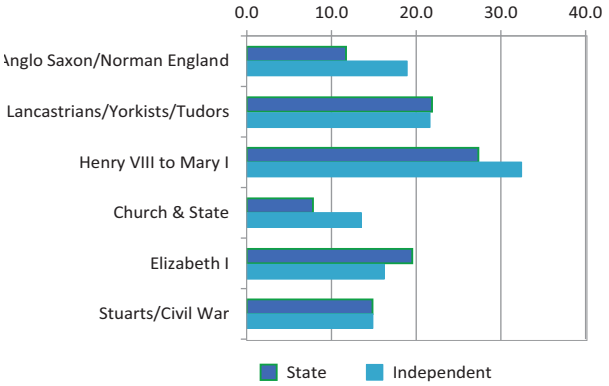


Figure 2: Percentage of schools choosing each topic by school type (option A)

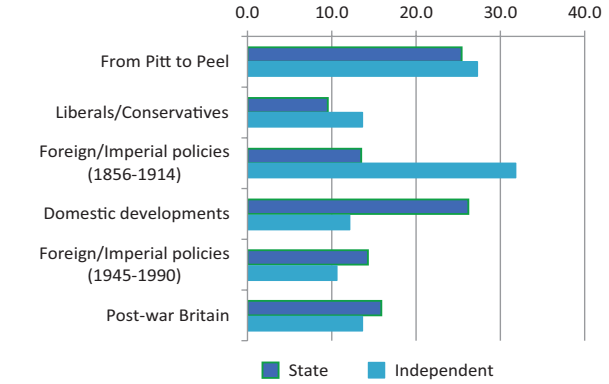


Figure 3: Percentage of schools choosing each topic by school type (option B)

Topic choice by school attainment

Figures 4 and 5 present the percentage of schools within each school attainment group choosing each topic.

High attaining schools were less likely to choose *Lancastrians, Yorkists and Tudors* than lower attaining schools. They were more likely to choose *Church and State*. Low attaining schools were less likely to choose *Henry VIII to Mary I* than higher attaining schools. However, none of these differences were statistically significant.

There were some substantial differences in option B. Low attaining schools were much less likely to choose *From Pitt to Peel* than medium or high attaining schools. They were also less likely to choose *Foreign & Imperial policies (1856–1914)* and more likely to choose *Domestic developments*. High attaining schools were much less likely to choose *Domestic developments* and more likely to choose *Foreign & Imperial policies (1856–1914)* or *Liberals and Conservatives*.

Two of these differences were statistically significant. Just 8.2% of low attaining schools chose *Foreign & Imperial policies (1856–1914)* compared to 17.7% of medium and 32.3% of high attaining schools (p<.005). In contrast, 32.8% of low attaining schools chose *Domestic developments*, compared to 21.0% of medium and 12.9% of high attaining schools (p<.005).

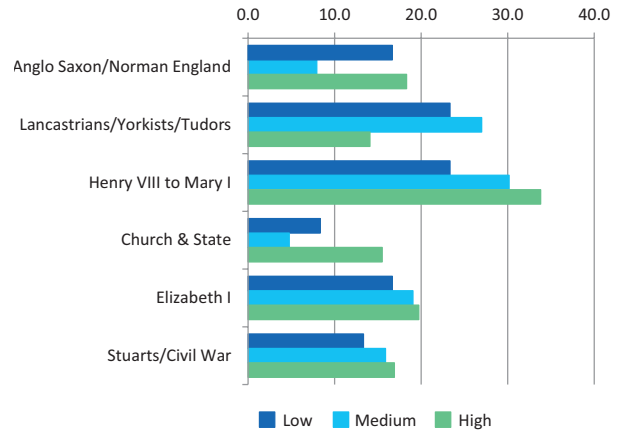


Figure 4: Percentage of schools choosing each topic by school attainment (option A)

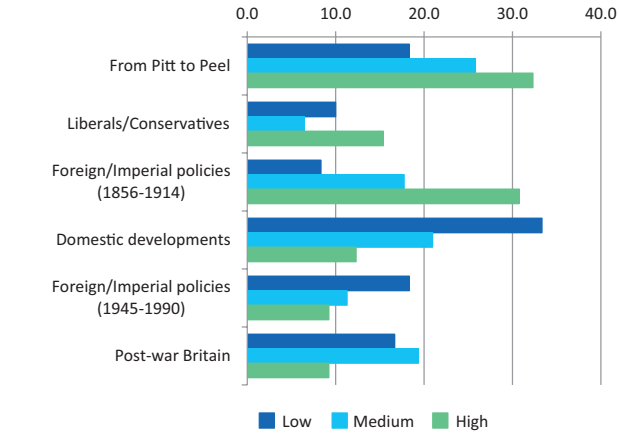


Figure 5: Percentage of schools choosing each topic by school attainment (option B)

Topic choice by school type and school attainment

Finally, an analysis of topic choice by attainment level within each school type was undertaken, to discover whether any of the differences observed were to do with the school type or the school attainment level or both. Using logistic regression, it was possible to investigate if either school type (state or independent) or school attainment (mean A level score) were significant predictors of whether each topic was taught or not.

Table 2: Logistic regression model summaries for each unit topic

Component	*Topic*	*Model summary*		*Variables in equation*			
				School attainment		*School type*	
		Chi-square[3]	*Nagelkerke*[4]	*Wald*[5]	*Exp*[6]	*Wald*	*Exp*
F961A	From Anglo Saxon to Norman England (1035–1087)	3.396	.032	.650	.905	2.342	2.371
	Lancastrians, Yorkists and Tudors (1450–1509)	1.392	.011	1.313	.678	.300	1.307
	Henry VIII to Mary I (1509–1558)	2.596	.022	2.638	1.658	.525	.725
	Church and State (1529–1589)	2.578	.028	.127	1.178	.175	4.388
	England Under Elizabeth I (1558–1603)	.137	.001	.005	.943	.189	3.476
	The Early Stuarts and the Origins of the Civil War (1603–1642)	1.125	.010	1.050	1.496	.258	.750
F961B	From Pitt to Peel (1783–1846)	6.246*	.049	5.732**	2.084	.900	.665
	Liberals and Conservatives (1846–1895)	3.472	.038	2.974	2.061	.176	.780
	Foreign and Imperial policies (1856–1914)	9.991*	.084	1.004	.316	4.236**	2.492
	Domestic developments (1918–1951)	12.138*	.098	6.559**	.429	.704	.672
	Foreign and Imperial policies (1945–1990)	.890	.009	.748	.719	.008	1.050
	Post-War Britain (1951–1994)	1.382	.013	1.118	.685	.004	1.033

* Overall model significant predictor of topic choice ($p<.05$) ** Individual predictor significant predictor within model ($p<.05$)

3. If the Chi-square is significant, it indicates that the model with the predictor variables included is a significantly better predictor than using the constant alone.

4. Gives an estimate for the proportion of the total variance that can be explained by the model.

5. Gives a Chi-square value that tests the unique contribution of each predictor.

6. Represents the change in odds of the topic being chosen by a centre, when a unit change of 1 occurs. For example for the topic From *Pitt to Peel (1783-1846),* centres that had an average A level grade of B were approximately twice as likely to choose this topic than centres that had an mean A level grade of C (Exp=2.084).

Twelve logistic regression analyses were run, using the enter method; one separate regression for each topic in Units F961A and B. For each regression, the outcome variable was dichotomous (topic taught – yes or no). The two predictor variables were the categorical variable school type (state or independent), and the continuous variable centre mean A level performance.

Table 2 shows the overall significance of each model, and the strength of each predictor variable within each model. Overall, the models accounted for less than 10% of the total variance, suggesting generally the models were not strong predictors. However, there were three topics where the regression model was a significantly better predictor than just using the overall proportion (*From Pitt to Peel, 1783–1846*; *Foreign and Imperial policies, 1856–1914*; and *Domestic developments, 1918–1951*). All of these topics were within Unit F961B. For the topic *From Pitt to Peel*, a one grade increase in mean A level performance doubled the likelihood that centres would teach this topic area (Exp.=2.084). However, school type did not have any significant influence on this topic choice. For the topic *Foreign and Imperial policies (1856–1914)*, independent schools were more than twice as likely to teach this topic compared to state schools (Exp.=2.492), although mean A level performance was not a significant predictor. Finally, for the topic *Domestic developments*, a one grade increase in mean A level performance reduced the likelihood that centres would teach this topic area by more than half (Exp.=.429). No significant interaction effects were found in any of the models.

Questionnaire findings

As full information on the topic choices was only available for one unit, the questionnaire allowed data to be gathered on choices made across the entire A level History course, for a sub-set of centres. The questionnaire findings are reported in two sections. The first section analyses the *unit* choices that were offered by schools at AS and A level. The second section analyses the *topics* within each AS and A level unit that were taught by schools.

Unit choices

The heads of department were asked to report which AS and A2 History unit combinations they offered to students. There were eight possible unit combinations at AS level, and two possible unit combinations at A2 level. The percentage of schools that offered each unit combination is provided in Tables 3 and 4.

The unit combinations that included a Modern History element (combinations 2, 4, 6 and 8 in Table 3) were generally the most commonly offered to students. At AS level the three most popular overall unit combinations (2, 4 and 8) included at least one unit that focused on Modern History, while the two most popular unit combinations included *only* Modern History units. At A2 level, again the unit combination that included a Modern History unit was the most commonly chosen, with over four out of five schools offering it to students.

There were some differences noted between different school types. At AS level there was a higher percentage of independent schools that offered unit combinations that comprised Medieval and Early Modern History units exclusively (combinations 1 and 5 in Table 3) , with approximately a quarter of independent schools offering each combination, compared to less than a tenth of state schools. In this sample, state schools were also more likely to offer a unit combination that comprised one Medieval and Early Modern unit and one Modern unit (combinations 2, 3, 6 and 7), although this was not common. For example, 28.3% of state schools offered AS level unit combination 2, while only 9.1% of independent schools did so.

Independent schools were more likely to offer more than one unit combination to students. At AS level, 27.3% of independent schools offered more than one unit, compared to only 13.2% of state schools. At A2 level, 21.2% of independent schools offered more than one unit combination, compared to only 1.9% of state schools.

Table 3: Unit combinations at AS level offered by different school type

Unit Combination	*Unit Number/Name*								*Overall number of schools that offer unit combination (%)*	*Number of **state** schools that offer unit combination (% of total state schools)*	*Number of **independent** schools that offer unit combination (% of total independent schools)*
	F961 – British Period History Studies		*F962 – European & World History Period Studies*		*F963 – British History Enquiries*		*F964 – European & World History Enquiries*				
	A – Medieval & Early Modern	*B – Modern & Early Modern*	*A – Medieval & Early Modern*	*B – Modern & Early Modern*	*A – Medieval*	*B – Modern*	*A – Medieval*	*B – Modern*			
1	✓						✓		13 (14.4)	5 (9.4)	8 (24.2)
2	✓							✓	19 (21.1)	15 (28.3)	3 (9.1)
3		✓					✓		5 (5.6)	2 (3.8)	3 (9.1)
4		✓						✓	30 (33.3)	13 (24.5)	16 (48.5)
5			✓		✓				10 (11.1)	2 (3.8)	16 (48.5)
6			✓			✓			4 (4.4)	2 (3.8)	8 (24.2)
7				✓	✓				10 (11.1)	7 (13.2)	3 (9.1)
8				✓		✓			24 (26.7)	17 (32.1)	6 (18.2)

Table 4: Unit combinations at A2 level offered by different school type

Unit Combination	*Unit Number/Name*			*Overall number of schools that offer unit combination (%)*	*Number of **state** schools that offer unit combination (% of total state schools)*	*Number of **independent** schools that offer unit combination (% of total independent schools)*
	F965 – Historical Interpretations & Investigations	*F966 – Historical Themes*				
		A – Medieval & Early Modern	*B – Modern*			
1	✓	✓		25 (27.8%)	13 (24.5)	12 (36.4)
2	✓		✓	73 (81.1%)	41 (77.4)	28 (84.8)

Topic choices

The heads of department were asked to report which AS and A2 History topics their centres offered to at least one class of students. These data were analysed to gather information on the popularity of topics across the entire course, and the most popular topics encapsulated within each unit. Table 5 lists the top 18 topic choices across all the units (out of a total of 54 choices), including the unit number, and the period of history it is linked to.

Of the top 18 topic choices within units, 13 of them were from a Modern unit. The most popular Modern units focused primarily on European History, specifically Russia (*Russian Dictatorship, 1855–1992*; *From Autocracy to Communism: Russia, 1894–1941*), Germany (*Dictatorship and Democracy in Germany, 1933–1963*; *Democracy and Dictatorship in Germany, 1919–1963*) and topics related to the two World Wars (*Churchill, 1920–1945*; *The Challenge of German Nationalism, 1789–1919*). The most popular non-European History topics focus primarily on the USA, both domestically (*Civil Rights in the USA, 1865–1992*; *The Origins of the American Civil War, 1820–1861*) and in respect to foreign relations (*The USA and the Cold War in Asia, 1945–1975*).

The three most popular topics offered within Medieval and Early Modern units were all related to the House of Tudor (*Mid-Tudor Crises, 1536–1569*; *Rebellion and Disorder under the Tudors, 1485–1603*; and *Henry VIII to Mary I, 1509–1558*).

Discussion

The current study intended to investigate the scope of one A level History course, and aimed to understand how schools utilised the optionality available to them. The optionality offered by examination boards at A level History is likely to be in response to the potential for History courses to serve multiple purposes including: covering content across a wide time span; the imperative to prepare students for later study; and the potential for History to aid students' identity formation (Harris, 2013).

The study used statistical information on students' question choices derived from Cambridge Assessment's internal databases, and a questionnaire sent to heads of History departments. Taken together, these two methods of data collection allowed school level analyses across the full scope of the course.

There have been a number of claims which argue that there is too great a focus on 20th century History in UK schools (Fitzgerald & Hodgkinson, 1994; Lang, 1990). Approximately 60% of centres sampled taught either a combination of F961B and F964B or F962B and F963B; the two unit combinations which permit Modern History to be studied exclusively. While, in this qualification at least, choosing a Modern History option does not necessarily mean having to select a 20th century topic, in practice the most popular topic choices were based in the 20th century. Furthermore, whilst the qualification structure permits schools to teach students a

Table 5: Top 18 topic choices offered by centres

Rank	Topic	Unit		Historical period		% Schools teaching that topic	% Schools teaching associated unit who teach that topic
		Code	Name	Medieval & Early Modern	Modern		
1	Russian Dictatorship (1855–1992)	F966	Historical Themes		✓	35.6	45.7
2	Civil Rights in the USA (1865–1992)	F966	Historical Themes		✓	28.9	37.1
3	Dictatorship & Democracy in Germany (1933–1963)	F964	European & World History Enquiries		✓	18.9	35.4
4	Mid-Tudor Crises (1536–1569)	F963	British History Enquiries	✓		14.4	65.0
5=	Churchill (1920–1945)	F963	British History Enquiries		✓	13.3	44.4
5=	Rebellion & Disorder Under the Tudors (1485–1603)	F966	Historical Themes	✓		13.3	50.0
7=	Henry VIII to Mary I (1509–1558)	F961	British History Study Periods	✓		11.1	37.0
7=	Democracy & Dictatorship in Germany (1919–1963)	F962	European & World History Period Studies		✓	11.1	25.6
7=	The Origins & Causes of the French Revolution (1774–1795)	F964	European & World History Enquiries		✓	11.1	20.8
10=	From Pitt to Peel (1783–1846)	F961	British History Study Periods		✓	10.0	26.5
10=	From Autocracy to Communism: Russia (1894–1941)	F962	European & World History Period Studies		✓	10.0	23.1
10=	The Origins of the American Civil War (1820–1861)	F964	European & World History Enquiries		✓	10.0	18.8
10=	The USA & the Cold War in Asia (1945–1975)	F964	European & World History Enquiries		✓	10.0	18.8
14	The Age of Gladstone & Disraeli (1865–1886)	F963	British History Enquiries		✓	8.9	29.6
15=	The First Crusade & Crusader States (1073–1130)	F964	European & World History Enquiries	✓		7.8	58.3
15=	The German Reformation (1517–1555)	F964	European & World History Enquiries	✓		7.8	58.3
15=	The Challenge of German Nationalism (1789–1919)	F966	Historical Themes		✓	7.8	10.0
15=	The Changing Nature of Warfare (1792–1945)	F966	Historical Themes		✓	7.8	10.0

combination of Modern and Medieval History, this is taken up by the minority. In the statistical analysis of Units F961A and B, independent schools were more likely than state schools to have students that answered questions on topics related to both Modern and Medieval and Early Modern periods (although this was not common). This may be due to the additional resources independent schools may have, which allow them to offer different routes through the course. Interestingly, however, the questionnaire analyses revealed both school types favoured units that matched in terms of the period of History studied (e.g. two Modern History units). The two most popular unit combinations at AS level, and the most popular unit combination at A2 level, studied Modern History exclusively.

History courses have also been criticised for their perceived focus on British and European History (Bowen *et al*., 2012; Evans, 2011; Tillbrook, 2002). The specification investigated in the current study attempts to negate this criticism by incorporating units that cover European and World History. However, the majority of the 23 topics within these units primarily focus on Europe. Non-European topics include the following:

- *Civil Rights in the USA (1865–1992)*
- *The USA in the 19th Century: Westward Expansion and the Civil War (1803–1890)*
- *Crisis in the Middle East (1948–2003)*
- *The Rise of China (1911–1990)*
- *The Origins of the American Civil War (1820–1861)*
- *The USA and the Cold War in Asia (1945–1975)*

The approach to unit and topic selection primarily observed in the study, where students cover increasing amounts about shorter periods of time, is referred to as the 'bore-hole effect' (Fisher, 1995), and has been identified as problematic due to its potential to narrow the scope of History. Indeed, the Smith (2013) review suggested that A level History students should study topics covering at least a 200-year period. The data collected in this project suggest that, in general, schools seek to teach in-depth within a historical era, rather than breadth over different historical periods. For example, the most popular unit combination comprised F961B and F964B, which was taught by one-third of the participants' schools. Within this combination, the most popular topics were *From Pitt to Peel (1783–1846)* and *Dictatorship and Democracy in Germany (1933–1963)*. Students that were taught both these topics studied a period of 180 years. Furthermore, for the second most popular combination observed in the present study (comprising units F962B and F963B), the most popular topics were *Democracy and Dictatorship in Germany (1919–1963)* and *The Age of Gladstone and Disraeli (1865–1886)* respectively, covering a period of only 98 years. Therefore, it is currently possible – and common – within this specification for students not to meet the suggestions made by Smith (2013). As such, whilst the specification does not promote the 'bore-hole effect', it is questionable

whether the optionality promoted by the specification meets the objectives underpinning the course.

The question that arises here is whether a broad coverage of historical periods, and a broad geographical context is indeed required, either for students to make a successful transition to university, or for future life and employment. Unit or topic choice in A level History is not currently a factor to differentiate between applications for university. As mentioned in the introduction, 9 of the 23 Russell Group universities which offer undergraduate degrees in History do not stipulate that applicants must have an A level in History. It is likely that the *skills* developed as part of the study of A level History are what is valued most by admissions tutors (Suto, 2012), as indicated by the fact that History is one of the most popular (Vidal Rodeiro & Sutch, 2013) and most useful (Russell Group, 2013; Suto, 2012) subjects for university applicants.

Conclusions and implications

The teaching of History, and History qualifications, are influenced by factors related to the personal, political and academic landscape (Harris, 2013), in addition to factors at the level of the school and classroom. The current study was a first attempt to determine the choices centres make, in relation to an A level History course.

In response to the potential for History courses to serve multiple purposes, an optionality approach to History qualifications has been adopted. This study has found that centres appear to favour particular historical periods and topics over others, and that these preferences are, at least in part, determined by the attainment level of schools, and the type of school. Given these observed differences, further research is required to investigate how and why centres prefer certain historical topics over others. Teachers may select topics based on their personal areas of interest or expertise (Bowen *et al*., 2012). Topic selection may also be guided by a desire for overlap between the current course content, and course content students had covered in previous qualifications. This course coherence may be seen as beneficial to students, as they have a platform of knowledge from which new information and understanding can be achieved. However, it could be problematic if students persist with academic behaviours that are not suitable for the new educational level (Conley, 2010). Furthermore, teachers may be influenced in their History topic choices by the availability (or quality) of curriculum support resources (Child, Devine, & Wilson, 2013; Devine & Wilson, 2013; Wilson & Devine, 2013a, 2013b).

A second avenue for future investigation concerns whether curriculum coherence across the different stages of education can be achieved in the study of History. If historical breadth is not currently being imparted through an optionality approach to A level History, a question arises about whether optionality should indeed be reduced. However, it is currently unclear as to what the appropriate History content at A level would be (Hibbert, 2006). An area for further study may be whether there is value in studying similar subjects at different stages of education (primary, early secondary, GCSE etc.), or whether optionality in History is utilised and valued differently by different populations taking History qualifications (e.g. different ethnic or socio-economic groups).

Acknowledgements

We wish to thank Tom Benton, Sylvia Green, Tom Bramley and Irenka Suto from Cambridge Assessment's Research Division, and Mike Goddard from OCR, for their helpful advice on this paper. We also wish to thank Jo Ireland for her administrative assistance during the study. Finally, we are grateful to the participants for engaging with this research.

References

Bowen, L., Bradley, K., Middleton, S., Mackillop, A., & Sheldon, N. (2012). History in the UK national curriculum: A discussion. *Cultural and Social History, 9*(1), 125–143.

Burn, K., & Harris, R. (2012). *Historical Association survey of history in schools in England 2011*. Retrieved from http://www.htai.ie/docs/2012_docs/pdf/History_in_English_Secondary_Schools_2011_Survey_Historical_Association.doc

Child, S. F. J., Devine, A., & Wilson, F. (2013). *"It's gold dust." Teachers' views on curriculum support resources*. Cambridge: Cambridge Assessment.

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

Department for Education. (2010). *The Importance of Teaching – the Schools White Paper*. Retrieved from https://www.education.gov.uk/publications/standard/publicationDetail/Page1/CM%207980

Department for Education and Skills. (2007). *Diversity and citizenship curriculum review*. London: DfES.

Devine, A., & Wilson, F. (2013). *Curriculum support resources in English: Opportunities for teacher learning*. Cambridge: Cambridge Assessment.

Evans, R. J. (2011). The Wonderfulness of Us (The Tory Interpretation of History). *London Review of Books, 33*(6), 9–12.

Fisher, T. (1995). The New Subject Core for A Level History. *Teaching History, 80,* 18–19.

Fitzgerald, I., & Hodgkinson, S. (1994). British University History Now. *History Today, 44,* 53–57.

Gove, M. (2010). *Speech to the Conservative Party Conference in Birmingham*. Retrieved from http://centrallobby.politicshome.com/latestnews/article-detail/newsarticle/speech-in-full-michael-gove/

Harris, R. (2013). The place of diversity within history and the challenge of policy and curriculum. *Oxford Review of Education, 39*(3), 400–419.

Hibbert, B. (2006). *The articulation of the study of history at General Certificate of Education Advanced Level with the study of history for an honours degree*. (Unpublished doctoral thesis), University of Leeds.

Joint Council for Qualifications. (2013). A, AS and AEA results 2013. Retrieved from http://www.jcq.org.uk/examination-results/a-levels

Lang, S. (Ed.). (1990). *A Level History: The Case for Change*. London: Historical Association.

Ofqual. (2013). *Consultation on new A level regulatory requirements*. Retrieved from http://comment.ofqual.gov.uk/a-level-regulatory-requirements-october-2013/

Russell Group. (2013). *Informed Choices: A Russell Group Guide to Making Decisions About Post-16 Education*. Retrieved from http://russellgroup.org/InformedChoices-latest.pdf

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

Suto, I. (2012). *What are the impacts of qualifications for 16 to 19 year olds on higher education? A survey of 633 university lecturers*. Cambridge: Cambridge Assessment.

Tillbrook, M. (2002). Content Restricted & Maturation Retarded? Problems with the Post-16 History Curriculum. *Teaching History, 109,* 24–26.

Vidal Rodeiro, C., & Sutch, T. (2013). *Popularity of A level subjects among UK university students*. Cambridge Assessment. Cambridge.

Wilson, F., & Devine, A. (2013a). *Curriculum support resources in mathematics: Opportunities for teacher learning*. Cambridge: Cambridge Assessment.

Wilson, F., & Devine, A. (2013b). *Curriculum support resources in science: Opportunities for teacher learning*. Cambridge: Cambridge Assessment.

Students' views and experiences of A level module re-sits

Tim Gill and **Irenka Suto** Research Division

Introduction

In this article, we report on a study exploring over 1,300 students' views and experiences of re-sits at A level[1]. We focus on two popular but contrasting A level subjects: Psychology and Mathematics. Anticipating reforms to A level assessment, our aim in collecting the data was to gain an understanding of what the likely effects of a system of reduced re-sits would be on students and their teachers. The findings of the study could help those seeking to support students and teachers during the current transition to linear assessment at A level.

Background

Historically, school qualifications in England, such as GCSEs and A levels, followed a linear approach, whereby students were assessed on what they had learnt at the end of a two-year course. Subsequently, an alternative, modular, structure became the norm, with the content of the course broken up into a series of 'chunks', to be taught and then assessed separately. In 2000 all A levels adopted this modular structure. A few GCSEs also became modular in 2003 and 2004, and the majority did so in 2009.

An important feature of the modular approach is the opportunity for students to re-sit modules if they are unhappy with the grade they received on that module, or want to try to improve their overall grade. Until recently, GCSE and A level students have been able to re-sit modules in multiple examination sessions (in January and June each year) and to keep the best result obtained. This contrasts with the linear approach, where the only way to improve on the grade is to re-take the whole qualification. When modular A level specifications were first introduced, a limit of one re-sit per module was imposed. However, this limit was removed in 2003 (BBC, 2003).

Module re-sits are a controversial issue. There is a widespread perception that they have (until recently) contributed to a year-on-year improvement in the A level pass rate and therefore to the perceived lowering of the A level standard (De Waal, 2009; Higton, Noble, Pope, Boal, Ginnis, Donaldson & Greevy, 2012). It has also been claimed that the modular system engenders a deleterious focus on exams and alleged 'teaching to the test' in the classroom at the expense of deeper learning (Poon Scott, 2010, 2012; Higton *et al*., 2012). Criticism has also come from within examination boards, with the Chief Executive of AQA claiming that too many re-sits may 'distort results' (BBC, 2010).

These views are shared by the current UK Government. At the start of its term in office, it raised concerns in an education White Paper (Department for Education, 2010) that the number of re-sits in GCSEs and A levels were "undermining" the qualifications. The national qualifications regulator, Ofqual, was asked to change the rules on assessment to prevent students re-sitting a large number of modules. Over the past two years, whilst re-sit opportunities have decreased in the interim (e.g. through the removal of the January examination session in 2014), the Secretary of State for Education has spearheaded a wider programme of qualifications reform which sees A levels and GCSEs return to a fully linear structure (Department for Education, 2014). In the majority of popular subjects, new fully linear A level syllabuses will be ready for first teaching in September 2015, with the first cohort of students being awarded their qualifications in the summer of 2017 (Ofqual, 2014).

There is certainly considerable evidence that many students have taken advantage of opportunities to re-sit, particularly at A level. Ofqual's predecessor, QCA (2007b), found that the percentage of students re-sitting the most popular modules in a range of A levels in 2006 was generally between 30% and 50%. Gill and Suto (2012) looked at re-sitting behaviour in A level Psychology and Mathematics in 2010 and found that 66.3% of Psychology students and 74.1% of Mathematics students re-sat at least one module.

There is less evidence for the claim that students carry on re-sitting each module until they reach a desired grade. It is only a small percentage of students who re-sit more than once. QCA (2007b) found that the percentage of students re-sitting the most popular modules multiple times in several A levels varied from 3.5% to 9.5%. Gill and Suto (2012) found slightly higher figures: of all students taking the OCR specification in the subject in 2010, 7.1% of Psychology students and 11.7% of Mathematics students re-sat the most popular module (in terms of re-sits) more than once.

It is certainly the case that re-sitting modules tends to lead to improvements in the grade achieved on the module, and sometimes to improvements in the overall grade. For a range of subjects, QCA (2007b) compared the percentage of A grades that would have been awarded had the students taken their AS results from the end of Year 12 (i.e. ignoring re-sits in Year 13), with the actual percentage of A grades awarded. Mathematics was the subject that showed the greatest improvement through re-sitting (7.8%), followed by French (7.2%), English Literature (5.0%) and Physics (4.5%).

Gill and Suto (2012) found that the percentages of students improving a module grade by re-sitting was between 51% and 65% of those who re-sat for A level Psychology, and between 54% and 79% for A level Mathematics. However, the impact on the overall grade was considerably less: of all students sitting the specifications in 2010, 26.5% of Psychology students and 34.8% of Mathematics students improved their overall grade through re-sitting.

This raw data is informative but does not reveal the reasons why students re-sit. If students have genuinely gained more knowledge by studying more advanced modules later in the course, or if they

1. The A level is the most popular qualification taken by students between the age of 16 and 18 in England (Years 12 and 13 of schooling). It is usually studied over two years and is made up of two parts; AS (whose modules are usually taken in Year 12) and A2 (modules usually taken in Year 13). The AS level is available as a stand-alone qualification, as well as contributing towards a full A level.

were feeling ill the first time they took an examination, then it seems reasonable that they should be allowed to demonstrate that they did not initially perform to their true ability. Many teachers interviewed by Higton *et al*. (2012) felt that not allowing re-sits in a modular course would disadvantage students who were slow starters. Some commented that the final grade achieved was always deserved, regardless of how many re-sits were involved, because it demonstrated a certain amount of knowledge and understanding. Poon Scott (2012) used a questionnaire and interviews to collect information on A level students' re-sitting experiences. She found that studying A2 modules in Year 13, the second year of the course, helped students with AS level re-sits (from the first year of the course), both through improved knowledge and through better exam technique.

Furthermore, levels of student motivation at the time when they first sit module examinations are not known. Some schools and colleges like to enter all (or most) students for a module exam at the earliest opportunity, to give them examination practice. Poon Scott (2012) found that for some A level students, their first sitting of a module exam was rather too soon, and they performed poorly. Others were more laid back about their first sitting because they knew that they had the opportunity to re-sit. Similarly, Vidal Rodeiro and Nadas (2011) interviewed students taking modular GCSEs and found that the knowledge that they could re-sit a module meant they worked less the first time they took the exam than they would have done without re-sit opportunities. It seems reasonable that any improvement these students made through re-sitting is valid. This conclusion fits with that of Al-Bayatti and Jones (2003) who found that students re-sitting AS level modules in January of Year 13 performed worse, on average, the first time they took the exam, than would be predicted by their GCSE grades. Their subsequent performance on the re-sit was much closer to their expected level.

However, others argue (De Waal, 2009) that the original intention of re-sits, to give students who performed below their best on the day another chance, has been superseded by students using them to play the system. For instance, there is a feeling that some students try to boost their overall grade by re-sitting 'easier' AS modules (studied in Year 12) rather than focusing on performing well in the A2 modules in Year 13. There is certainly evidence that students re-sit AS modules in far greater numbers than they do A2 modules (QCA, 2007a; Gill & Suto, 2012). However, this is not to say that students are deliberately targeting the AS modules in this way; just the fact that there are more opportunities to re-sit AS modules means it is more likely that they will be re-sat. Poon Scott (2010), found this tactic to be a rare occurrence, with only 2.5% of students giving it as a reason for re-sitting. It is also worth noting that the introduction of the A* grade in June 2010 means that this approach would not apply to the very best students, who require high marks on the A2 modules in order to reach the highest grade.

A further concern with modularisation and re-sitting is that it has led to a focus on exams at the expense of deeper learning. Students interviewed by Poon Scott (2012) made comments about their approach to exams being to revise hard, but then they fail to retain the information after the exam. Teachers interviewed by Higton *et al*. (2012) often complained that their students were disrupted by re-sits and lost their focus on what they were studying. The teachers also felt they had less time to teach beyond the syllabus. These views were similar to those reported by teachers surveyed in other studies (De Waal, 2009; Williams, 2009; NASUWT, 2008).

It is not only teachers who are concerned about an excessive focus on exams and re-sits. Media reports suggest some universities will not accept A level results that are achieved with the use of re-sits (Grimston, 2010). Poon Scott (2010) spoke to several university admissions tutors who believed that re-sitting meant that deep learning had been compromised and students were therefore not ready for university. One admissions tutor said that he would not consider students who achieved their grade through re-sitting, whilst two others said they would want to know the reasons for re-sitting. Ofqual (2013) reported on the perceptions of A levels amongst various stakeholders and found that the biggest concern from representatives of higher education institutions was "too many re-sits". This was also a major concern among most of the 633 university lecturers surveyed by Suto (2012). Teachers interviewed by De Waal (2009) also believed that grades were less worthy if achieved by re-sitting modules, and could lead to students going to the wrong universities.

Anticipating the current A level reforms, we conducted a questionnaire-based study in 2011, exploring students' views and experiences of re-sits in two popular A level subjects. Our aim was to provide our examination board colleagues with an understanding of what the likely effects of a system of reduced re-sits would be on students and their teachers. This would potentially help colleagues to provide stakeholders with maximum support during the transition period and beyond. In the study, we investigated how A level re-sits were being used, whether students were playing the system, and whether the reasons behind decisions to re-sit were genuine and valid. We also explored whether the amount of time spent on re-sit exams was such that it interfered with learning new subject content for other modules. In a few years' time, once a reformed system of new linear A levels has bedded down, data from this study may prove useful in comparative research.

Method

Subjects

Two A levels offered by the Oxford, Cambridge and RSA (OCR) exam board were selected as the focus of the questionnaires: Mathematics and Psychology. They were chosen because they were popular A level subjects and they contrasted in some important ways. First, at the time of the data collection, to obtain a Mathematics A level students were required to complete six modules (three AS level and three A2 level), whereas for Psychology A level only four modules were required (two AS level and two A2 level). The larger number of modules in Mathematics meant there was more opportunity to re-sit. A further difference was in the extent of choice of modules. In Psychology, all four modules were compulsory (although there was some choice of topic within one A2 module). In Mathematics, students were required to study four core Mathematics modules (two at AS level and two at A2 level), but then had a choice of a combination of Mechanics, Statistics or Decision Mathematics modules for their other two modules. Finally, there was some difference in the way the two subjects were structured; in Mathematics much of the learning in later modules built upon knowledge gained in earlier modules and may have helped with the understanding of the content of earlier modules. This meant that students could benefit from re-sitting some of the earlier modules late in the course. This was less the case in Psychology, where modules were more stand-alone. Overall, these differences suggested that differences in re-sitting behaviour between the two subjects would be likely.

Questionnaire design and piloting

A questionnaire was developed for Year 13 students with alternative versions for Mathematics and Psychology. Year 13 students were targeted because Year 12 students would not have had the opportunity to re-sit any modules at the time the questionnaire was sent out. It was decided to keep the questionnaire as short as possible so that students would not feel daunted by its length. This meant focusing on a few core aspects of the re-sitting experience: the reasons why students re-sit: who influences their decision; how they prepare; and their general views of re-sits.

The content was partly determined by reviewing the literature and considering which issues were covered in other questionnaires (e.g. Poon Scott, 2010). Some of the possible responses to the questions were based on media and public perceptions of re-sits. This included the following reasons for re-sitting: treating the first sitting of an exam as practice; those just below a grade boundary re-sitting on the off chance they might go up a grade; being unlucky with the questions the first time; and re-sitting 'easier' AS modules to boost overall grade (De Waal, 2009). Other questions were also informed by the literature, including the view that too much time is spent preparing for re-sits, eating into teaching time for other modules (De Waal, 2009; Higton *et al.*, 2012). More positive views on re-sits were also investigated, such as the belief that they reduce exam pressure on students by acting as a safety net, or that they enable students to demonstrate that they have improved their knowledge by studying later modules. Finally, more practical aspects of the re-sitting experience were explored, such as the time spent on preparing for re-sits and the extra support that is taken up by students.

The questionnaire was successfully piloted in two schools, one for each subject. Following this, letters of invitation were sent to heads of department in all schools and colleges taking the OCR specifications, along with ten copies of the questionnaires. (Contact details were provided so that further copies of the questionnaire could be requested, as required.) The teachers were asked to give the questionnaires to students in Year 13 who had re-sat or were planning to re-sit modules. The questionnaires were sent two months after the January examination session, to allow for results to have been received by students. Schools and colleges were given four weeks to complete the questionnaires and return them.

Responses

Questionnaires were sent to all schools and colleges taking the OCR qualifications (329 in Psychology and 400 in Mathematics). Responses were received from 87 schools for Psychology and 75 for Mathematics (response rates of 26.4% and 18.8% respectively). Overall, there were more responses from Psychology students (737) than Mathematics students (614). An analysis of the background characteristics of the students and their schools/colleges confirmed their overall representativeness in terms of the OCR A level populations in the subjects, and an absence of any notable response biases.

Results

Influences on re-sit decisions

In the questionnaire, a multiple choice question was used to ask students:

> *Which person most influences your decisions about whether to re-sit modules?*

This question was asked because the way in which decisions are made may impact on students' views and experiences of re-sits, in terms of the control they feel they can exert and how happy they are with the decision. The responses (and response options) are presented in Figure 1. Despite instructing students to tick only one box for this question, some Psychology students (n = 48) ticked multiple boxes. These might be students who genuinely found it too difficult to make one choice only. However, their responses were excluded since we did not know how many other students had a similar desire to tick more than one box but felt unable to do so.

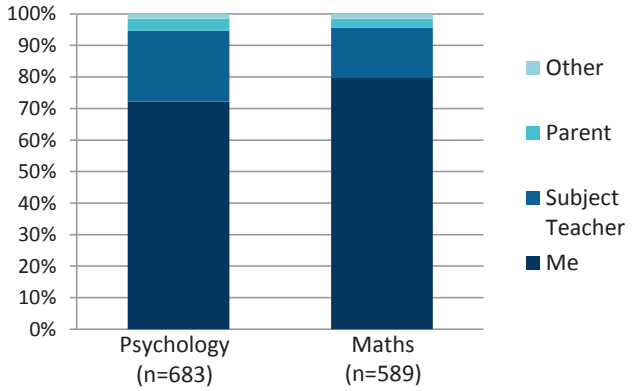


Figure 1: Influence over re-sit decision, according to students (% of responding students)

A large majority of students (72.0% for Psychology and 79.2% for Mathematics) believed that they had the greatest influence on their re-sitting decisions. Of the remaining students, some felt they were most influenced by their teachers (22.5% and 15.8% respectively) whilst a small minority felt they were influenced mainly by their parents (3.8% and 2.9%).

Reasons for re-sitting

The students were asked to choose, from a list, their reasons for re-sitting AS level modules (if they had done so). Multiple reasons were permitted. These questions focused on the AS modules (and on the compulsory ones only in Mathematics) as they were the most likely modules to have been re-sat. The opportunity to give reasons for a second re-sit of a module ('Psychological Investigations' in Psychology and 'Core Mathematics 1' in Mathematics) was included. These modules were the most likely to have been re-sat more than once (Gill & Suto, 2012).

Figures 2 and 3 present the percentages of students (who gave at least one reason) choosing each of the possible responses, for the AS modules in Psychology and Mathematics respectively. It can be seen that most students gave multiple reasons for re-sitting, with between 68% and 78% giving two or more reasons, and between 47% and 55% giving three or more. Psychology students were slightly more likely to give two or more reasons than Mathematics students.

For all modules considered, the three most popular reasons were: "I needed a higher grade for university/college"; "I thought I could do better because I had improved my knowledge through studying other modules"; and "It would be easier to boost my overall grade by re-sitting an AS level module than by doing well in A2 modules". For each module, only a very small percentage of students said that they treated the first exam as a practice, or that they had no choice in the matter.

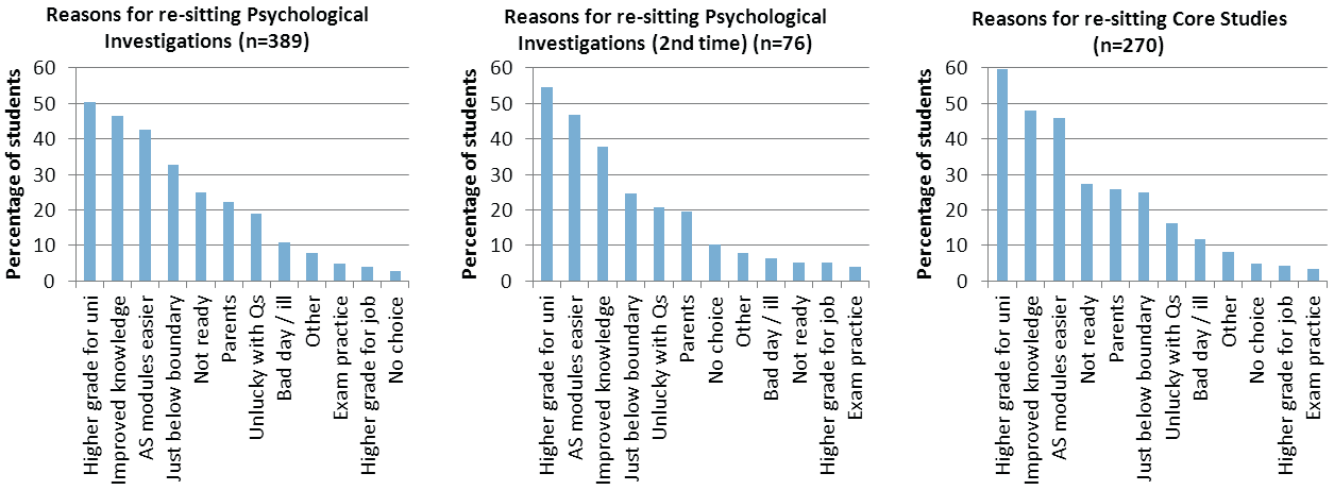


Figure 2: Reasons given for re-sitting Psychology AS modules (% of responding students)

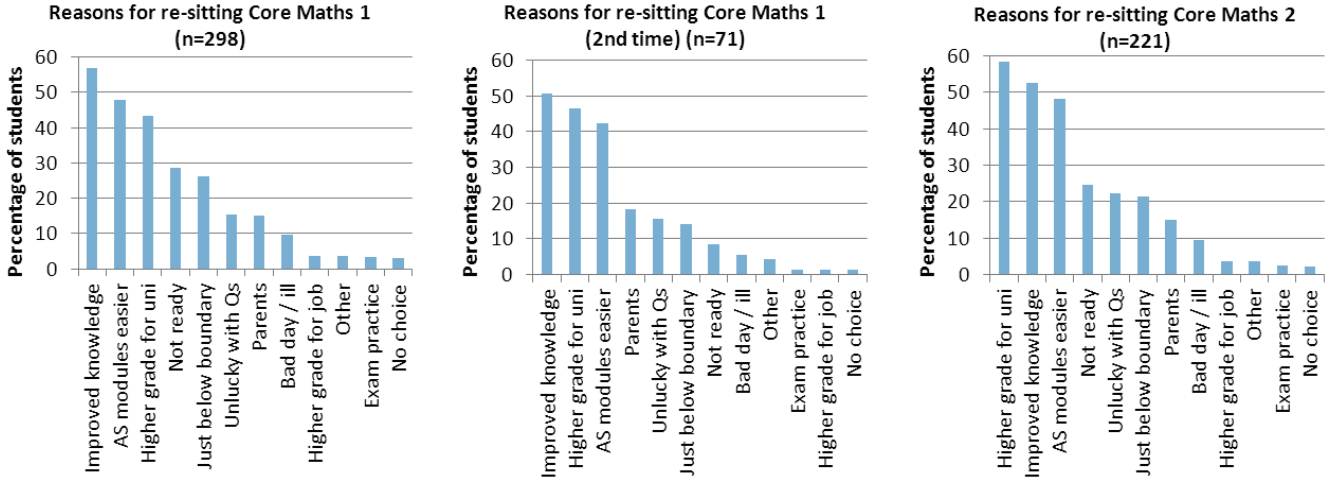


Figure 3: Reasons given for re-sitting Mathematics AS modules (% of responding students)

For the first AS module in each subject, it is noticeable that the percentage of students re-sitting, because they were not ready, was much lower for the second re-sit than for the first. This is not surprising, as it seems less likely for students to still not be ready when taking a re-sit.

The students were asked whether they intended to re-sit one of the A2 modules ('Approaches and Research Methods in Psychology' or 'Core Mathematics 3') and if so why, giving the same options as for the previous question. A common approach for A2 modules is to take one in January of Year 13, allowing for the possibility of re-sitting in June of Year 13. Therefore the A2 modules chosen for this question were those most likely to be sat for the first time in January of Year 13. The numbers of students planning to re-sit were 192 in Psychology and 242 in Mathematics.

As with the AS modules, most students gave more than one reason for re-sitting, with only around 30% giving one reason only. Their reasons given were slightly different for this planned A2 re-sit than for the completed AS re-sits. Large proportions of students (67.2% in Psychology and 68.3 in Mathematics) were planning to re-sit to get a higher grade for university. This may be partly due to the influence of the A* grade, for which students need to get 90% of UMS on A2 modules. Smaller proportions (28.9% and 34.5% respectively) had improved their knowledge by studying other modules, which is perhaps to be expected for an A2 module.

Nearly 50% of students re-sitting Core Mathematics 3 (the first A2 module) believed they were unlucky with the questions they got, a much higher percentage than for the first Psychology A2 module. This suggests it may have been a particularly difficult paper, or that there is less predictability in Mathematics exams in general than in Psychology. As with the AS modules, a higher percentage of Mathematics students (34.5%) than Psychology students (28.9%) gave improved knowledge as a reason. A slightly higher percentage than in the AS modules (35.3%) gave 'not being ready' as a reason, which may be due to some students struggling with the shift up from AS to A2 modules.

Time spent preparing for re-sits

The students were asked:

> *When preparing for exams, what proportion of your time do you spend on re-sits?*

Figure 4 displays the results.

In both subjects, just over half of the students estimated that they split their exam preparation time equally between new modules and re-sits. Almost one third of Psychology students and almost two fifths of Mathematics students spent more time on new modules. Only 10.9% of Psychology students and 6.2% of Mathematics students spent more time on re-sits. This suggests that for most students, revising for re-sits is seen as being important but does not take over to such a degree that they spend more time on this than on preparing for other module exams.

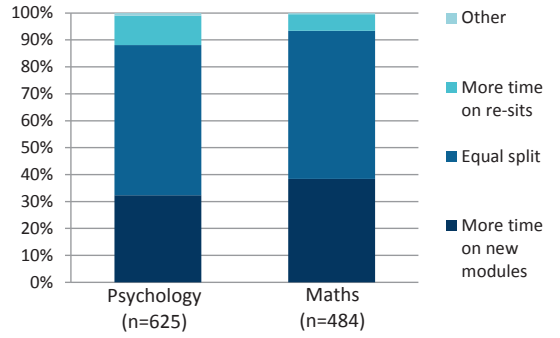


Figure 4: Proportion of time spent preparing for re-sits (% of responding students)

Ways of preparing for re-sits

Students were asked to select, from a list, the ways in which they prepared for re-sit exams.

The majority of students chose multiple preparations, with around 50–55% in each subject getting either two or three different types of help. Almost 20% in each subject indicated four or more types of preparation. The percentages of responding students selecting each option are presented in Figure 5.

The most popular methods for preparing for re-sits in both subjects were: to obtain past papers; to study with other students; and to get extra help, either informally or by attending extra lessons. Obtaining past papers was more common amongst Mathematics students (75.2%) than Psychology students (63.5%). Private tutoring was also more popular amongst Mathematics students (21.7%, compared with 4.8%).

Attitudes to re-sits and their impact on learning

To assess more general attitudes to re-sits and how they impact on learning, the students were asked to use five-point Likert scales to indicate their level of agreement with each of seven statements. Figures 6 and 7 present each statement and the percentage of students responding with each level of agreement.

Around half of the students (49.8% in Psychology and 47% in Mathematics) agreed they felt under less pressure the first time they sat an exam because they knew they could re-sit (Statement 1). However, over a third of students (34.5% in Psychology and 36.2% in Mathematics) disagreed, indicating that modular assessment did not invariably reduce stress levels.

The vast majority of students indicated that they did not treat their first sitting of an exam as a practice (Statement 2). Only 3.8% of

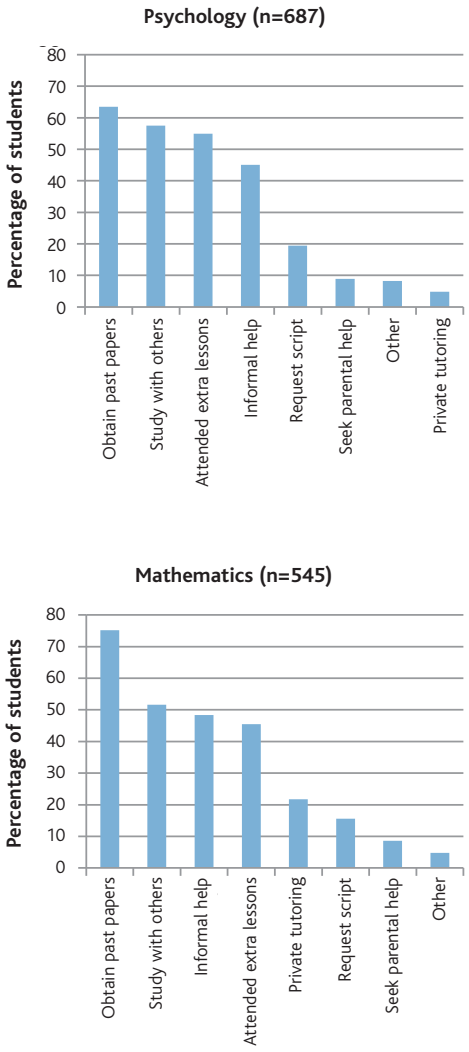


Figure 5: Preparation for re-sits (% of responding students)

Psychology students and 1.9% of Mathematics students agreed with the statement. The Mathematics students were more likely than the Psychology students to strongly disagree (60.9% compared with 48.8%). For both subjects, the difference in responses to Statements 1 and 2 indicates the existence of a group of students for whom initial attempts at A level exams are serious but less stressful events than final attempts.

Students generally agreed that re-sitting meant that they had to work harder (Statement 3), with only 5.6% of Psychology students and 6.7% of Mathematics students disagreeing. The students were also likely to agree (over 60% in both subjects) that re-sitting had improved their understanding of the subject (Statement 4), suggesting that module assessments may be being used formatively as well as summatively.

Students were less decisive in their response to Statement 5: "I feel I did less well in later modules because I spent too long preparing for re-sits of earlier modules". Around 39% in each subject neither agreed nor disagreed. This may be because they found it hard to judge the effect of re-sit preparation on their performance in other exams. Of those that did voice an opinion, the majority disagreed, with only 17.3% of all Psychology students and 13.4% of all Mathematics students feeling that re-sits led to them doing less well on other modules.

In general, students did not think that re-sitting module exams had wasted their time (Statement 6). This response fits with that for Statement 4 in supporting the idea that module assessments may be

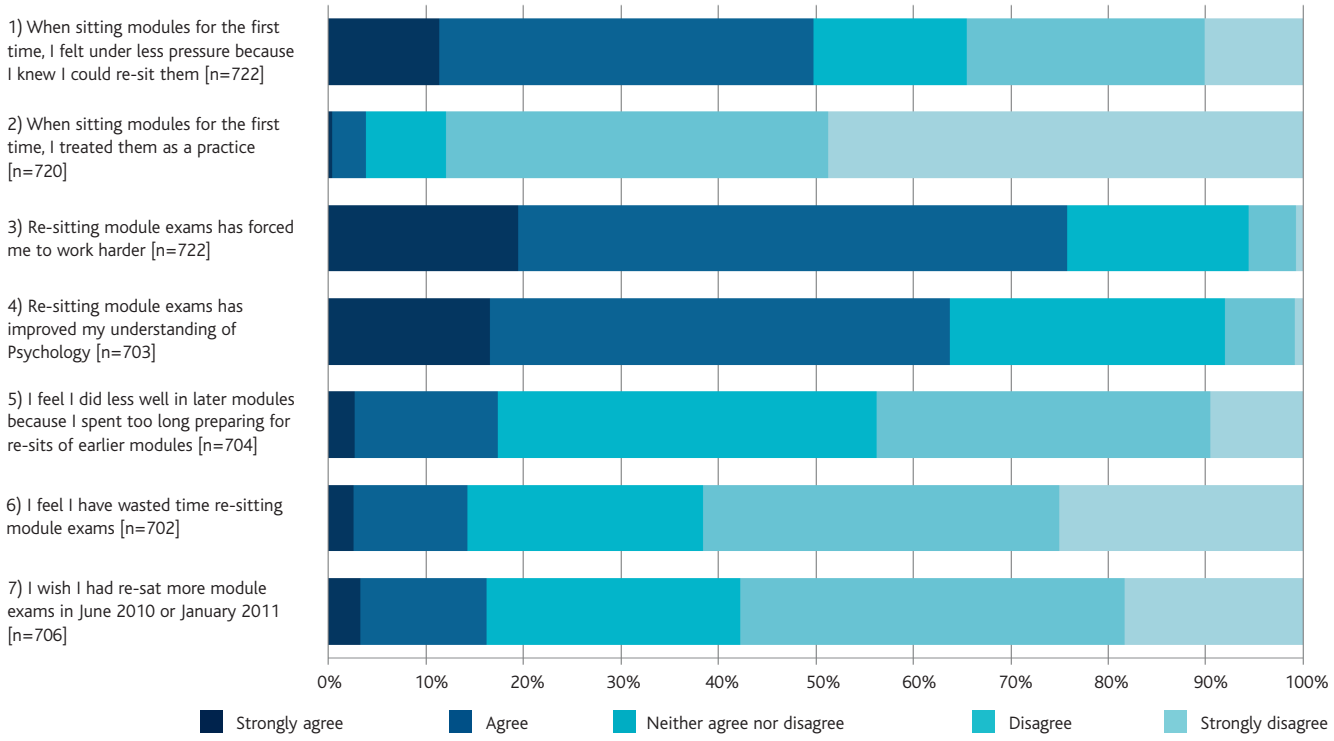


Figure 6: Percentage of responding students agreeing with statements (Psychology)

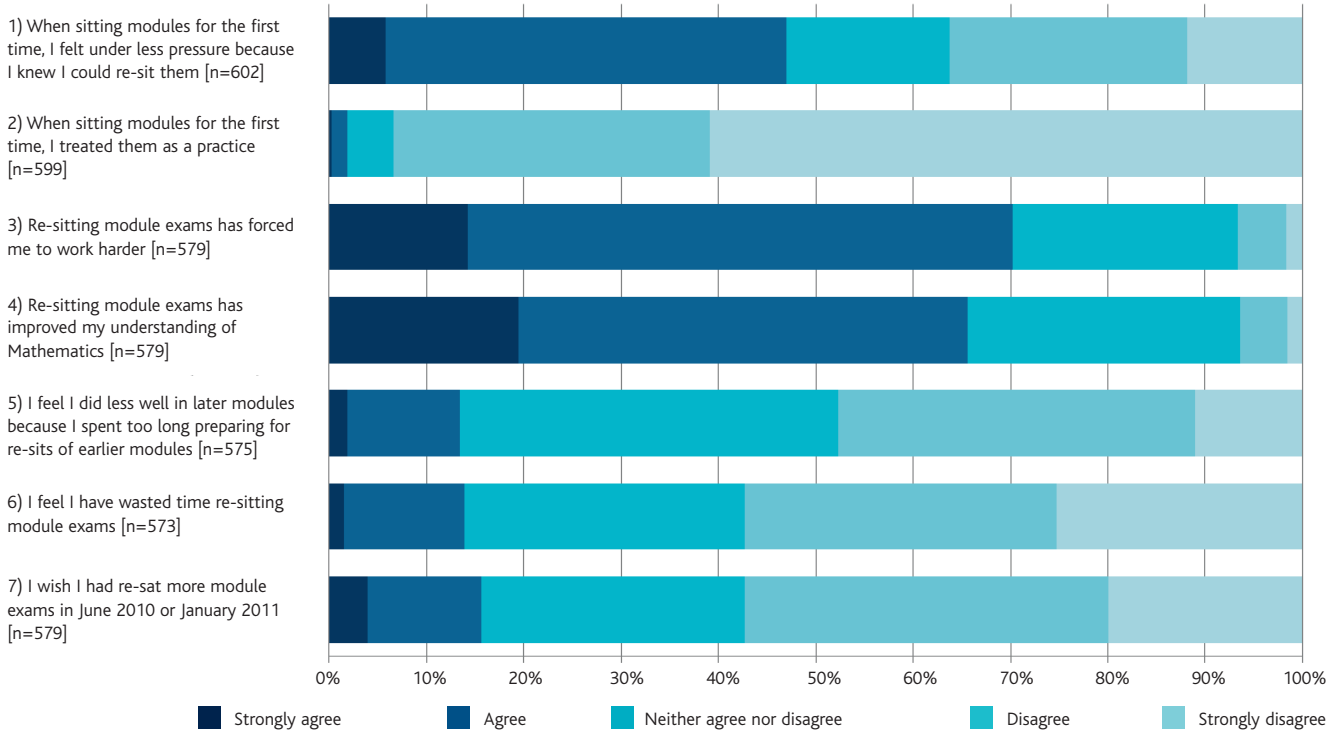


Figure 7: Percentage of responding students agreeing with statements (Mathematics)

being used formatively as well as summatively. However, a minority of students (14.3% in both subjects) believed they had wasted time in re-sitting (although they may have been only to some of their re-sits and not all of them). Another minority (16.2% of Psychology and 15.6% of Mathematics students) felt they should have done more re-sits in previous sessions (Statement 7). Over half (57.8% of Psychology students and 57.3% of Mathematics students) actively disagreed with the statement.

Further views from students

In the final section of the questionnaire, students were asked if they had any further comments they would like to share. There were over 100 comments from Psychology students and around 80 from Mathematics students, although many of these were not related to re-sits. However, of those that did relate to re-sits, there were two main themes that were common to both subjects.

1. Re-sits are good because they mean you can improve your grade (nine comments by Psychology students, five by Mathematics students). For example:

   Re-sits are a helpful way to gain extra marks and lifting your overall grade for A level. (Female Mathematics student, independent school)

   I found re-taking these exams very beneficial and it has completely changed and improved my grade beyond what I thought possible. (Female Psychology student, FE/tertiary college)

2. Re-sits are valid because they allow students who had a 'bad day' to have another chance to show what they know (seven comments by Psychology students, five by Mathematics students). For example:

   I agree with re-takes as students shouldn't be punished for having 'bad days' on the day of the exam. It also allows for another chance if students have an unfortunately planned exam table not allowing them enough time to prep as they would like. (Male Psychology student, comprehensive school)

   It gives you a second chance if you had a bad day/didn't feel well or questions were really hard. Re-takes give you a chance to do better. (Female Mathematics student, sixth form college)

Among the responses from Psychology students, two other themes stood out:

1. Re-sits cost too much, which is unfair on those who cannot afford them (eight comments).

   Although I agree with the idea of re-taking exams in order to obtain a better grade, I feel it is an unfair system and think re-taking shouldn't be allowed as it depends on how much money the students and school has. (Female Psychology student, comprehensive school)

   Re-sits are too expensive. I feel as if I've wasted my money trying again and again to improve my grades. (Female Psychology student, comprehensive school)

2. Re-sits should only be undertaken if they do not interfere too much with preparation for new modules (four comments).

   Only re-sit exam if you are definitely not jeopardising the newer modules. Make sure you don't over burden yourself. Pace yourself. (Female Psychology student, comprehensive school)

   People should only re-sit an exam if it doesn't or won't affect or interfere with any other exams which will be coming up also. (Female Psychology student, comprehensive school)

In Mathematics there were another two topics that students commented on.

1. Re-sitting AS modules later in the course is a good strategy as it is easier to get top marks on these modules (four comments).

   I am waiting until the end of course (June 2011) to re-sit Core 1 and Core 2, as I will have a better understanding of Mathematics, so hopefully will do better, as on borderline A grade. (Male Mathematics student, comprehensive school)

   Core three and four seem much harder than one or two, so it seems logical to re-sit core one or two to improve UMS. (Female Mathematics student, grammar school)

2. Re-sits are unfair on those students who do well the first time and therefore do not have to re-sit. They don't reflect the true ability of students (four comments).

   I do not think it's fair that people can re-sit C1 and C2 in Year 13 and get close to full UMS when they got much lower first time round. (Female Mathematics student, grammar school)

   Re-sits are an unfair advantage and do not reward people who attain good grades at the first time. (Male Mathematics student, comprehensive school)

Finally, in both subjects there were some interesting comments on how best to overcome the feeling that re-sits undermine A levels:

> *I believe a culture of re-sits has been a major contributor to grade inflation. If it were in anyone's interest to combat this, one could ensure a candidate definitely takes the grade they get from re-sits – allowing those who had terrible exams to re-take, but ensuring no one would casually re-take on the chance of increasing their grade.* (Male Mathematics student, comprehensive school)

> *Re-sits should be limited to one re-sit per candidate per subject. Continual re-sitting of A level modules particularly AS modules at the end of Y13... devalues A levels.* (Male Psychology student, comprehensive school)

Discussion

In this study we obtained A level students' views and experiences of re-sits in Psychology and Mathematics, prior to a reduction in re-sit opportunities taking effect nationally. Our aim was to provide examination board colleagues with an understanding of the likely effects of a system of significantly reduced re-sits on students and their teachers. Since the data was collected, the move back to linear assessment at A level has begun, beginning with the abolition of the January examination session at the start of 2014. The findings of our study indicate several important consequences for students and other stakeholders in the new assessment regime.

Traditionally, it has been argued that re-sitting individual modules in a qualification gives a second chance to students who, for one reason or another, did not initially demonstrate their knowledge, skills and understanding. Re-sitting also enables students who were underprepared the first time to become more knowledgeable about a topic, and to demonstrate this improved knowledge. Few would claim that students who genuinely benefit in this way do so unfairly. In a linear system, after all, students are assessed only when they have covered the entire course content, when knowledge acquired in Year 12 has been reinforced and augmented by knowledge acquired in Year 13. However, it has also been argued that re-sits enable some students to achieve a higher grade than they deserve by playing the system, for example by becoming more 'exam-savvy' through practice or by being 'lucky' with the questions on one of the versions of the exam. In line with these arguments, we found that one of the students' most common reasons for re-sitting could be seen as a valid means of getting a higher grade ("I had improved my knowledge through studying other modules") whilst another might be seen as playing the system to some degree ("It would be easier to boost my overall grade by re-sitting an AS level module than by doing well in A2 modules"). These findings concur with those of previous research (for example, De Waal, 2009; Poon Scott, 2010, 2012).

However, further findings from our study suggest that, in reality, the situation is perhaps more complex than this simple split would suggest. Most students who responded to the questionnaire gave multiple reasons for re-sitting a module. In each subject, a majority thought that re-sits *per se* had:

1. made them work harder, and
2. increased their knowledge of the subject.

These views indicate that module examinations do not only provide summative assessment, but are also used for formative assessment purposes too. The implication is that those responsible for new linear A levels need to think carefully about the need to offer students suitable interim assessments, for reasons of both formative assessment and motivation. The motivational feeling that module exams 'actually count for something' could be difficult to replicate in lower stakes internal exams.

It is interesting to note that the percentages of students in the study who thought they had improved their knowledge through studying other modules, were higher for both of the two Mathematics AS modules than for the two Psychology AS modules. This may be due to differences in the nature of the subjects and the course structures. Mathematics is a subject that is 'spiral' in nature, in that later modules build on knowledge gained in earlier modules. In contrast, Psychology modules tend to be more separate from each other in their content and in the background knowledge assumed. Thus it is more likely that Mathematics students will learn things in later modules that would help them in earlier modules, and will therefore struggle less than Psychology students will in the move to linear assessment. Subject differences of this kind may need to be taken into account when grade boundaries are determined for the first cohorts of students taking the new linear A levels.

We found that over half of the students in the study claimed to spend at least half of their exam preparation time on re-sits rather than on new modules. The return to a linear assessment system should free up this time (as well as the time spent preparing for first attempts at module exams partway through the course), but consideration should be given to how this time is used instead. If at present, modular assessment motivates students to spend part of their exam preparation time looking back at where they went wrong during previous module attempts and addressing gaps in knowledge revealed by those attempts, then arguably, it could be challenging for teachers to get students to do something equally or more worthwhile. On the other hand, it seems plausible that the amount of time spent on re-sit exams interferes with learning new subject content for other modules. Moreover, the freed-up time in the linear regime could be spent teaching beyond the syllabus, an activity which has been found to be associated with higher results in A level Mathematics, relative to students' performances in their other subjects (Suto, Elliott, Rushton & Mehta, 2011).

This study suggests that an important benefit of linear assessment could be the resolution of some equity problems (either perceived or actual). Several participating students felt that re-sits cost too much, and that the modular system was therefore unfair on those students who were less well off financially. The linear system should ensure that all students taking a particular A level course will sit the same number of examinations, and that their assessment costs will be uniform.

Finally, the study had several notable limitations. First, due to the self-report nature of questionnaires we cannot be sure about the honesty or accuracy of all of the responses. Secondly, the analysis of response data was limited to fairly simple descriptive statistics. No statistical tests were undertaken to determine if any of the differences observed were statistically significant. Furthermore, some of the questions allowed multiple responses, but the analysis was only carried out for each response separately. Had a larger data set been obtained, it would have been interesting to investigate the combinations of responses that were most commonly selected (for example, influence over re-sit decision and reasons for re-sitting). Finally, we looked at two A level courses only. It would be useful to know how generalisable the findings are to A levels in the same subjects offered by other exam boards or to other A level subjects. Additional ideas for further research include a longitudinal study. In a few years' time, once the linear A levels have bedded down, data from the present study may prove useful in comparative research.

References

Al-Bayatti, M.F. & Jones, B. (2003). *Statistical study of the differences in candidates' results between first and second attempts in some GCE AS modules*. Paper presented at the British Educational Research Association Annual Conference, September 2003

BBC (2003). *A level students get unlimited re-sits*. Retrieved from http://news.bbc.co.uk/1/hi/education/3160852.stm

BBC (2010). *AQA exam board to bring in exam-only GCSEs in England*. Retrieved from http://www.bbc.co.uk/news/education-11419483]

De Waal, A. (2009). *Straight As? A level teachers' views on today's A levels*. London: Civitas. Available at http://www.civitas.org.uk/pdf/straight_a's.pdf

Department for Education. (2010). *The Importance of Teaching: The Schools White Paper 2010*. London: The Stationery Office.

Department for Education (2014) *Curriculum, exam and accountability reform*. Retrieved from http://www.education.gov.uk/schools/teachingandlearning/qualifications/examsadmin/news/a00221428/curriculum, examandaccountabilityreform

Gill, T. & Suto, I. (2012). *Students' and teachers' views and experiences of A level unit re-sits*. Cambridge Assessment Network Seminar. Retrieved from http://www.cambridgeassessment.org.uk/Images/student-teacher-views-and-experiences-tim-gill-dr-irenka-suto-presentation.pdf

Grimston, J. (2010). Universities reject A level grades obtained in re-sits. *The Sunday Times*, May 30th. Retrieved from http://www.thesundaytimes.co.uk/sto/news/uk_news/Education/article304056.ece

Higton, J., Noble, J., Pope, S., Boal, N., Ginnis, S., Donaldson, R. and Greevy, H. (2012). *Fit for Purpose? The view of the higher education sector, teachers and employers on the suitability of A levels*. Ofqual, Coventry. Retrieved from http://ofqual.gov.uk/documents/fit-for-purpose-the-view-of-the-higher-education-sector-teachers-and-employers-on-the-suitability-of-a-levels/

NASUWT (2008). *Testing to Destruction: Examination and Testing Overload*. Birmingham: NASUWT. Retrieved from http://www.nasuwt.org.uk/consum/groups/public/@education/documents/nas_download/nasuwt_000721.pdf

Ofqual (2013c) *Perceptions of A levels, GCSEs and other qualifications – Wave 11 – summary report*. Office of Qualifications and Examinations Regulation: Crown copyright. Retrieved from http://www.ofqual.gov.uk/files/2013-05-03-perceptions-A levels-GCSEs-and-other-qualifications-wave11-summary-report.pdf

Ofqual (2014) *Changes to qualifications: timeline of reforms*. Retrieved from http://ofqual.gov.uk/qualifications-and-assessments/qualification-reform/

Poon Scott, E. (2010). *Re-sits in high-stakes examinations: the unusual case of A levels*. Paper presented at the 36th International Association for Educational Assessment (IAEA) Annual Conference, August 2010. Retrieved from http://www.iaea.info/documents/paper_4d33f52.pdf

Poon Scott, E. (2012) Short-term gain at long-term cost? How resit policy can affect student learning. *Assessment in Education: Principles, Policy & Practice 19*(4), 431–449.

Qualifications and Curriculum Agency (QCA), (2007a). *A level resitting: summary of research findings*. Retrieved from http://www.ofqual.gov.uk/files/qca-07-3387-Resit-report.pdf

Qualifications and Curriculum Agency (QCA), (2007b). *Evaluation of participation in GCE Mathematics: Appendices A–D.* Retrieved from http://www.ofqual.gov.uk/files/GCE_mathematics_-_Appendices_A_to_D.pdf

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.

Suto, I., Elliott, G., Rushton, N., & Mehta, S. (2011). Going beyond the syllabus: A study of A level Mathematics teachers and students. *Educational Studies, 38*(4), 479–483.

Vidal Rodeiro, C. & Nadas, R. (2010). *Effects of modularisation*. Research report. Cambridge: Cambridge Assessment.

Williams, D.A. (2009). *What has been the impact of re-sitting AS-Level examinations in Economics and Business Studies on students at a boys' independent school in the West Midlands?* PhD thesis, University of Warwick: Coventry

Do Cambridge Nationals support progression to further studies at school or college, to higher education courses and to work-based learning?

Carmen Vidal Rodeiro Research Division

Introduction

The number of students taking vocational qualifications in England has risen dramatically in the last few years (Ofqual, 2012). This can be partly attributed to the growing availability of vocationally orientated/related qualifications aimed at 16 to 19 year-olds. However, whilst in the past the completion of a vocational programme would have been seen as an end in itself, there is now an expectation that all forms of education and training provide progression. In particular, it has been argued that vocational qualifications must be designed to ensure they provide a sufficient platform for progression to higher level of study or to employment (e.g. Bowers-Brown & Berry, 2005; Cowan, 2012; Fuller & Unwin, 2012).

OCR National qualifications, now called Cambridge Nationals, are exam-free, vocationally related qualifications at levels 1 to 3 of the National Qualifications Framework[1] that have an engaging and practical approach to learning and assessment. They are primarily aimed at young people aged 14–19 in full-time or part-time study, although they are also appropriate for adult learners, therefore suiting a wide range of learning styles across the whole ability range. As well as providing practical insight into industry sectors, OCR Nationals help students develop valuable workplace skills, such as team working, communication and problem-solving.

OCR National qualifications have been gaining popularity since their introduction in 2004 (e.g. awards rose from 14,620 in 2006/07 to around 300,000 in 2011/12) and currently around 3,000 education establishments in England are delivering OCR Nationals alongside other qualifications. In fact, more than 1.5 million students of all abilities have been awarded OCR National qualifications over the past few years and the ICT version of the qualification is currently one of the most popular courses in English schools, delivered by more than half of secondary schools. The growth of these qualifications is expected to continue because teachers enjoy teaching them and pupils find them motivating, very relevant and very clear in explaining what is expected of them and what they are trying to achieve (mc^2 market research, 2008; EdComs, 2009).

OCR National qualifications are made of units, which are centre-assessed and externally moderated and as a result, there are no timetabled exams. Candidates receive assessment and learning support throughout the course, giving them a clear indication of their progress, which can increase levels of success and motivation as students can see their own progress through the course, rather than waiting until the end to sit an exam. Furthermore, OCR Nationals offer teachers the flexibility to incorporate work experience, to use their own assignments, and to deliver units in any order. However, some of the OCR National qualifications have been described as having little value and being used simply as a way to take low achievers off academic subjects or to boost schools' league table positions (e.g. Civitas, 2010; Sharp, 2010; Williams & Shepherd, 2010). However, OCR Nationals are a distinctive and important contribution to the 14–19 curriculum. In fact, recent research (mc^2 market research, 2008) provided evidence to support the view that OCR Nationals should have a significant role in 14–19 education. This research consisted of a survey carried out in schools and colleges across the country where the respondents taught or managed the teaching of at least one OCR National qualification. Most respondents said that OCR National qualifications had helped students engage with the subjects in ways that had not been possible before. Furthermore, with the pressure of exams taken off them, the confidence of many students was boosted to allow them to develop themselves. Although it was acknowledged that this did not work for every single student, the overall

1. Each regulated qualification in England has a level between entry level and level 8. Qualifications at the same level are of a similar level of demand or difficulty. To find out more about qualification levels visit http://www.ofqual.gov.uk/help-and-advice/comparing-qualifications/.

view was that OCR Nationals provided opportunities for students who would otherwise be underachievers and/or leaving with lower prospects.

Further research on OCR National qualifications (EdComs, 2009) recommended raising awareness of the progression routes and the qualification value in the sense of affecting employability and progression towards higher education, as employers and Higher Education Institutions (HEIs) do not always place the right value on OCR Nationals due to the poor perceptions of level 2 and level 3 vocational study and also due to the limited understanding of the qualification. In fact, research by Connor *et al*. (2006) found that there was a lack of parity of esteem between vocational and academic qualifications, leading to prejudice against and negative valuing of vocational qualifications. This research also highlighted a need for more knowledge of the content and assessment of vocational qualifications among higher education admissions staff. On the same lines, Carter (2009) reported that universities tend to favour applicants with academic qualifications as opposed to those with vocational qualifications. Similarly, Sinclair and Connor (2008) and Hodgson and Spours (2010) suggested that the potential of vocational qualifications to become a major route to higher education was constrained by their low uptake and the low understanding and recognition of the qualifications.

To date, there have been some attempts to quantify the numbers of young people entering higher education with vocational qualifications (e.g. Connor & Little, 2007; Vickers & Bekhradnia, 2007; Ertl *et al*., 2010). However, those focussed on vocational qualifications as a whole and there is very little information about progression to higher education of learners with specific vocational qualifications, or about progression to further study at school or college, and to work-based learning. Therefore, further evidence regarding the numbers and types of candidates with OCR National qualifications and where they progressed on completion was needed.

The present research set out to investigate if OCR National qualifications enabled successful progression into the labour market (e.g. via work-based learning) and into higher level education. In particular, the research looked at:

- the types of learners who were awarded OCR National qualifications (age, prior attainment, socio-economic background and centre attended);
- the progression of learners with OCR National qualifications in terms of further studies (at school, college or higher education) or work-based learning (e.g. apprenticeships).

Data and methodology

Data

OCR National qualifications are available at levels 1 to 3 of the National Qualifications Framework in a wide range of subjects. The focus of the research presented in this article was on students who were awarded an OCR National qualification in the academic year 2008/09 in subjects listed in Table 1.

At the time this research was carried out, no single dataset existed in England which tracked students over the different stages of their education from completion of compulsory secondary education to completion of an undergraduate degree or a work-based learning programme. Therefore, to investigate the uptake of and the progression from OCR National qualifications, different datasets had to be combined.

Table 1: OCR National subjects included in the research

Level 1	*Level 2*	*Level 3*
Business and ICT	–	–
–	Business	Business
Health and Social Care	Health and Social Care	-
–	–	Health, Social Care and Early Years
ICT	ICT	ICT
–	Media	Media
–	Science	–
–	Sport	Sport
Leisure and Tourism	–	–
–	Travel and Tourism	Travel and Tourism

This work used data from four different data sources: data on OCR National qualifications obtained from the OCR awarding body; data on attainment at school and college obtained from the Department for Education; data on work-based learning programmes obtained from the Learning and Skills Council; and data on students enrolled in HEIs obtained from the Higher Education Statistics Agency. An overview of each of the data sources used in this research is presented below.

OCR National qualifications: Data on OCR National qualifications awarded in the academic year 2008/09 was obtained directly from the OCR awarding body. This data comprised personal characteristics (e.g. name, gender, date of birth) and assessment characteristics (e.g. centre, subject, level, grade, award date) for all students who obtained an OCR National qualification.

National Pupil Database: The National Pupil Database (NPD), which is compiled by the Department for Education, is a longitudinal database for all children in schools in England, linking student characteristics (e.g. age, gender, ethnicity, attendance, and exclusions) to school and college learning aims and attainment. Data for the analyses carried out in this research was extracted from the NPD for the academic year 2009/10.

Individualised Learner Record: The Individualised Learner Record (ILR) contains data for post-16 students in all forms of provision with the exception of schools and is sourced by the Learning and Skills Council. Every college course started is recorded in the ILR. This dataset also records programmes such as apprenticeships and courses offered by non-school learning providers (e.g. carried out in a work-based learning environment or delivered by private/independent learning providers). Data for the analyses carried out in this research was extracted from the work-based learning extract of the ILR dataset for the academic year 2009/10.

Higher Education Statistics Agency: The Higher Education Statistics Agency's (HESA) student record dataset contains students' qualifications prior to starting a higher education course, the course studied, and the

institution where the student was enrolled. For the analyses presented in this article, data on first year undergraduate students in the academic year 2009/10 was provided by HESA[2].

Together, the four datasets mentioned provide information about the gender, age and socio-economic status of learners, as well as the qualifications obtained (or courses enrolled on), and the educational establishments attended at different stages of their education.

Methodology

The analyses presented in this article were carried out in two stages. Stage 1 consisted of an analysis of the entries for OCR National qualifications; and stage 2 looked into the progression from OCR Nationals towards other qualifications in schools, colleges, work-based learning providers, and HEIs.

Stage 1: Entries for OCR National qualifications

The research addressed this issue mainly through a descriptive analysis that looked into candidates' characteristics such as age, prior attainment, socio-economic background, and centre where the qualification was obtained.

Prior attainment: A measure of students' general attainment (proxy for ability) was computed using data from the National Pupil Database.

An average GCSE[3] score was used as a measure of general attainment for students with OCR National qualifications at level 3. By assigning scores to the GCSE grades (A*=8, A=7, B=6, C=5, D=4, E=3, F=2, G=1, U=0) it was possible to arrive to a total GCSE score for each student. An average GCSE score was calculated by dividing the total score by the number of subjects attempted. The mean GCSE score ranges from 0 to 8.

For students with OCR National qualifications at levels 1 or 2, Key Stage 3 scores were used instead. At the end of this stage pupils, generally aged 14, are tested and awarded attainment levels depending on what they are able to do. The tests cover English, Mathematics and Science. The average of the total marks in these three subjects was used as a general attainment measure for candidates with OCR Nationals at levels 1 and 2. The Key Stage 3 score ranges from 0 to 100. The distribution of the students' general attainment measure was used to divide the students into three attainment groups: low, medium and high.

Socio-economic background: The students' socio-economic background was determined by the students' level of deprivation using the Income Deprivation Affecting Children Index (IDACI)[4]. The distribution of this index was used to divide the students into three deprivation groups: low, medium and high.

Type of centre: Centres have been categorised into five different groups: schools; sixth form colleges; further education colleges; adult education and training providers; and 'other'.

Stage 2: Progression from OCR Nationals

This stage of the research investigated the types and numbers of qualifications candidates who obtained OCR Nationals progressed to in:

a. schools and colleges;

b. a work-based learning environment; and

c. HEIs. Descriptive analyses, such as frequency tables and simple descriptive statistics, were used as the main analysis technique.

Schools and colleges: Candidates with OCR National qualifications were identified in the NPD extracts and matched to the qualifications obtained two years later. This permitted identifying the qualifications these candidates progressed to in schools and colleges.

Work-based learning environment: In this case, progression after just one year was investigated. This differs from the progression in schools and colleges due to, principally, two reasons. Firstly, the ILR extracts record enrolments (the NPD only records completed qualifications) and therefore it is not necessary to wait until the end of the programme to know if a candidate is studying towards a work-based learning programme. Secondly, work-based learning programmes can last between one and five years, depending on the particular course and level, and therefore many of the candidates considered in this research would not have had the time to complete the programme and would not be included in the analyses if only completed qualifications were looked at. Candidates with OCR National qualifications in the period of study were matched to the work-based learning extract of the ILR datasets one year later. This permitted identifying the qualifications these candidates progressed to in a work-based learning environment.

Higher Education Institutions: Data on candidates with OCR National qualifications in 2008/09 was linked to the HESA student records dataset. This matching allowed the identification of candidates with at least one level 3 OCR National qualification who enrolled on higher education courses in the academic year 2009/10. For these candidates the following information about their higher education courses was available: highest qualification on entry, subject of study, level of study, and institution.

Results

Entries for OCR Nationals

Overall entries for OCR National qualifications by level and subject are presented in Table 2. This table shows that in the academic year 2008/09, the most popular OCR National qualifications were level 2 qualifications in ICT, Science, Business and Health and Social Care, and the least popular were level 1 OCR National qualifications (in all subjects) and level 3 OCR Nationals in ICT and Travel and Tourism.

It should be noted that entries for level 2 qualifications in ICT and Science have been rising considerably in the last few years (Vidal Rodeiro, 2010a; 2010b) making them by far the most popular OCR National qualifications.

Age of candidates

Figure 1, which displays the age profile of candidates taking these qualifications at each level, shows that, although OCR National qualifications at level 1 are aimed at 14-16 year olds, the majority of the candidates who were awarded a level 1 qualification were at least 17 years old. At level 2, Figure 1 shows that the majority of the

2. Source: HESA Student Record 2009/10. Copyright Higher Education Statistics Agency Limited 2011. HESA cannot accept responsibility for any inferences or conclusions derived from the data by third parties.
3. General Certificate of Secondary Education. This is a qualification taken by the majority of 16 year olds in England.
4. This index is based on the percentage of children in a small area who live in families that are income deprived (in receipt of Income Support, Income based Jobseeker's Allowance, Working Families' Tax Credit or Disabled Person's Tax Credit below a given threshold).

Table 2: Overall entries for OCR National qualifications by subject and level, 2008/09

OCR National subject	*OCR National level*	*2008/09*	
		Candidates	*Percentage*
Business and ICT	1	747	0.45
Health and Social Care	1	481	0.29
ICT	1	302	0.18
Leisure and Tourism	1	194	0.12
Level 1		**1,724**	**1.04**
Business	2	4,799	2.91
Health and Social Care	2	2,930	1.78
ICT	2	138,453	84.02
Media	2	1,269	0.77
Science	2	8,563	5.20
Sport	2	1,198	0.73
Travel and Tourism	2	1,472	0.89
Level 2		**158,684**	**96.30**
Business	3	934	0.57
Health, Social Care and Early Years	3	1,231	0.75
ICT	3	82	0.05
Media	3	817	0.50
Sport	3	774	0.47
Travel and Tourism	3	548	0.33
Level 3		**4,386**	**2.67**
All		**16,4794**	**100.00**

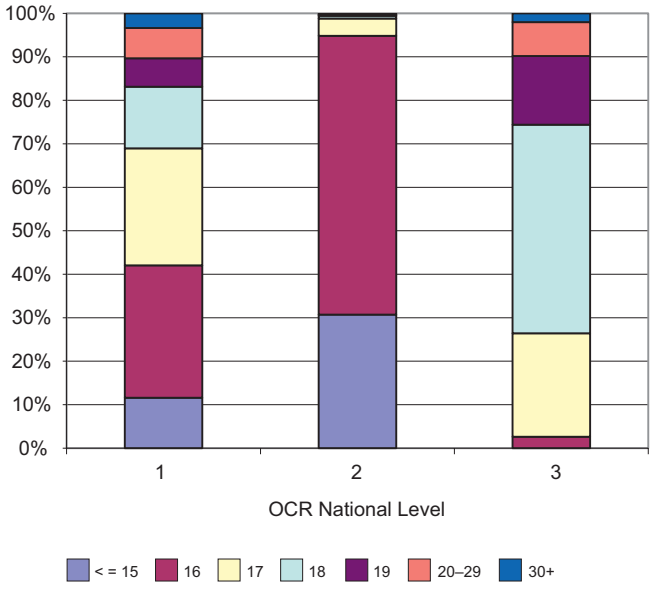


Figure 1: Age profile of candidates by the level of their OCR National qualification

candidates were below 16. In particular, less than 5% of the candidates who obtained a level 2 qualification were 17 or older. At level 3, the majority of the candidates were aged between 18 and 19 years old, and about 10% of the candidates were at least 20 years old. The proportion of candidates at level 1 aged at least 17 years old might suggest that OCR National qualifications provide learners with a second opportunity to find out more about a certain sector, or to introduce themselves to the skills, knowledge and understanding required to prepare for work in a particular area.

Prior attainment of candidates

Figure 2 displays the prior attainment of OCR Nationals candidates. Although OCR National qualifications are designed to suit candidates across the whole ability range, this figure shows that more low ability (prior attainment) candidates than medium or high ability ones obtained OCR National qualifications. However, it should be noted that about a quarter of the candidates who obtained OCR National qualifications had high prior attainment and more than half had medium or high prior attainment. This contrasts with the belief that OCR National qualifications are offered to low ability students instead of other more traditional or academic subjects.

Figure 2, which also shows the prior attainment of candidates by the level of their OCR National qualification, highlights that at levels 1 and 3 the majority of the candidates had low prior attainment. By contrast, at level 2 the percentages of candidates with low or medium prior attainment were not much higher than the percentages of candidates with high prior attainment.

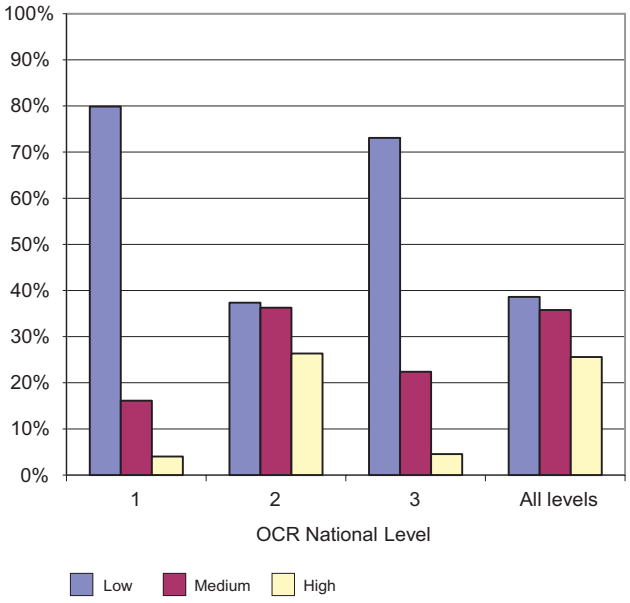


Figure 2: Prior attainment of candidates with OCR National qualifications

Level of deprivation of candidates

It has been claimed that vocational qualifications, such as the OCR Nationals, have been increasingly taken by students from deprived areas. Figure 3, which displays the level of deprivation of OCR National candidates, shows that more highly deprived candidates than medium or low deprived ones obtained OCR National qualifications.

At each level of the qualification, the same patterns of uptake as discussed already are present (that is, more highly deprived candidates than medium or low deprived ones obtained OCR National qualifications). The pattern was slightly more prominent at levels 1 and 3 than at level 2.

Type of centre where the OCR National qualification was obtained

Given the age profile of the candidates with OCR Nationals (see Figure 1), it is not surprising that the overwhelming majority obtained these qualifications in schools (87% of candidates who were awarded an OCR National in the academic year 2008/09 attended schools). Further education and sixth form colleges followed schools as the second and third most popular types of centres where OCR National qualifications

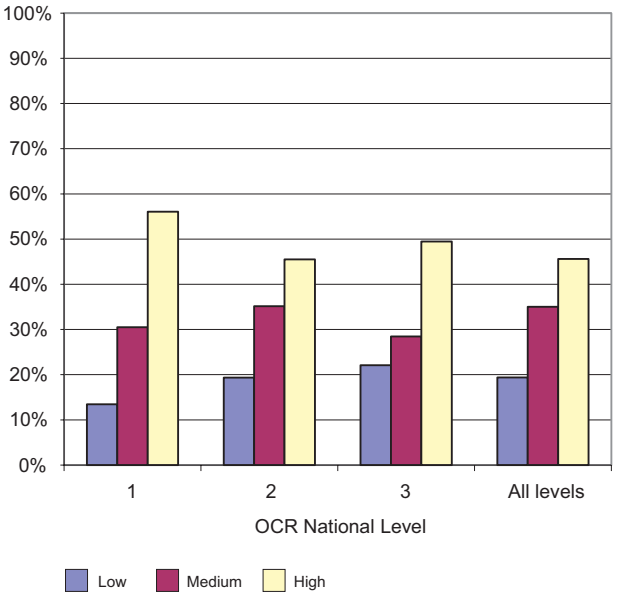


Figure 3: Level of deprivation of candidates with OCR National qualifications

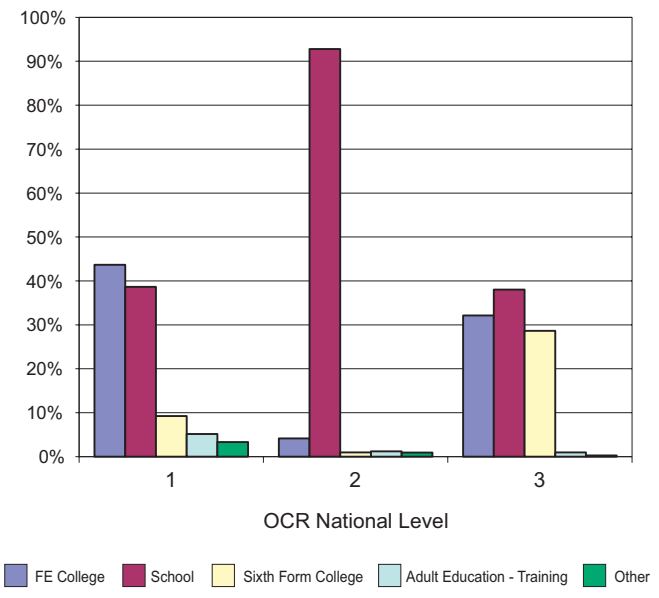


Figure 4: Type of centre where the OCR National qualification was obtained, by the level of the OCR National qualification

Table 3: Progression in schools and colleges from OCR National qualifications at levels 1 and 2 (percentage of students)

OCR National subject	*OCR National level*	*Qualifications in 2009/10*										
		A level	*AS level*	*Applied AS/A level*	*NVQ Level 1&2*	*NVQ Level 3*	*VRQ Level 1&2*	*VRQ Level 3*	*BTEC*	*OCR Nationals*	*DiDA*	*Key Skills*
Business and ICT	1	5.38	11.83	3.23	0.00	0.00	2.15	0.00	3.23	3.23	0.00	0.00
Health and Social Care	1	3.13	5.00	2.50	0.00	0.00	1.25	1.25	5.00	0.00	0.00	0.00
ICT	1	–	–	–	–	–	–	–	–	–	–	–
Leisure and Tourism	1	1.30	3.90	3.90	0.00	0.00	3.90	1.30	6.49	0.00	0.00	0.00
Business	2	27.77	29.69	12.29	0.08	0.08	1.76	0.88	12.37	11.73	0.00	0.24
Health and Social Care	2	13.74	17.82	7.43	1.49	0.62	1.86	2.97	15.10	9.28	0.00	0.00
ICT	2	32.26	36.34	8.91	0.50	0.49	2.25	1.24	14.97	1.98	0.01	0.48
Media	2	20.57	25.96	4.37	0.26	0.51	1.29	1.29	13.37	3.86	0.00	0.00
Science	2	14.60	19.03	8.17	0.52	0.87	2.45	1.74	21.29	1.41	0.02	0.20
Sport	2	17.61	23.06	5.87	1.26	0.84	4.61	0.00	14.68	6.29	0.00	0.00
Travel and Tourism	2	14.13	17.73	9.70	0.28	0.83	1.11	2.77	13.57	3.05	0.00	0.00
Students with no OCR National qualifications		42.13	80.05	11.83	1.55	1.47	15.65	4.93	22.14	4.17	5.11	9.72

were awarded (6% and 2%, respectively). Other types of establishments, such as adult education providers, training organisations, HEIs or prisons, accounted for the remaining 5% of the awards.

Figure 4 shows the type of centre where the OCR National award was obtained by the level of the qualification. It highlights that around 40% of level 1 qualifications were obtained in further education colleges. At level 2, the vast majority of qualifications (over 90%) were obtained in schools. Finally, at level 3 similar percentages of OCR National qualifications were obtained in further education colleges, sixth form colleges and schools.

Progression from OCR Nationals

Schools and colleges

The qualifications to which candidates with OCR Nationals have progressed to in schools and colleges are shown in Table 3. Other qualifications, such as the International Baccalaureate, were taken by very small numbers of candidates and are not presented here.

Very small percentages of candidates with an OCR National qualification at level 1 progressed to further study at school or college, with the most popular qualification being an AS level[5].

Candidates with OCR National qualifications at level 2 progressed predominantly to AS/A levels, Applied AS/A level qualifications, VRQs (vocational-related qualifications) and BTECs[6]. Higher percentages of candidates progressed towards an AS level than towards a full A level. The highest percentage of candidates progressing towards AS/A level qualifications was among candidates with an OCR National in ICT (32%). The next highest percentages were among candidates with an OCR National in Business (about 28%) and with an OCR National in Media (about 21%).

5. AS and A levels are qualifications taken by students between the age of 16 and 18 in England. A levels are usually spaced out over two years and made up of two components: AS and A2 levels. AS levels can stand as a qualification on its own or can be carried on to A2 to complete a full A level qualification.
6. BTEC stands for Business and Technology Education Council which used to award the qualification. BTECs are now awarded by the Edexcel exam board.

Relatively high percentages of candidates with level 2 OCR National qualifications in Business, Health and Social Care and Sport progressed to a level 3 OCR National qualification.

Very few candidates progressed to NVQs (National Vocational Qualifications), DiDA (Diploma in Digital Applications) or Key Skills qualifications in schools and colleges after obtaining an OCR National qualification at levels 1 or 2.

It is essential that secondary school pupils, and in particular those pursuing vocationally-related qualifications, are adequately advised about the implications of the different routes open to them post-16 and, in particular, that taking some of them will open up or close off certain options in the future. In England, around 73% of students who are accepted on to degree courses at HEIs do so with A levels (e.g. Connor, Banerji & Sinclair, 2006). In this context, it is therefore important to know the percentages of candidates with OCR National qualifications at levels 1 or 2 who obtained, two years later, at least three A levels. Table 4 shows that around 30% of the candidates who did not take an OCR National qualification at level 2, had at least three A levels. This figure was much smaller among candidates with OCR National qualifications, ranging from about 5% among candidates with the qualification in Health and Social Care to about 21% among candidates with the qualification in ICT.

Table 4: Percentages of candidates with three or more A levels by the OCR National qualification

OCR National subject	*OCR National level*	*3+ A levels 2009/10*	
		Candidates	*Percentage*
Business and ICT	1	3	3.23
Health and Social Care	1	3	1.88
ICT	1	-	-
Leisure and Tourism	1	1	1.30
Business	2	162	12.93
Health and Social Care	2	36	4.46
ICT	2	11805	20.65
Media	2	40	10.28
Science	2	287	6.23
Sport	2	47	9.85
Travel and Tourism	2	21	5.82
Candidates with no OCR National qualifications		182886	30.51

Work-based learning environment

The following types of programmes in the work-based learning (WBL) sector were considered here: apprenticeship, advanced apprenticeship, and other.

Only 3.8% of candidates who were awarded an OCR National (at any level) in the academic year 2008/09 progressed to work-based learning programmes after completing such qualifications. Table 5 presents the percentages of candidates enrolled in each of the WBL programmes by the OCR National qualification they progressed from.

Progression rates varied greatly by subject. The highest percentages of candidates with level 1 OCR National qualifications progressing towards work-based learning and towards apprenticeships in particular, obtained the OCR National in Health and Social Care. The highest percentages of candidates with level 2 OCR National qualifications progressing towards work-based learning, and towards apprenticeships in particular, obtained the OCR National in Health and Social Care, Sport or Science. At level 3, candidates who obtained OCR National qualifications in Business, Health, Social Care and Early Years, or Travel and Tourism were the most likely ones to be enrolled on a WBL programme, in particular on a programme of study leading to an apprenticeship.

Table 5: Progression to work-based learning programmes from OCR National qualifications at levels 1 to 3 (percentage of students)

OCR National subject	*OCR National level*	*WBL programme in 2009/10*			*Total*
		Apprentice-ship	*Advanced Apprentice-ship*	*Other*	
Business and ICT	1	2.68	0.27	1.20	4.15
Health and Social Care	1	7.28	0.00	2.08	9.36
ICT	1	1.66	0.33	0.33	2.32
Leisure and Tourism	1	2.06	0.00	0.52	2.58
Business	2	2.79	0.31	0.88	3.98
Health and Social Care	2	5.49	0.48	2.59	8.56
ICT	2	2.60	0.51	0.43	3.54
Media	2	1.73	0.55	0.24	2.52
Science	2	4.61	0.55	0.60	5.76
Sport	2	3.76	1.17	0.58	5.51
Travel and Tourism	2	2.79	0.34	1.02	4.15
Business	3	3.21	0.21	1.50	4.92
Health, Social Care and Early Years	3	2.92	1.14	1.30	5.36
ICT	3	1.47	0.24	0.49	2.20
Media	3	1.47	0.24	0.49	2.20
Sport	3	2.20	0.65	0.39	3.24
Travel and Tourism	3	2.92	0.36	0.73	4.01
All		2.77	0.51	0.52	3.80

Higher Education

Previous research on progression from vocational courses into higher education (e.g. UKCES, 2010) provided evidence of all types of vocational qualifications being recognised for entry purposes by HEIs[7], although often when achieved alongside other qualifications. However, recognition of vocational qualifications is more extensive amongst the Post-92 universities and, not surprisingly, vocational qualifications are more likely to lead to vocational degrees (Hoelscher *et al.*, 2008).

Research carried out by the OCR awarding body (EdComs, 2009) showed that students with OCR National qualifications were progressing to higher education and that teachers of these qualifications felt that the students were well-prepared for that level of study. Furthermore, case studies showed that having OCR Nationals opened higher education access to students who had previously not considered it suitable for them. It should be noted that OCR Nationals are unlikely to be the sole qualification taken as part of a young person's programme of learning, and students could take, for example, A levels or other VRQs alongside them.

At the end of the academic year 2008/09, there were 4,386 candidates with at least one level 3 OCR National qualification in the subjects considered in this research. Of those, 1865 (42.5%) were enrolled in a higher education course in the following academic year, 2009/10. This

7. Post-92 universities are former polytechnics or colleges of higher education that were given the name 'university' in 1992.

Table 6: Participation in HE courses by the subject of the level 3 OCR National qualification[8]

OCR National subject (at level 3)	*Higher education entries (2009/10)*		*Same/related subject area at HEI*	
	Candidates	*Percentage*	*Percentage (out of those with an OCR National and in HE)*	*Percentage (out of those with an OCR National)*
Business	480	51.28	67.43	34.58
Health, Social Care and Early Years	480	38.99	63.33	24.70
ICT	25	28.05	'..'	'..'
Media	375	45.90	33.87	15.54
Sport	325	41.99	59.38	24.94
Travel and Tourism	185	33.39	59.02	19.71
Overall	1865	42.52	56.94	24.21

Table 7: Highest qualification on entry for students registered in an HEI by the subject of level 3 OCR National qualification (percentage of students)

OCR National subject (at level 3)	*GCE A level or A level equivalents*	*OCR National Diploma*	*OCR National Extended Diploma*	*OCR National Certificate*	*Other qualifications*
Business	18.58	36.12	7.10	0.00	38.20
Health, Social Care and Early Years	22.50	38.13	14.79	0.00	24.58
ICT	'..'	'..'	'..'	'..'	'..'
Media	9.33	38.67	20.00	0.00	32.00
Sport	18.46	44.31	11.69	0.00	25.54
Travel and Tourism	26.23	32.24	7.10	0.00	34.43

Table 8: HEI by the subject of the level 3 OCR National qualification (percentage of students)

OCR National subject (at level 3)	*Russell Group*	*1994 Group*	*University Alliance*	*Million+ Group*	*UKADIA*	*Other*
Business	3.76	3.76	34.24	35.70	0.84	21.71
Health, Social Care and Early Years	6.25	4.58	29.38	34.38	0.00	25.42
ICT	'..'	'..'	'..'	'..'	'..'	'..'
Media	1.07	4.27	37.33	35.73	5.07	16.53
Sport	1.85	3.38	31.38	29.85	0.31	33.23
Travel and Tourism	1.64	3.83	43.17	31.15	1.64	18.58
Overall	3.32	3.97	33.78	34.05	1.55	23.32

Table 9: Level of study at HEIs of candidates with a level 3 OCR National qualification (percentage of students)

OCR National subject (at level 3)	*First Degree*	*Other undergraduate*	*Higher National Diploma (HND)*	*Higher National Certificate (HNC)*
Business	84.55	10.44	4.80	0.21
Health, Social Care and Early Years	66.46	33.13	0.42	0.00
ICT	'..'	'..'	'..'	'..'
Media	93.07	5.33	1.60	0.00
Sport	90.15	7.69	2.15	0.00
Travel and Tourism	83.06	15.30	1.64	0.00
Overall	82.25	15.39	2.31	0.05

section of the article focuses on this group of candidates. It should be noted though that in 2009/10 there were 208,170 first year student enrolments on HE courses (HESA, 2011) and therefore the percentage of students in HEIs who had completed OCR National qualifications at level 3 was small (0.9%).

Table 6 shows numbers and percentages of candidates in HE courses by the subject of the OCR National qualification they held. The highest participation rate was among students progressing from an OCR National qualification in Business. The lowest participation rate was among those progressing from an ICT qualification, closely followed by those progressing from a qualification in Travel and Tourism.

Table 6 presents some encouraging figures for the OCR National qualifications, as in subjects such as Sport, Media or Business, over 40% of the students progressed towards a course in an HEI.

Table 6 also shows the percentages of candidates who progressed to an HE course in the same/related subject of their level 3 OCR National

qualification. In all subjects, with the exception of ICT and Media, more than half of the candidates who enrolled into an HEI did so in a course in the same (or related) subject as their OCR National qualification. Furthermore, around a quarter of the students with OCR National qualifications in Business, Sport or Health, Social Care and Early Years were enrolled in HE courses in the same area.

Candidates with OCR National qualifications registered at HEIs could have obtained other qualifications at level 3 alongside their OCR Nationals (e.g. A levels, NVQs, BTEC). It is therefore important to investigate how many of those candidates had an OCR National qualification as the highest qualification on entry. It should be noted though that a student's highest qualification on entry is not necessarily that which was required for entry to the programme of study. Table 7 shows that higher percentages of candidates with OCR National qualifications and registered for an higher education course had the OCR National as the highest qualification on entry rather than having other qualifications. The OCR National qualification with the highest percentage of students having it as the highest qualification on entry was the qualification in Sport, followed by the qualifications in Business and in Health, Social Care and Early Years. The OCR National qualification with the lowest percentage of students having it as the highest qualification on entry was Travel and Tourism.

There are many different classifications of HEIs. For the purpose of this research, the following groups[9] were considered: The Russell Group, The 1994 Group, University Alliance, The Million+ Group, UKADIA and Other.

The most popular destination for candidates with OCR National qualifications were HE institutions in the Million+ Group and University Alliance. Institutions in UKADIA Group, followed by institutions in the Russell Group and the 1994 Group were the least common among candidates with OCR National qualifications (Table 8). It should be noted that the choice of institution could have been influenced by the type of course/degree that the candidate wanted to pursue. In fact, many 'selecting' universities in the Russell or 1994 Groups do not offer the range of vocational and work-based courses that are likely to be of interest to the vocational learners.

Table 9 shows the level of study of candidates with level 3 OCR National qualifications. The overwhelming majority of students were registered for a first degree, with percentages being higher among students with OCR National qualifications in Media or Sport. The highest percentages of students registered on other undergraduate courses were among those with OCR National qualifications in Health, Social Care and Early Years and Travel and Tourism. Small percentages of students registered for an HND[10] at an HEI. Those percentages were higher among students with an OCR National qualification in Business and lower among students with a qualification in Health, Social Care and Early Years.

8. Numbers of students have been rounded to the nearest multiple of 5 throughout the article and percentages calculated on groups which contain 52 or fewer individuals were suppressed and represented as '..', following HESA's rounding strategy.

9. Some universities formed groups through which they share ideas and resources regarding issues and procedures in the higher education sector. For this research, HESA provided the following university groups: The Russell Group, The 1994 Group, University Alliance, The Million+ Group and UKADIA. Universities that have not joined any of these groups were included in a separate group, labelled Other. A list of members of each group can be obtained from the websites of each group.

10. HNCs and HNDs are work-related (vocational) higher education qualifications. They are designed to give students the skills to put knowledge to effective use in a particular job.

Conclusions and discussion

This research aimed to gather detailed information about the learners enrolled on OCR National qualifications and where they progressed on completion. In brief summary, it showed that OCR National qualifications enable learners to progress in a variety of ways (to further studies at school or college, to work-based learning and to higher education) and therefore are an important contribution to the 14–19 curriculum.

Uptake of OCR Nationals

In recent years, the popularity of OCR National qualifications has really taken off. In the academic year 2011/12 around 300,000 learners were awarded OCR National qualifications, compared to around 13,000 in 2006/07. Although these qualifications are available in a broad range of subjects, the most popular ones were ICT, Business, Science and Health and Social Care.

Similarly to work carried out by Carter (2009) and UKCES (2010), this research shows that the majority of those learners tend to come from lower socio-economic groups. This has therefore implications for progression. A White Paper by the Panel on Fair Access to the Professions (2009) noted that more than twice as many young people from lower socio-economic groups choose vocational routes as do young people with parents in professional occupations. UKCES (2010) suggests that encouraging vocational progression to higher level learning is fundamental to social mobility and that better support for individuals on vocational pathways, who have the aspirations and ability to achieve higher level skills, should have a positive impact on social mobility.

Although OCR National qualifications are designed to suit candidates of all abilities, more low ability candidates than medium or high ability ones were awarded these qualifications. This contrasts with the belief that OCR National qualifications are being offered to low ability students instead of other more traditional or academic subjects. Nevertheless, it should also be noted that about a quarter of the candidates who obtained OCR National qualifications had high prior attainment and more than a half had medium or high prior attainment.

The number of centres offering OCR National qualifications has also been increasing rapidly over time. In particular, there were 137 educational establishments delivering OCR Nationals in 2006/07 and 1,693 in 2008/09. In the latter academic year, 87% of the centres were schools, 6% were further education colleges and about 2% were sixth form colleges. Other types of establishments accounted for the remaining 5%.

Progression routes from OCR Nationals

The question of what learners go on to do following the completion of their OCR National qualifications was a crucial one in this research. Overall, this research showed that OCR Nationals enabled progression to further study at schools or colleges and at university. There were also many instances when learners progressed towards work-based learning programmes and, in particular, apprenticeships.

In terms of progression at school or college, very few candidates with OCR National qualifications at level 1 progressed to further studies. Candidates with OCR National qualifications at level 2 progressed predominantly to AS/A and Applied AS/A level qualifications, VRQs at different levels and BTECs.

The research showed that there appears to be a reasonably consistent

pattern of students carrying on with the OCR National subject from level 2 to level 3, particularly in Business, Sport, Travel and Tourism and Health and Social Care.

It should be noted that previous research (EdComs, 2009) reported that some teachers felt that students could struggle with the demands of A levels even after succeeding at level 2 OCR Nationals. This could have resulted in students being advised to continue on a vocational route at level 3.

The numbers of candidates with OCR National qualifications who progressed to work-based learning programmes were small in comparison to the numbers progressing to AS/A level or to higher education courses. However, apprenticeships and other work-based learning programmes offer a clear route into employment and OCR National candidates progressed to these programmes. This might have been due to the fact that schools delivering OCR Nationals have connections with the local community and employers and therefore learners can be encouraged to progress to work-related programmes. It should be noted that apprenticeship opportunities for young people have been quite limited in recent years (e.g. Bowers-Brown & Berry, 2005). However, apprenticeship numbers are set to grow considerably in future years as the Apprenticeships, Skills, Children and Learning Act, (2009) provided a statutory entitlement to apprenticeships for suitable qualified 16 to 18 year-olds in England from 2013.

In terms of progression to higher education, whilst the route from academic qualifications to full-time degree programmes is one that is well defined and well respected in England, the route from vocational and applied qualifications is less clear and one that far fewer individuals follow. In particular, previous research (e.g. Hoelscher *et al*., 2008) has shown that A levels provide the major access route into university, in particular, into the most prestigious ones and that students with vocational backgrounds are more likely to start their higher education studies at Post-92 universities, if at all. However, there is also evidence (e.g. UKCES, 2010) of all types of vocational qualifications being recognised for entry purposes by all types of HEIs, although those are often achieved alongside other qualifications.

The present research showed that relatively high percentages of candidates with OCR National qualifications at level 3 enrolled on courses in HEIs. In subjects such as Sport, Media or Business over 40% of these students progressed to a course in an HEI and, in the majority of cases, the candidates enrolled on a course in the same subject area as their OCR National qualification.

However, OCR Nationals are unlikely to be the sole qualification taken as part of a young person's programme of learning and students will take, for example, A levels or VRQs alongside them. Nevertheless, for the majority of the candidates who progressed to higher education, having taken OCR Nationals, the OCR National qualification was the highest qualification on entry.

With regards to the institution attended, Carter (2009) argued that vocational progression routes are often best developed in the newer parts of the higher education sector. Many Post-92 universities, further education and higher education colleges have rich experience in developing learning programmes and recruitment procedures that are tailored to the needs of vocational learners. This research confirms the above argument as the most popular destinations for candidates with OCR National qualifications were HEIs in the University Alliance, followed by institutions in the Million+ Group, which are constituted by the newest universities and colleges. Institutions in the Russell Group and the 1994 Group were the least common among candidates with OCR National qualifications. However, about 6% of the candidates still enrolled on courses in institutions in these two groups.

UCAS research into vocational progression to higher education (Papageorgiou, 2007) revealed that whilst 93% of the higher education institutions give information about entry requirements for applicants with academic qualifications such as A levels, much smaller percentages do so for applicants with vocational qualifications such as the OCR Nationals. Carter (2009) suggests that the admissions process should be made less daunting for applicants with vocationally-related qualifications and UKCES (2010) asks for more robust information to be made available about how many and what types of learners progress from vocational education to higher level skills and that this information is used to plan provision.

In the current educational climate it is important to be clear about the value that the OCR National qualifications bring to learners in terms of their future progression. On these lines, there are plans to publish detailed data on the destinations of school leavers in England. In fact, the Department for Education has recently published 'Experimental Statistics' on education destination measures which show the percentage of students progressing to further learning in a school, further education or sixth form college, apprenticeship, work-based learning provider, or higher education institution (DfE, 2012).

This practice has long been in place in Scotland and Northern Ireland (ONS, 2010; DENI, 2011), where each year information is released on the destinations of school leavers (e.g. higher education, further education, employment) and on the highest level of qualifications obtained by school leavers. In the meantime, this research provides evidence that OCR Nationals are valuable qualifications and indeed support progression to further learning at school or college, to work-based learning and to university.

Limitations

There were a number of limitations regarding the data used for this research. Firstly, most of the data interrogated in the analyses presented in this article are routinely collected for administrative rather than for research purposes. Therefore, although the data is a rich source of information, it is limited in different ways (e.g. Davies, Barnes and Dibben, 2010). For example, these data are constantly evolving to meet the needs of the people and organisations who primarily use them and to accommodate the changing policy and aims of the data holding bodies. Furthermore, new variables are added and changes to definitions are made over time.

Secondly, linking between candidates with OCR National qualifications and candidates recorded in the ILR dataset was carried out using candidates' full name and date of birth. Similarly, data on candidates with OCR Nationals was linked to the HESA student records dataset using a process of 'fuzzy' matching on name, date of birth, gender and, where available, location of school. Matching in these ways is not perfect and it would have been impossible to achieve a 100% matching rate. Therefore, some candidates who obtained an OCR National qualification and were pursuing a work-based learning programme or were enrolled on higher education courses might not have been included in the analyses.

Thirdly, analyses using data from HEIs must adhere to the HESA standard rounding methodology in order to ensure that no data where

living individuals can be identified are published. In particular, percentages based on 52 or fewer individuals had to be suppressed and therefore it was not possible to fully report on the progression towards HE courses of candidates with level 3 OCR National qualifications in Media and ICT.

Other limitations/issues in relation to the analysis presented in this article must also be acknowledged. Firstly, it is important to recognise that progression to further studies or employment cannot be attributed solely to the OCR National qualifications as, in most cases, learners will have completed other qualifications alongside their OCR Nationals.

Secondly, it should be noted that following the Wolf Review of Vocational Education (Wolf, 2011), new criteria are to be set from 2014 for vocational courses, including OCR National qualifications, to be included in the performance tables (see e.g. BBC News, 2011; or DfE, 2011). The patterns of uptake of OCR National qualifications reported in this research can therefore change in the near future.

References

BBC News (2011). *Many vocational courses axed from league tables*. BBC News, 20 July. Retrieved from http://www.bbc.co.uk/news/education-14218920

Bowers-Brown, T. & Berry, D. (2005). Building pathways: Apprenticeships as a route to higher education. *Education+Training, 47,* 270–282.

Carter, J. (2009). *Progression from vocational and applied learning to higher education in England*. Bolton: University Vocational Awards Council.

Civitas (2010). *Unqualified Success: Investigating the state of vocational training in the UK*. London: CIVITAS, The Institute for the Study of Civil Society.

Connor, H., & Little, B. (2005). *Vocational ladders or crazy paving? Making your way to higher levels*. London: Learning and Skills Development Agency.

Connor, H., Banerji, N., & Sinclair, E. (2006). *Progressing to higher education: vocational qualifications and admissions*. Ormskirk: Action on Access.

Cowan, T. (2012). Is there a universal right to higher education? *British Journal of Educational Studies, 60*(2), 111–128.

Davies, J., Barnes, H. & Dibben, C. (2010). *Education administrative data: exploring the potential for academic research*. St Andrews: Administrative Data Liaison Service.

DENI (2011). *Qualifications and destinations of Northern Ireland school leavers 2009/10*. Bangor: Department of Education.

DfE (2011). *Wolf Review of Vocational Education: Government's response* (DfE-00038-2010). London: Department for Education.

DfE (2012). *Destinations of Key Stage 4 and Key Stage 5 pupils, 2009/10*. London: Department for Education.

EdComs (2009). *OCR Nationals: 'Voice of the market research'*. London: EdComs Ltd.

Ertl, H., Hayward, G. & Hoelscher, M. (2010). Learners transition from vocational education and training to higher education. In David, M. (Ed.). *Improving learning by widening participation in higher education*. London: Routledge.

Fuller, A. & Unwin, L. (2012). *Banging on the door of the university: the complexities of progression from apprenticeship and other vocational programmes in England*. Cardiff: Skope Publications.

Her Majesy's Stationery Office (2009). *Apprenticeships, Skills, Children and Learning Act*.

HESA (2011). *Statistical First Release 153: Higher education student enrolments and qualifications obtained at higher education institutions in the UK for the academic year 2009/10*. Cheltenham: Higher Education Statistics Agency.

Hodgson, A. & Spours, K. (2010). Vocational qualifications and progression to higher education: the case of the 14–19 Diplomas in the English system. *Journal of Education and Work*, *23*(2), 95–110.

Hoelscher, M., Hayward, G., Ertl, H. & Dunbar-Goddet, H. (2008). The transition from vocational education and training to higher education: a successful pathway? *Research Papers in Education, 23*(2), 139–151.

mc² market research (2008). *OCR 'Nationals' and the future*. Nottingham: mc² market research Ltd.

Ofqual (2012). *Statistics bulletin. vocational and other qualifications*. Coventry: Office of Qualifications and Examinations Regulation.

ONS (2010). *Destinations of leavers from Scottish schools 2009/10*. Newport: Office for National Statistics.

Panel on Fair Access to the Professions (2009). *New Opportunities White Paper*. London: Department for Business, Innovation and Skill.

Papageorgiou, J. (2007). *Progression to higher education for applicants with vocational qualifications*. Cheltenham: Universities and Colleges Admissions Service.

Sinclair, E., & Connor, H. (2008). *University admissions & vocational qualifications: two years on*. Ormskirk: Action on Access.

Sharp, H. (2010). *Vocational education has 'lost its way'*, says Gove. BBC News, 9 September. Retrieved from http://www.bbc.co.uk/news/education-11229469

UKCES (2010). *Progression from vocational and applied learning to higher education across the UK: A comparative study*. London: UK Commission for Employment and Skills.

Vickers, P. and Bekhradnia, B. (2007). *Vocational A levels and university entry: is there parity of esteem?* Oxford: Higher Education Policy Institute.

Vidal Rodeiro, C. L. (2010a). *Provision of science subjects at GCSE*. Statistics Report Series No 15. Cambridge: Cambridge Assessment.

Vidal Rodeiro, C. L. (2010b). *Uptake of ICT and computing qualifications in school in England 2007–2009*. Statistics Report Series No 25. Cambridge: Cambridge Assessment.

Williams, R. and Shepherd, J. (2010). GCSE results: university crisis to hit school students, union warns. *The Guardian*, 24 August. Retrieved from http://www.guardian.co.uk/education/2010/aug/24/university-crisis-gcse-students?INTCMP=SRCH.

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

An investigation of the effect of early entry on overall GCSE performance, using a propensity score matching method

Tim Gill Research Division

Introduction

A report by Gill (2013) found that certain groups of students performed worse than expected in some GCSE subjects when they were taken early, even taking into account any improved performance from re-sitting. In particular, high attaining students (those achieving level 5 at Key Stage 2 [KS2] tests in the subject) were less likely to achieve a grade A in GCSE English or GCSE Mathematics if they took the exam early (even if they re-sat at the expected time). However, it may be that one reason for taking an exam early is to 'get it out of the way' to enable increased focus on other subjects in Year 11. An Ofsted survey (Ofsted, 2013) asked schools their reasons for entering students early and 44% responded that they did so "to allow students to focus on other subjects". Furthermore, schools were asked what they felt the benefits of early entry were given their experience, and 51% responded "the freed time allowed students to do better in other subjects".

If early entry leads to better than expected performance in the other exams then the overall impact of early entry may not be detrimental and could even be advantageous. This article investigates this issue by looking at whether students entering early for GCSEs perform better or worse across all their GCSEs (or equivalents) than those who do not enter for any GCSEs early.

Data and methods

The data for this analysis came from the National Pupil Database (NPD) for 2011. This is a database of student level attainment and personal characteristics compiled by the Department for Education from data supplied by centres and awarding bodies. The Key Stage 4 (KS4) extract, which records all attainment by students who are at the end of KS4, was used. The database includes exams taken by these students in previous years, meaning it was possible to identify early entry.

To compare the overall GCSE performance of early entry students with non-early entry students, three different outcome measures were used:

1. Mean GCSE score. This was calculated by transforming each GCSE grade into a number (A*=8, A=7 etc.) and then generating a mean value for each student. The grade used was the best grade attained in each subject (i.e. after re-sits).
2. Indicator of whether or not the student passed the statutory target of five or more GCSEs (or equivalents) at A* to C including English and Mathematics. This is an important accountability measure for schools, and is used in school performance tables. The outcome measure to compare between the different groups was therefore the percentage of students passing this threshold.
3. Total KS4 points score. For all KS4 qualifications a score is allocated to each grade (for example, an A* grade at GCSE is worth 58 points, an A grade 52 points and a B grade 46 points)[1]. The total points score is the sum of the points received on all qualifications taken by the candidate. This was included as an alternative to the GCSE mean score because it gives more value to a candidate with, for instance, nine A* grades than one with eight A* grades. This might be an important difference in particular circumstances (e.g. allowing more options at A level).

Early entry was defined as having taken the exam for the first time prior to starting Year 11. This means that students taking an exam for the first time in January of Year 11 were not considered to be early entry. Students taking the qualification early and then re-sitting in Year 11 were counted as early entry, despite the fact that some of them will not have had more time in Year 11 to focus on other subjects (as hypothesised). However, counting these students as not early entry would potentially have been more problematic because of the way in which they would have been 'allocated' to this group. Had we done this, it is likely that anyone who didn't achieve at least a C would probably be entered again, thus moving from an early entry group to a non-early entry group. In effect this could mean that the outcome measure (GCSE performance) determines which group students are in, which would invalidate the analysis. Students taking fewer than five GCSEs were excluded from the analysis.

Propensity score matching

A propensity score matching method was used in this research (see Caliendo & Kopeinig, 2008; Morgan & Harding, 2006). This method is useful when we have a 'treatment', and want to compare the outcomes for a 'treated' group with those of a 'non-treated group' but we are not able to randomly assign people to the groups. For this research, treatment refers to early entry in at least one GCSE, and the outcome refers to each of the three performance measures detailed above. In theory, to know for certain the effect of a treatment, we would need to compare the outcomes for the same participants with and without treatment at the same time. In practice this is not possible, so other methods are necessary. The treated and non-treated groups could just be compared in terms of their mean outcomes, but this would not be comparing like with like because of differences between the two groups in terms of background characteristics (covariates). The propensity score method attempts to overcome this by manipulating the data such that the treated and non-treated groups are made similar enough for comparisons between the groups to be valid.

There are a number of different ways of doing this, the most common of which is to 'match' each individual in the treated group with one (or more) individuals in the non-treated group in terms of covariates. However, this is a computationally demanding method when dealing with large data sets and so a different method was employed here, involving

1. For a full list of qualifications and points scores visit http://register.ofqual.gov.uk/Qualification, enter the qualification and click on "View performance measures".

the creation of subgroups in the treated and non-treated groups such that the members of each subgroup were very similar in the two different groups in terms of covariates. Weights were then used to compensate for the imbalance of treated and non-treated individuals in each subgroup. This method is now described in some detail.

First, it was necessary to identify individuals in each group who were similar in terms of covariates. To do this, individuals were classified by their 'propensity' for being in the treated group. A logistic regression model was run, with being in the treated group (i.e. early entry) as the dependent variable and all the covariates of interest as independent variables. The coefficients from this model allowed us to estimate the probability an individual with any particular set of background characteristics would be in the treated group. This probability is referred to as the propensity score. Groups of students with similar propensity scores are very likely to be similar in terms of their background characteristics.

Once the propensity score measure was calculated, individuals were classified into ten subgroups[2], based on their propensity score. Thus, subgroup 1 consisted of those with the lowest propensity score (lowest probability of being treated) and subgroup 10 those with the highest propensity score (highest probability of being treated). The equivalent subgroups in the treated and non-treated groups should now have been similar in terms of their background characteristics, enabling comparisons to be made. However, within each subgroup the balance of the number of treated and non-treated individuals was not even (particularly in groups 1 and 10), and thus it was also necessary to apply weights to the non-treated individuals to account for this imbalance.

Following the application of the weights, the distribution of covariates in the treated and non-treated group should have been approximately the same. This was checked to make sure that the weighting had worked correctly. Then, the outcome variable (weighted in the non-treated group) was compared in the treated and non-treated groups. The results for the weighted non-treated group could be thought of as the outcome for the treated group *had they not been treated*. Using the technical language, we were estimating the *average treatment effect for the treated* (ATT).

This method was applied in a number of different situations. The first of these had just one treated group (entering early for at least one GCSE). The same method was then applied to a situation with two treatments, either taking one GCSE early or two or more GCSEs early. For this analysis the principle was the same but the method was modified in two important ways. Firstly, a different method for generating propensity scores was used. This was necessary because using logistic regression with two treatment groups generated propensity scores that meant the groups were not well-matched on covariates. Instead, a Generalised Boosted Model (GBM) was used to generate the propensity scores. GBMs use an automated, data-adaptive algorithm to estimate a smooth function, by adding together a large number of simple functions (see McCaffrey, Ridgeway & Morral, 2004, for an example application to propensity score estimation). They are flexible because they allow the function being modelled to be non-linear, and generate propensity scores that are well-matched to the empirical probability of treatment.

The second modification was in the meaning of the propensity score, which now referred to a student's propensity for being in the *non-treated* group. Students were then classified by their propensity score into 15 subgroups and, in contrast to the single treatment situation, data from the *treated* groups was weighted to match the non-treated group in terms of the number of students in each subgroup. This means that the results for the weighted, treated groups can be thought of as the outcome for the non-treated group *had they been treated*. In the technical language, we are estimating the *average treatment effect for the non-treated* (ATNT).

Finally, both the single treatment and two treatment models described were applied to subgroups of students to see if there were different treatment effects within different groups. Students were classified by three variables; gender, school type and prior attainment. The methods described were then applied to each subgroup in turn.

Subgroup analyses

To investigate the effect of prior attainment, students were classified by their mean KS2 level across the three tests (English, Mathematics and Science) into three approximately equally sized groups. Normal KS2 levels range from 2 to 5 and students given either a level 'B' ('Working below the level assessed by the test') or 'N' ('No test level awarded') were allocated a level 1 so that these results could be included in the calculation of their mean KS2 level.

For the school type analysis, the schools that students attended were classified into four types – comprehensives (including academies and free schools), grammar schools, independent schools and secondary modern schools.

Covariates

For the logistic regression models only the covariates that had a statistically significant effect on the probability of being in the treated group were included. These were selected from the following list:

- Total number of GCSEs taken
- Number of other qualifications taken
- Number of BTECs taken
- Number of OCR Nationals taken
- Prior attainment (as measured by KS2 levels in English, Mathematics and Science)
- Deprivation measure (IDACI)
- Ethnicity
- Gender
- Age
- School type

These variables were chosen because they were available in the NPD, and were potentially influential in determining both a student's likelihood of entering early and the outcome measures.

When calculating the propensity scores, only variables with a statistically significant parameter estimate were included in the final logistic regression model. Furthermore, when analysing by subgroup, the relevant subgroup variable was excluded from the model. So, for instance, in the analysis of comprehensive students the school type variable was excluded because all students were in the same category. For other subgroups some variables were missing for all students, so these were excluded. For example, for the analysis of independent school students', ethnicity and Income Deprivation Affecting Children Index (IDACI) score were removed because these are not recorded in the NPD for these students.

2. According to Rosenbaum and Rubin (1983), subclassification into five subclasses is enough to remove 90% of the bias for many distributions, so ten subclasses should be more than sufficient.

Results

Data exploration

Table 1 presents a breakdown of the number of GCSEs entered for early by students in the 2011 cohort taking at least five GCSEs in total. Thus, almost 38% of students entered early for at least one GCSE, with most of those just entering one early (25%). Only a very small minority entered for three or more GCSEs early.

Table 1: Number of GCSEs entered early

Number of early entry	*Students (n)*	*Students (%)*
0	328,246	62.7
1	130,738	25.0
2	43,286	8.3
3	13,704	2.6
4	3,919	0.8
5+	3,635	0.7

Figures 1 and 2 present the distribution of mean GCSE for early entry and non-early entry students (and for different numbers of GCSEs entered early).

Overall, students entering at least one GCSE early had a higher mean GCSE (5.33) than those who did not enter any GCSEs early (5.16). Figure 3 shows that students who entered early for one, two or three GCSEs had the highest mean GCSE scores (5.34, 5.33 and 5.33 respectively).

However, this analysis takes no account of differences in the background characteristics of students in each category. If these characteristics have an impact on the variable of interest (GCSE mean grade) then it is important to account for any differences in them between the groups.

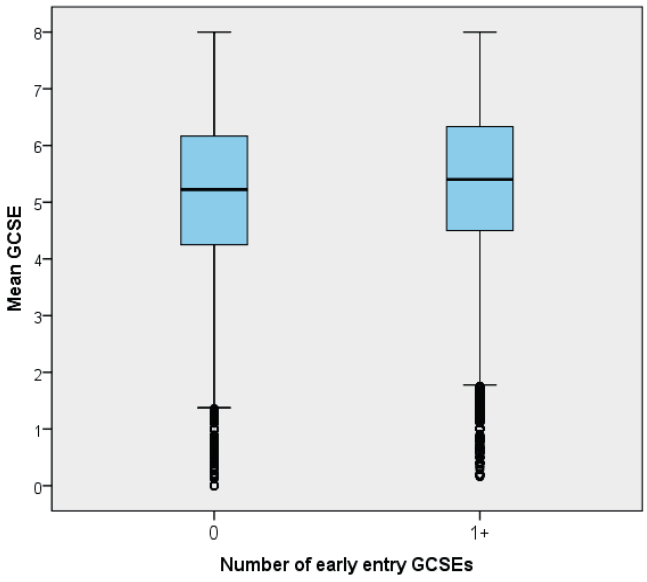
Analysis 1 – Mean GCSE, single treatment group

The first set of analyses compared the mean GCSE scores for those taking at least one GCSE early with those not taking any early. It should be noted that there were fewer students in this analysis than in Figure 1 because it was not possible to estimate a propensity score for students with missing data for any of the covariates used in the logistic regression model. For example, there were a substantial number of students (40,759) with missing KS2 levels across all three tests (mainly from independent schools). For these analyses there were 453,421 students of which 38.2% entered early for at least one GCSE.

Checking quality of matching

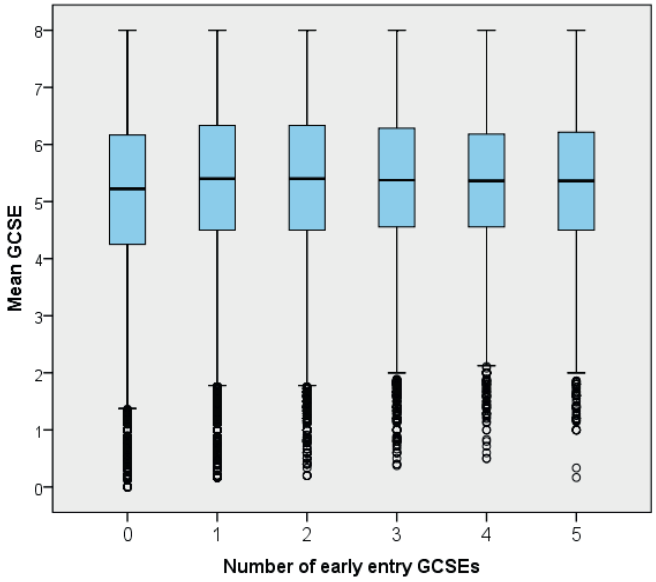
Before presenting the results for each of the analyses in terms of the outcome variable (GCSE mean), it is important to check the quality of the process undertaken to match the treated and non-treated group in terms of their covariates. This was done by comparing the mean values for all covariates between the non-treated and treated groups, before and after weighting. It is normally recommended that statistical tests are undertaken to check whether there are significant differences and, should any be found, the model may need to be re-specified. However, because of the very large sample sizes involved, statistical tests (e.g. t-test) are highly likely to come out as significant even if the differences are very small. Thus, the approach used here was to note any particularly large differences and take them into account when interpreting the results.

For the sake of brevity the full results of this checking are not presented here. However, we note that for this analysis the matching worked very well for all the covariates, with the values for the weighted non-treated group very close to the values for the treated group.



	0	*1+*
Mean	5.16	5.33
Std Dev.	1.41	1.38
N	328,246	195,282

Figure 1: Distribution of mean GCSE scores for early entry and non-early entry students



	0	*1*	*2*	*3*	*4*	*5*
Mean	5.16	5.34	5.34	5.33	5.29	5.26
Std Dev.	1.41	1.38	1.39	1.35	1.32	1.36
N	328,246	130,738	43,286	13,704	3,919	3,636

Figure 2: Distribution of mean GCSE scores by number of early entry GCSEs

Table 2: Mean GCSE performance of early entry and non-early entry groups

Analysis	*Subgroup*	*No of students*	*% early entry*	*GCSE mean (T)*	*GCSE mean (NT–UW)*	*GCSE mean (NT–W)*	*Prob*
Main		453,421	38.2	5.28	5.03	5.27	0.4598
Gender	Female	223,024	39.0	5.50	5.26	5.48	0.5074
	Male	215,973	37.8	5.19	4.95	5.18	0.7363
Prior attainment	High	107,946	45.8	6.44	6.40	6.44	0.9600
	Medium	160,455	39.1	**5.40**	5.37	**5.43**	**0.0314**
	Low	188,234	33.0	4.21	4.12	4.22	0.6532
School type	Comp	414,712	38.0	5.24	4.97	5.23	0.7576
	Independent	23,426	22.2	6.57	6.47	6.52	0.4186
	Selective	20,538	36.7	6.75	6.62	6.74	0.8995
	Secondary Modern	16,912	46.9	4.92	4.69	4.94	0.7907

Comparison of means

The results in terms of the mean GCSE variable are shown in Table 2. This shows the number of students in each subgroup, the percentage taking at least one GCSE early and the mean of the mean GCSE variable for the treated (T) and non-treated groups (both un-weighted [NT-UW] and weighted [NT-W]). The difference in means between the treated and non-treated (weighted) groups was assessed with a test of statistical significance[3]. Statistically significant differences are highlighted in bold.

In terms of GCSE mean the students in the treated group performed better than those in the non-treated group (un-weighted), both overall and in each of the subgroups. However, after weighting these differences almost disappeared and in three of the subgroups the students in the non-treated group performed better. There was only one statistically significant difference, for students in the medium prior attainment group, where students entering early had a significantly lower mean GCSE (5.40) than those not entering early (5.43). However, this difference was very small (only 1/33rd of a grade), or the equivalent to 0.3 of a grade in one GCSE for someone who took ten GCSEs in total.

Analysis 2 – Mean GCSE, two treatment groups

The second analysis investigated whether there was an effect of different numbers of early entries on GCSE performance. Students were classified into groups based on how many GCSEs they entered for early (none, one, or two or more). Table 3 presents the number and percentage of students in each group.

Table 3: Number of students in early entry groups

No of early exams	*Number of students*	*% of students*
0	280,121	61.8
1	115,257	25.4
2+	58,098	12.8
Total	453,476	

3. The 'Surveyreg' procedure in SAS was used to test for differences in the means. This accounts for the effect of clustering of students within schools.

Thus, the majority of students did not take any exams early (61.8%) and about 13% took two or more. As before, analyses were undertaken comparing the performance for the whole cohort of students and then separately for students in each subgroup.

Checking quality of matching

In contrast to the first analysis, the data from the treated groups (one early entry and two or more early entry) were weighted to match the data from the non-treated. The quality of this procedure was checked by comparing the mean values for all covariates between the non-treated and treated groups, before and after weighting.

Overall, the matching was very good, and there were no issues at all between the first treated group (one early entry) and the non-treated group. However, for two of the subgroups the matching between the second treated group (two or more early entry) and the non-treated group was not ideal on all variables. Specifically, for the high attaining subgroup, there was a mismatch on the school type variable after weighting, with 76.10% of the second treated group attending a comprehensive school, compared with 71.68% of the non-treated group. For the selective school subgroup the matching was poor on the gender variable after weighting, with 58.69% of the second treated group being female, compared with 50.83% of the non-treated group. Furthermore, for this subgroup, 75.50% of the second treated group were white, compared with 81.15% of the non-treated group. Therefore, we need to acknowledge these differences when interpreting the results for these subgroups.

Comparison of means

Table 4 presents a comparison of mean GCSE scores for the non-treated group (NT), the treated, un-weighted group (T-UW) and the treated, weighted group (T-W). Separate rows in the table compare the students in the treated groups (one GCSE early or two or more GCSEs early) with those not taking any GCSEs early. Thus for the analysis of all students ('Main') the non-treated group had a mean GCSE of 5.10. The mean for the group taking one GCSE early was 5.29 (un-weighted) and 5.07 (weighted). The figures for the group taking two or more GCSEs early were 5.30 (un-weighted) and 4.98 (weighted).

Again, statistical tests were undertaken to assess whether differences in the mean between the non-treated group and the treated, weighted groups were significant.

Table 4: Mean GCSE performance of (multiple) early entry and non-early entry groups

Analysis	*Subgroup*	*No of early entry*	*No of students*	*GCSE mean (NT)*	*GCSE mean (T–UW)*	*GCSE mean (T–W)*	*Prob*
Main		1	126,562	5.10	5.29	5.07	0.2968
		2+	63,288	**5.10**	5.30	**4.98**	**0.0087**
Gender	Female	1	63,899	5.27	5.44	5.24	0.3948
		2+	33,647	**5.27**	5.46	**5.17**	**0.0332**
	Male	1	62,633	4.93	5.13	4.90	0.3422
		2+	29,641	**4.93**	5.12	**4.80**	**0.0079**
Prior attainment	High	1	33,820	6.49	6.50	6.52	0.2251
		2+	18,521	**6.49**	6.43	**6.41**	**0.0266**
	Medium	1	42,985	**5.42**	5.44	**5.39**	**0.0166**
		2+	21,649	**5.42**	5.37	**5.30**	**<0.0001**
	Low	1	43,613	4.15	4.23	4.13	0.3615
		2+	19,917	**4.15**	4.18	**4.00**	**0.0002**
School type	Comp	1	111,805	4.95	5.20	4.93	0.4433
		2+	55,321	4.95	5.23	4.87	0.0624
	Independent	1	3,971	6.45	6.58	6.49	0.573
		2+	1,320	6.45	6.47	6.37	0.534
	Selective.	1	4,807	6.67	6.81	6.68	0.931
		2+	3,178	6.67	6.72	6.63	0.845
	Secondary Modern	1	5,406	4.58	4.90	4.57	0.909
		2+	3,115	4.58	4.76	4.38	0.151

For the analysis using all data ('Main') there was a statistically significant difference in means for the second treated group only (p=0.0087). The performance of the non-treated group (5.10) was better than the second treated group (4.98) after weighting had been applied. Similar results were also found in each of the subgroups, with the differences being statistically significant except in the school type subgroups. In each case the performance of the non-treated group was better than the second treated group after weighting. The differences varied from 0.07 of a grade (comprehensives) to 0.20 of a grade (secondary moderns). This suggests that early entry (of two or more subjects) had a negative impact on overall performance at GCSE, for students overall and for several of the subgroups that were analysed.

Comparing students in the first treated group with those in the non-treated group, the differences were very small. The only statistically significant difference was for students in the medium attaining group, where students in the non-treated group performed better (5.42) than those in the treated group (5.39).

Analysis 3 – Accountability measure, single treatment group

This analysis is with a single treatment group, but with the outcome measure being the percentage of students passing the school accountability target of five or more GCSE grades A* to C including English and Mathematics. As before, this analysis was done for all students and then each of the subgroups. The same propensity scores were used as in the GCSE mean analyses, so there was no need to check the quality of matching.

Comparison of percentages

Table 5 compares the percentage of students passing the threshold measure in the treated group with the percentage[4] passing in the non-treated groups (weighted and un-weighted). A test of statistical significance was undertaken of the difference in percentage between the treated group and the weighted non-treated group.

There is a clear pattern in these results with a significantly higher percentage of students in the treated group passing the threshold measure than those in the non-treated group (after weighting), overall and in most of the subgroups. Amongst all students, 73.77% of the treated group passed, compared with 70.97% of the non-treated group.

The exceptions to this pattern were in the low attaining and the independent school groups, where a significantly higher percentage of the non-treated group passed (62.55% and 91.10% respectively) than the treated group (57.97% and 88.73% respectively).

These results suggest that there may have been some advantage in schools entering students early for some GCSEs, in terms of getting more students to pass the threshold measure (except for low attaining students and those in independent schools).

Analysis 4 – Accountability measure, two treatment groups

This analysis investigated whether there was an effect of different numbers of early entries on the percentage of students passing the accountability measure.

Comparison of percentages

Table 6 compares the percentage of students passing the target measure in the non-treated group with the percentage passing in the two treated groups.

4. The 'surveylogistic' procedure in SAS was used to test for differences in the proportions, taking into account the clustering of students within schools.

Table 5: Threshold measure success rate of early entry and non-early entry groups

Analysis	*Subgroup*	*No of students*	*% early entry*	*% passing (T)*	*% passing (NT–UW)*	*% passing (NT–W)*	*Prob*
Main		453,421	38.2	**73.77**	63.19	**70.97**	**<0.0001**
Gender	Female	223,024	39.0	**76.61**	66.75	**73.84**	**<0.0001**
	Male	215,973	37.8	**72.20**	61.95	**69.99**	**<0.0001**
Prior attainment	High	107,946	45.8	98.49	97.69	98.42	0.5747
	Medium	160,455	39.1	**85.23**	80.39	**83.62**	**<0.0001**
	Low	188,234	33.0	**57.97**	66.62	**62.55**	**<0.0001**
School type	Comp	414,712	38.0	**73.06**	61.89	**70.31**	**<0.0001**
	Independent	23,426	22.2	**88.73**	90.72	**91.10**	**0.0325**
	Selective	20,538	36.7	99.23	98.55	99.35	0.6485
	Secondary Modern	16,912	46.9	66.81	54.59	64.13	0.2613

Table 6: Threshold measure success rate of (multiple) early entry and non-early entry groups

Analysis	*Subgroup*	*No of early entry*	*No of students*	*% passing (NT)*	*% passing (T–UW)*	*% passing (T–W)*	*Prob*
Main		1	126,562	64.22	71.60	64.60	0.5910
		2+	63,288	64.22	76.62	65.84	0.1570
Gender	Female	1	63,899	66.73	73.54	67.21	0.5335
		2+	33,647	66.73	78.82	68.90	0.0825
	Male	1	62,633	61.67	69.63	61.97	0.7220
		2+	29,641	61.67	74.12	62.57	0.5030
Prior attainment	High	1	33,820	97.68	98.24	97.60	0.7230
		2+	18,521	97.68	98.75	97.29	0.1510
	Medium	1	42,985	81.14	84.18	81.21	0.8800
		2+	21,649	81.14	87.46	81.99	0.2990
	Low	1	43,613	**34.23**	40.51	**36.10**	**0.0052**
		2+	19,917	**34.23**	45.93	**37.28**	**0.0096**
School type	Comp	1	111,805	61.25	70.31	62.21	0.1936
		2+	55,321	**61.25**	75.82	**64.50**	**0.0085**
	Independent	1	3,971	90.13	88.74	88.95	0.3279
		2+	1,320	**90.13**	86.89	**85.37**	**0.0348**
	Selective	1	4,807	98.82	99.11	98.53	0.5224
		2+	3,178	98.82	99.34	97.36	0.0943
	Secondary Modern	1	5,406	51.54	63.41	52.31	0.7870
		2+	3,115	51.54	65.14	50.79	0.8430

For the main analysis and most of the subgroup analyses there was very little difference in the percentages of students passing the threshold after weighting, although the treated groups tended to do slightly better. There was only one subgroup with a statistically significant difference between the non-treated and first treated group. This was for the low attaining students, with 36.10% of the first treated group achieving the threshold compared with 34.23% of the non-treated group (p=0.0052). For this subgroup, students in the second treated group were also significantly more likely to achieve the threshold than the non-treated group (37.28%, p=0.0096). There were two other subgroups with significant differences between the second treated group and the non-treated group. For comprehensive school students, 64.60% of the second treated group achieved the threshold, compared with 61.25% of the non-treated group. In contrast, a lower percentage of independent school students in the second treated group achieved the threshold (85.37%) than those in the non-treated group (90.13%).

These results suggest that there seemed to be little advantage for those taking just one GCSE early (except for low attaining students), and the advantage for those taking two or more GCSEs early was limited to comprehensive school students and low attainers. Independent school students were disadvantaged if they took two or more GCSEs early.

These results are somewhat at odds with the results for the single treatment group (Table 5), which had significant differences in most of the subgroups and larger differences in percentage of students passing. This finding is discussed further in the conclusion.

Analysis 5 – Total points score, single treatment group

For the final two analyses the outcome measure was the total points score, across all GCSEs and equivalents. As before this analysis was undertaken for all students and then each of the subgroups. Again, the same propensity scores were used as in the GCSE mean analyses, so there was no need to check the quality of matching.

Comparison of means

Table 7 compares the mean total points score in the treated and non-treated groups.

For the main analysis and each subgroup analysis the treated group had a higher mean total points score than the non-treated (weighted) group. This difference was statistically significant in the main analysis and in the female, male and comprehensive schools subgroups. However, the differences were not large, being about 6 points, equivalent to one GCSE grade in one GCSE.

This suggests that students entering early for some GCSEs, whilst not doing significantly better on their GCSEs (see Table 2), tend to perform better on the GCSE equivalent qualifications, leading to a higher total points score.

Analysis 6 – Total points score, two treatment groups

Table 8 presents the results of the analysis of mean total points score with two treatment groups.

The differences between the first treated group and the non-treated

Table 7: Mean total points score of early entry and non-early entry groups

Analysis	*Subgroup*	*No of students*	*% early entry*	*Mean points total (T)*	*Mean points total (NT–UW)*	*Mean points total (NT–W)*	*Prob*
Main		453,421	38.2	**540.9**	474.0	**534.4**	**0.0183**
Gender	Female	223,024	39.0	**557.5**	491.2	**551.2**	**0.0318**
	Male	215,973	37.8	**528.3**	463.1	**522.3**	**0.0424**
Prior attainment	High	107,946	45.8	629.3	568.3	624.6	0.1185
	Medium	160,455	39.1	549.7	496.8	547.3	0.3809
	Low	188,234	33.0	458.4	411.4	453.8	0.1137
School type	Comp	414,712	38.0	**538.2**	470.6	**532.3**	**0.0393**
	Independent	23,426	22.2	483.0	455.5	479.1	0.5723
	Selective	20,538	36.7	628.5	557.1	623.6	0.6991
	Secondary Modern	16,912	46.9	518.4	467.8	515.6	0.7758

Table 8: Mean total points score of (multiple) early entry and non-early entry groups

Analysis	*Subgroup*	*No of early entry*	*No of students*	*Mean points total (NT)*	*Mean points total (T–UW)*	*Mean points total (T–W)*	*Prob*
Main		1	126,562	470.77	520.12	469.44	0.5587
		2+	63,288	**470.77**	575.42	**463.96**	**0.0459**
Gender	Female	1	63,899	485.15	533.78	484.13	0.6730
		2+	33,647	485.15	589.73	479.75	0.1590
	Male	1	62,633	456.25	506.18	454.72	0.5536
		2+	29,641	**456.25**	559.18	**447.12**	**0.0184**
Prior attainment	High	1	33,820	559.08	605.57	560.31	0.6090
		2+	18,521	559.08	660.27	556.22	0.4680
	Medium	1	42,985	493.24	531.52	491.52	0.4180
		2+	21,649	**493.24**	582.47	**485.70**	**0.0178**
	Low	1	43,613	409.72	445.52	408.90	0.7373
		2+	19,917	**409.72**	489.71	**399.58**	**0.0075**
School type	Comp	1	111,805	468.51	519.37	467.46	0.6680
		2+	55,321	468.51	575.43	464.44	0.2660
	Independent	1	3,971	453.37	479.31	453.76	0.9520
		2+	1,320	453.37	487.83	442.76	0.3230
	Selective.	1	4,807	565.06	607.53	561.68	0.6810
		2+	3,178	565.06	667.01	571.03	0.6550
	Secondary Modern	1	5,406	459.59	502.23	452.73	0.4743
		2+	3,115	459.59	526.22	435.87	0.0528

group were generally very small, and none were statistically significant. However, students in the non-treated group did perform significantly better than those in the second treated group, for the main analysis and in three of the subgroup analyses (males, medium attainers and low attainers). In the main analysis, the difference was around 7 points, equivalent to just over one grade in a GCSE. The differences were slightly larger amongst males (9 points), medium attainers (7.5 points) and low attainers (10 points). There was also a considerably larger difference in the secondary modern subgroup (24 points), although this was not quite large enough to be statistically significant. Thus, students entering early for two or more GCSEs seemed to be disadvantaged in terms of their overall KS4 points score.

Again, these results are somewhat at odds with the results with a single treatment group (Table 7), which had students in the treated group performing better, on average, than those in the non-treated group. This finding is discussed further in the next section.

Discussion

The purpose of this research was to investigate whether students who take one or more GCSEs before Year 11 perform any differently than those not doing so, across all GCSEs (and equivalents). Many schools enter some students early for at least one GCSE and this may be for a number of reasons, such as trying to get students over the crucial threshold grade C, or getting some qualifications out of the way to allow students to focus on other subjects in Year 11. Students may also be entered early in order to give them a chance to re-sit if they do not do as well as expected.

By looking at three different measures of success at GCSE it was possible to investigate the effect of early entry on individual students and also on schools' performance through the percentage of students passing the important threshold measure of five grades A* to C including English and Mathematics.

For individual students there does not seem to be any advantage in early entry in terms of overall GCSE performance. Comparing those taking at least one GCSE early with those not doing so showed there to be almost no difference in mean GCSE. Indeed, Table 4 showed that for students taking two or more GCSEs early there was a significant disadvantage compared with non-early entry students. The analysis of the whole cohort found a difference of 0.12 of a grade, equivalent to one grade lower in one GCSE for a student taking eight GCSEs. This significant effect was also present for all the subgroups investigated apart from the school type subgroups.

However, it seems there may be some advantage in early entry if we consider other measures of performance. Early entry students had a statistically significantly higher mean total points score than those not taking any GCSEs early (Analysis 5). This amounted to 6.5 points in the cohort as a whole, equivalent to about one GCSE grade (which is, perhaps, not a large difference over ten or more GCSEs, but significant nonetheless). This difference was present in all subgroups (although only significant in the male, female and comprehensive school subgroups). When the analysis was limited to students with two or more early entry GCSEs a different pattern emerged, with these students tending to perform worse on this measure than those not taking any GCSEs early. This was the case for all students taken together and amongst males, low attainers and medium attainers (Analysis 6).

When looking at the percentage of students passing the threshold measure (Analysis 3), students in the early entry group performed significantly better (73.77% passed) compared with the non-early entry students (70.97% passed). This was also the case for most of the subgroups. In the low attaining and independent school subgroups, however, the early entry students performed significantly worse than non-early entry students. When comparing students with different numbers of early entry GCSEs the positive effect of early entry on the threshold measure was only present for low attainers and for comprehensive school students taking two or more early (Analysis 4).

It should be noted that there is an important difference in the interpretation of any differences between treated and non-treated groups depending on whether we are talking about the single treatment or multiple treatment case, because of the different weighting methods in each case. In the single treatment case we were estimating the average treatment effect for the treated (ATT). This was done by comparing the actual results for the treated group with the estimated results for the treated group *had they not been treated* (the weighted, non-treated group). In contrast, the two treatment case involves estimating the average treatment effect for the non-treated (ATNT). This is done by comparing the actual results for the non-treated group with the estimated results for the non-treated group *had they been treated* (the weighted, treated groups). So any observed differences are only really relevant for the group in question.

This distinction may be the reason why there was a positive effect of early entry when looking at the single treatment group (in terms of accountability measure and total points score) and a less positive (for accountability measure) or even negative (for total points score) effect when looking at the two treatment groups. In other words, the effect of early entry seems to have been more positive (on average) for the treated, than it would have been for the non-treated. This suggests that, to a degree, teachers are correctly choosing the early entry students as those most likely to benefit from it.

It is interesting that early entry seems to be successful in getting a larger percentage of students to pass the threshold measure, but it is not better for individual students (at least in terms of GCSE mean). This apparent contradiction is presumably because early entry is more successful in getting students around the C boundary to improve their grade than getting A and B grade students to perform to their potential. These results also corroborate the findings from previous studies (Gill, 2013) that high attaining students are least likely to benefit from early entry (in individual subjects).

It is also interesting that independent school students who take exams early have a higher mean GCSE than those that don't (although this difference is not significant), but are significantly less likely to pass the threshold measure, suggesting that independent schools' focus is on individual students rather than the threshold measure (which they are not judged on).

One interesting hypothesis that may be worthy of further research is whether students who were disadvantaged by early entry in individual subjects (e.g. high performing students in English and Mathematics) are able to make this up in Year 11. From the results presented here we cannot know whether this is the case because we do not identify students who performed below expectations on their early entry exams. All we can say is that, on average, students who take at least one GCSE early are not disadvantaged in terms of overall GCSE, and actually perform better in terms of overall points score. In contrast, it is estimated that those who did

not enter early would have performed worse if they had taken two or more GCSEs early. Further research could also estimate the average treatment effect for the treated in the case of two treatment groups, to see if taking two or more GCSEs early is beneficial to these students or not.

Finally, it will be interesting to see the impact of GCSE reforms on the amount of early entry. Students will still be able to sit GCSEs in Year 10, but changes to accountability measures mean that only the result from the first sitting of a GCSE will count in performance tables. This is likely to lead to a fall in early entry because schools may want to wait until students are ready to achieve their best possible grade, rather than getting them to sit GCSEs early and then re-sit if they underperform.

References

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72.

Gill, T. (2013). Early entry GCSE candidates: Do they perform to their potential? *Research Matters: A Cambridge Assessment Publication, 16*, 23–40.

McCaffrey, D.F., Ridgeway, G., & Morral, A.R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*(4), 403–425.

Morgan, S.L., & Harding, D.J. (2006). Matching Estimators of Causal Effects: Prospects and Pitfalls in Theory and Practice. *Sociological Methods & Research, 35*(1), 3–60.

Ofsted (2013). Schools' use of early entry to GCSE examinations. Its usage and impact. Manchester: Ofsted.

Rosenbaum, P.R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Big data and social media analytics

Vikas Dhawan and **Nadir Zanini** Research Division

Introduction

'Big data' is fast becoming an area of great importance for businesses in many areas, including education. In simple terms it refers to the combination of data from various sources and understanding patterns in the data which can be used for various purposes such as improving market intelligence and educational research. Businesses, large and small, are implementing (or planning to implement) big data strategies. Apart from market intelligence, it is being applied in diverse areas such as healthcare and other scientific research, complex manufacturing industries such as aviation and heavy machinery, improving public utilities and traffic management, oil and gas exploration, telecoms, retail, banking and insurance, defence and security.

In this article we give an introduction to big data and some of its applications in various fields, including education. We also describe the use of big data for the monitoring of social media (for instance LinkedIn, Facebook and Twitter) for market growth and brand management. Some training courses in big data offered by various universities are mentioned in the article.

Applications in the education industry mentioned in this article include the combination of various sources of information about pupils such as test records, behaviour patterns, and teacher observations over a period of time for providing more accurate and timely interventions. In addition to this, we discuss new forms of assessment such as e-assessment and adaptive testing which will provide new streams of data which could be tapped for studying the performance of test takers in more detail and for monitoring and evaluation of tests.

Big data

Technological advances in recent years have led to a significant amount of data which is now generated in everyday life, such as shopping, travelling, banking, manufacturing and trading, public utilities, state and governance, sports, entertainment, science, education and health. Commercial organisations, research bodies and governments have started to realise the importance of using this data for their growth. As a result, the study of big data has gained prominence among scholars in different areas of research (Einav & Levin, 2013; Mayer-Schönberger & Cukier, 2013) as well as generating interest from the non-academic world (BBC, 2013; Lohr, 2012).

The concept of big data encompasses the collection of data, the combination of the data collected from various sources, processing it and using the results so obtained. Specifically, big data is a term used for large databases requiring complex processing and visualisation which cannot be efficiently handled by traditional data processing software (Wikipedia, 2014a). According to the McKinsey Global Institute, "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika *et al*., 2011). A well-known model (known as 3V's model) of big data attributed to Gartner Inc. defines it as "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" (Beyer & Laney, 2012). The term 'volume' here indicates the complexity of datasets and not necessarily their size. 'Variety' refers to the different type of structured or unstructured data such as text and

numeric, video and audio and log files. 'Velocity' refers to the speed with which data can be made available for analysis. Sometimes other V's such as 'Veracity' (aiming at data integrity and the ability of the organisation to confidently use the data) or 'Value' (does new data enable an organisation to get more value?) are highlighted as well (Swoyer, 2012; Villanova University, 2014).

The rising potential of big data has led to the funding of several new initiatives by governments in recent years. The European Union has recently launched the Big Data Public Private Forum (called the BIG Project) to engage with academics, companies and other stakeholders to formulate a clear strategy for research and innovation in big data. The outcomes of the project will be used as input for Horizon2020 – an initiative aimed at securing Europe's global competitiveness and creating new growth and jobs in Europe (BIG, 2014; European Commission, 2014). The US Government announced funding of $200 million for the "Big Data Research and Development Initiative" in 2012 which aims to gain insights from large and complex collections of data in the fields of science and engineering, national security and teaching and learning (Kalil, 2012). The United States National Security agency is constructing a data centre in Utah to handle information they collect over the internet. There may be some concerns over privacy related to this development because it might result in the collection of personal data of individuals, such as internet access history, private communications, credit card usage and health records, etc.

The amount of data which is expected to be processed (not stored) at the facility in Utah is likely to be in 'yottabytes' – the largest unit prefix in the International System of Units (SI) and which was added in 1991. One yottabyte (prefixed as YB) is equivalent to 10^{24} bytes. Table 1 gives the data storage units in use. Gigabyte is still the most commonly used measure for the capacity of hard disk, however terabyte and petabyte have started to be used as well. Today a 1-terabyte disk drive (about 2.5 inches wide) can fit within a laptop. It is fascinating to note that, according to one estimate, storing a yottabyte on terabyte sized drives would require a million city block size data-centres, as big as the US states of Delaware and Rhode Island (Wikipedia 2014c; 2014d; Diaz, 2010).

Table 1: Data storage units (Wikipedia, 2014d).

Metric prefixes (multiples of bytes)

Decimal		*Binary*		
Value	*Metric*	*Value*	*JEDEC*[1]	*IEC*[2]
1000	kB Kilobyte	1024	KB Kilobyte	KiB kikibyte
1000^2	MB megabyte	10242	MB Megabyte	MiB mebibyte
1000^3	GB gigabyte	10243	GB Gigabyte	GiB gibibyte
1000^4	TB terabyte	10244	TB Terabyte	TiB tebibyte
1000^5	PB petabyte	10245		PiB pepibyte
1000^6	EB Exabyte	10246		EiB exbibyte
1000^7	ZB zettabyte	10247		ZiB zebibyte
1000^8	YB yottabyte	10248		YiB yobibyte

1. Joint Electron Device Engineering Council memory standards
2. International Electrotechnical Commission units

This gives an idea of how much traffic is likely to flow through the internet in the coming years, and the investment being made by governments (and private organisations) realising the potential impact of this data revolution (Wikipedia, 2014a).

According to CompTia (The Computing Technology Industry Association), in 2013, 28% of UK companies were using big data, 36% were planning a big data initiative that year and 95% see data as crucial to success over the next two years (Raconteur Media, 2013). They also report that there was a 5% annual global growth in IT spending in 2013 compared to a 40% growth in data. There has been a phenomenal explosion of data available from online usage in recent years. According to some estimates (IBM, 2013):

- 1.43 billion people worldwide visited a social networking site in 2012;
- nearly one in eight people worldwide have their own Facebook page;
- one million new accounts were added to Twitter every day in 2012;
- three million new blogs come online every month;
- 65% of social media users say they use it to learn more about brands, products and services.

The amount of data collected in organisations is expected to grow in the coming years. This could be due to an increase in the efficiency and declining cost of data storage and processing capabilities, the spread of digital technologies, and volume of data available from internet and digital devices and sophistication of algorithms for processing. A significant amount of this data would be generated online which would require substantial investment in data storage facilities. It has been recently reported that Facebook is currently building a data centre in Sweden the size of 11 football fields, along with two others in America, to collect and process their data (Bradbury, 2013).

There is a considerable amount of interest in educational organisations in exploiting the applications of big data and analytics, which is expected to rise in the near future. However, in order to make the most of big data, organisations should be clear about what exactly they want to investigate and how they plan to use the information. We believe that businesses need to consider the following questions while implementing big data/ social media policies:

1. Are we future ready?
2. Is it hype or necessity?
3. Are there any simpler and/or more economical ways of getting similar results?
4. Is it better to develop in-house capability or hire external resource?
5. Would our customers/stakeholders be comfortable with such monitoring?
6. Do we need to disseminate our policy to the stakeholders? If yes, have we done that?
7. What is the state of preparedness of our competitors?
8. Are we adhering to the data privacy laws?
9. How much value can be placed on the online behaviour of people?
10. Are we also using traditional sources of information (such as interviews and focus groups) to complement online metrics?
11. Are we also relying on human judgement for interpreting the data (and not only on software-generated results)?
12. Are we working with other departments within the organisation to develop a comprehensive policy?

Applications of big data

There are many examples of how big data is being used in various fields. Whilst these are not directly associated with the field of education, they give us a picture of the impact of data in our day-to-day lives (Raconteur media, 2013). Examples include:

- **IBM's Deep Thunder weather analytics package:** helps farmers know when to irrigate their crops;
- **SAS:** uses big data to identify fraud in the insurance sector;
- **British Airways' Know Me Programme:** uses the data collected to get a better insight into personal preferences and buying patterns of its frequent fliers;
- **Transport for Greater Manchester:** uses real-time traffic information to avoid congestion on roads;
- **Bank of America Merrill Lynch:** creates practical and effective solutions for clients based on a more comprehensive and holistic understanding of their requirements;
- **East Kent Hospitals University NHS Foundation Trust:** staff given access to data to adapt to real-time changes such as re-allocation of doctors and nurses between sites based on changes in demand across sites;
- **Citi:** estimates targeted predictive analytics according to customer behaviour;
- **Public Health England:** creates highly targeted treatments according to how patients respond in real-time through recently announced national cancer database (the data contains 11 million historical records and 350,000 new entries added every year);
- **Ocado:** delivers groceries purchased online. It keeps track of vehicle location, driving styles and petrol consumption while delivering 1.1 million items every week;
- **Royal Dutch Shell:** spends £650 million a year compiling big data across a number of sites so that they can more accurately predict presence of hydrocarbon resources at a site – this may help save them drilling costs (which for a single offshore drilling can cost up to £65 million);
- **Accenture:** collects social media analytics for the purposes of sentiment analysis by using data and text mining, semantics, linguistics and syntax processing;
- **Facebook:** recently started to decode the content of photographs (identifying faces and objects) and video;
- **Apple:** granted a patent to collect data on body temperature and heart rate through audio buds;
- **Google:** tunes algorithms in language processing to be culturally relevant (for instance differentiating between American and British idioms) and also improving its speech recognition capabilities;
- **Temetra:** collates information on how people use gas and water in their homes and businesses, giving them data after every 15 minutes rather than an annual reading;
- **Modak Analytics:** mined about 18 terabytes of data of a 810 million electorate during the general elections in India held in April to May 2014 on various demographics such as gender, age, and economic status for their client, a political party (Kurmanath, 2014).

An interesting application of the use of big data in developing government policy is the Behavioural Insights Team (www.behaviouralinsights.co.uk) which is jointly owned by the UK Government and Nesta www.nesta.org.uk. This organisation brings together data from a range of inter-related academic disciplines (Behavioural Economics, Psychology, and Social Anthropology) to understand how individuals make decisions in practice and how they are likely to respond to options so as to enable the Government to design its policies or interventions accordingly.

Applications of big data in education

A large amount of data is being generated in schools and higher education. Big data in education could be used to:

- understand performance and behaviour patterns of students;
- keep track of student progress throughout their education, allowing timely intervention if any anomalies are noticed;
- develop personalised content and instructional methodologies for each student in order to provide remedial help without stigmatising or isolating students or embarrassing them in front of their peers;
- estimate how students will perform on standardised tests (i.e. predictive assessment);
- find out which instructional techniques work best for students and to provide customised teaching (i.e. diagnostic assessment);
- feedback in real-time to help improve student performance;
- conduct adaptive testing;
- merge systems such as learning management and curriculum management;
- integrate ICT devices used by students in classrooms and homes leading to a large amount of useful information about them under initiatives such as bring your own device (BYOD);
- combine various data sources such as course records, student attendance, class rosters, programme participation, degree attainment, discipline records and test scores which could enable more efficient management of student recruitment, administration and academic research; (Hoit, 2012; West, 2012).

In addition to the applications mentioned above, awarding bodies could use data for more comprehensive research in areas such as test development and marker monitoring. They could also make use of large amounts of data which is likely to be generated by the use of computerised assessment and through other IT-enabled initiatives such as computerised, interactive systems for producing questions.

Educational courses in big data

McKinsey reports that by 2018 the United States alone will face a shortage of up to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data (Manyika *et al*., 2011). A recent report prepared by *e-skills UK*[3] for SAS suggests that over the next five-year period the average annual growth rate of demand for big data professionals in the UK is expected to be about 18% per annum (compared to 2.5% for IT staff). This would equate to the generation of approximately 28,000 job opportunities per annum (a total of 132,000) by 2017 (e-skills, 2013).

Various universities in the UK are offering MSc courses in big data/

3. The Sector Skills Council for Business and Information Technology based in the UK.

analytics/data science/business intelligence/marketing analytics. These include University College London (UCL), Imperial College, Royal Holloway, Sheffield Hallam University, University of Dundee, Warwick University, Aston University and the University of Westminster. Bournemouth University is offering an MSc in Applied Data Analytics in partnership with SAS. SAS has also launched the SAS student academy in collaboration with Birmingham City University to tackle the demand for big data specialists (Shah, 2012; Orater, 2013).

Internationally, universities offering similar courses are the National University of Singapore (in collaboration with IBM), George Washington University, Columbia University, the Big Data Institute – University of Virginia, University of San Francisco, and New York University. Online courses in this field are also being provided by various institutes and MOOC (Massive Open Online Course) providers such as the Stanford University, University of California, Berkeley School of Information, Big Data University, MIT, Coursera and Statistics.com (KDnuggets, 2014). Short term professional courses are being run by the University of Oxford and Harvard. Technology vendors such as IBM, SAS, SAP and Google are also running various academic programmes in this field (Nerney, 2013).

Big data and social media

Businesses thrive on understanding their customers to the greatest extent possible. The monitoring of people's online behaviour is therefore becoming important for their success. Organisations are investing in gathering such analytics using big data as a key component for monitoring social media activity, particularly on social networking websites such as Facebook, Twitter and LinkedIn.

Social media analytics are the synthesis of the behaviour of internet users. The availability of data on consumers' web browsing, online shopping behaviour, customers' feedback and marketing research on social networks allow organisations to gain timely and extensive insights into consumers. Therefore, organisations can focus their market intelligence strategies based on different objectives such as advertising and product launches; publicity and brand management; promoting customer loyalty; providing personalised services to customers; keeping a tab on market trends and competitors; minimising risk; saving cost and business expansion in general.

The big data phenomenon applied to social media is fuelling a new, growing area of study known as 'sentiment analysis'. Its aim is to be aware of what people say or share in their everyday life. Businesses mine this information to understand their customers and to improve their operations accordingly. Educational organisations could also 'listen' to students and gain further insights into their perceptions. Using students' activity on social networking websites, sentiment analysis provides a useful tool to gather information about their online behaviour and, most importantly, their feedback on different aspects of the educational system, such as the university admissions process, features of qualifications, examinations and their aspirations.

Organisations could feed this information into developing their marketing strategies. This could be done in a number of ways, such as targeting the countries/regions with lower than expected online activity from their students, monitoring their examination experiences based on discussions in online forums, understanding what their brand means to students and getting feedback on new products.

Tools and metrics

The availability of more sources and forms of online data has also led to the development of new tools to access information and produce metrics about visibility of websites. It is possible to gather metrics such as countries/cities where website visitors were based, the web browsers they were using, the keywords they had used to search for a website and the webpages they had visited before and after accessing a particular website. Some such metrics are presented below.

Website rankings

Websites can be ranked to get an estimate of a website's popularity relative to all other websites over a specified period of time (for instance, six months or one year). The ranks are provided by tools such as www.ranking.com and www.alexa.com. The lower the rank, the higher the popularity of the website (for instance, the rank of Google.com is 1 followed by Facebook.com and YouTube.com). The ranks could be used by organisations to estimate the popularity of their websites in general, as well as in comparison to their competitors. Figure 1 shows a comparison of the ranks of two websites www.education.gov.uk and www.parliament.uk from November 2013 to May 2014.

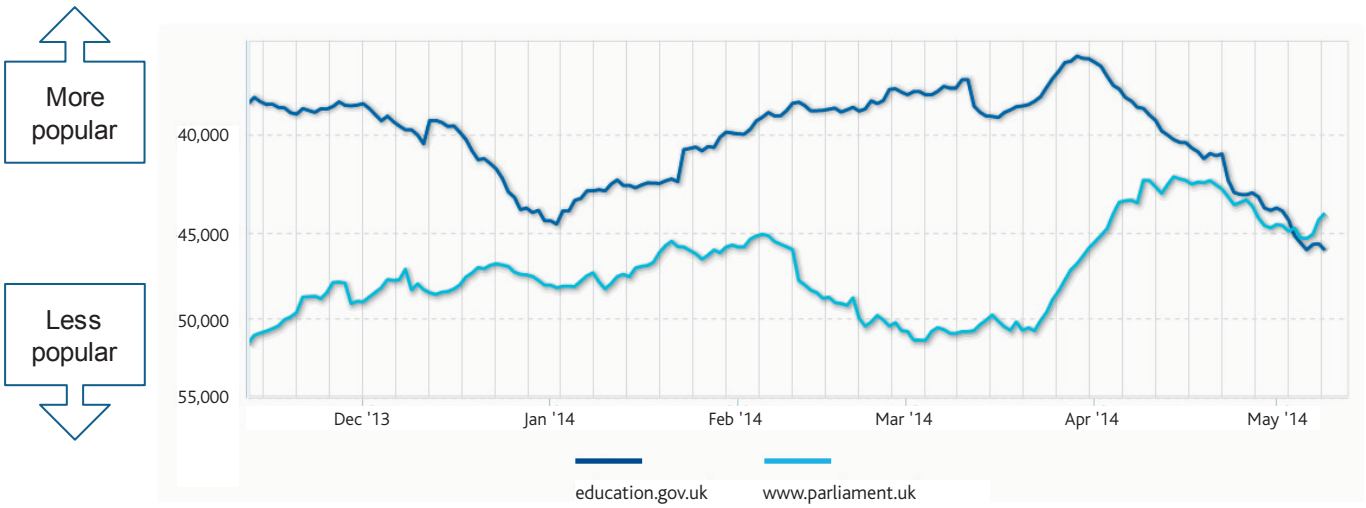


Figure 1: Historical traffic trends for the two websites from 12th November 2013 to 9th May 2014. Source: www.alexa.com (retrieved 12th May, 2014).

Online traffic analytics

Online tools such as Google Analytics and www.alexa.com provide traffic metrics of websites in the form of tables and interactive graphs which could be customised by the users. Some tools also provide the data collected in a spreadsheet, which can be used by organisations for producing graphs of their own. Some of the metrics provided are: the total number of visits to the website during a particular time period, the number of unique visitors, the total number of webpages viewed, the average number of webpages viewed in each visit, the average visit duration, and the bounce rate which is the percentage of single-page visits (i.e. visits in which a user left the website from the first page without continuing to view other pages within the site). Generally the lower the bounce rate the better the ability of the website to hold the interest of the visitors. A bounce rate of 50% is considered as an average value (Wikipedia, 2014b). All such metrics help organisations to get a more detailed understanding of the visitors to their websites which could be used for targeting their products and services. The metrics also allow the identification of those website sections which are popular with the visitors and those which are not, which in turn could help organisations improve their websites.

Social media monitoring

Organisations are able to be in regular touch with their customers through social media websites such as Facebook, Twitter, LinkedIn, YouTube and blogs. Organisations could also interact with their employees and other stakeholders (e.g. students, customers, external consultants) using tools such as Yammer, a private social network that aids collaboration across departments, locations and business applications.

Organisations can also monitor the news and find out who the key contributors are in online conversations about them. They can measure the results of their campaigns and monitor potential problems. Training providers can use such website monitoring tools to help them to understand and improve the reach of their training courses. Businesses can benefit from understanding the interconnections between their online users.

The use of some of the monitoring tools which offer basic metrics is free. However, most of the services that can actually help a business can be very costly ranging from a few hundred to several thousand pounds per month. It is therefore important for businesses to strategically plan their requirements and expectations from online monitoring tools. This might not be an easy task, because social media is a new and very fast changing area. In addition, the number of service providers in this area is growing rapidly and it might be difficult to find a reliable provider. The trialling of some tools might be required before selecting the most appropriate solution. Not all tools would fulfil the requirements of every organisation. The reports produced by the tools should be easy to interpret and worth the cost.

Tools for social media monitoring

Some popular tools for monitoring of social media are: Yomego, Ubervu, Hootsuite and Vocus. Other tools which social media managers may find useful are given in Table 2 and 3. Table 2 lists several web analytics reporting tools which can be used for producing insights from users' own websites. Data is visually presented using graphs and tables that can be customised through dashboards. Table 3 provides web-traffic estimation services which help gather how much traffic websites are receiving.

Table 2: Web analytics tools

Service	*Description*	*URL*
Google Analytics	Perhaps the most widely used website metrics service. It generates detailed metrics about a website's traffic. It's easy to use and is specifically designed for marketing research.	http://www.google.com/analytics/
AWStats	An open source web analytics reporting tool where users are encouraged to contribute to its development.	http://awstats.sourceforge.net/
Amung.us	Provides widgets to be included in the websites which show the number of live readers viewing a webpage and the location of current and previous visitors, in real time.	http://whos.amung.us/
WebSTAT	Its distinctive trait is the measure of visitors' behaviour once on the website. This includes their drivers and conversions; such as, the degree to which different landing pages are associated with online purchases.	http://www.webstat.com/

Table 3: Web traffic estimation tools

Service	*Description*	*URL*
Alexa	Provides an estimate of the percentage of internet users that may have visited a website during the last six months and allows comparisons with other websites.	http://www.alexa.com/
Compete	Helps to monitor online competition and to benchmark performance against the industry.	https://www.compete.com/
Website trafficspy	Makes use of data from a number of external sources to estimate traffic of a business' website or of their competitors.	http://websitetrafficspy.com/

Though this kind of data might not be completely accurate, it can be extremely useful to get an overall picture for marketing research.

Discussion

Data is changing our world – and fast. There is no denying this fact. What we buy, what we eat, how we communicate, how we are governed, how we live are all affected by the use of data. However, it should be noted that using data in day-to-day life is not a new concept. Ancient civilisations designed their calendars by predicting planetary movements based on data from prior recordings. More recently the advancement in digital and telecommunication technologies has led to an explosion of the amount of data available. The world has never been so interconnected. Each person who uses the internet, the telephone, or credit cards leaves a trail of information which can be used by organisations to predict their behaviour and adapt accordingly. The same is true of anyone who pays a utility bill, files a tax return or is registered

with government in some way (electoral registration office, health services, etc.). Big data is also being used in government initiatives as well as in all areas of research including health, economics, manufacturing, defence and security and education.

Organisations should plan their big data and social media policies carefully and with a long term view in mind. Due to the hype created in this area companies appear be in a rush to collect huge amounts of data, both text and non-text. However, not all of the data which they collect is necessarily meaningful or required. In essence, big data means combining data from various sources. There is a risk that accumulating very noisy data and making sense of it may require more resources than the returns it creates. Organisations also need to be aware of the increasingly high costs of hiring 'big data' scientists. It would therefore be advisable to carry out a cost-benefit analysis at the outset. The risk of data policies being unsuccessful can prove to be very costly for an organisation – both to its balance sheet as well as to its brand.

Schools and educational organisations hold huge amounts of data about students. This may include biographical information (such socio-economic status and ethnicity) and performance history (marks/grades/teacher observations) in summative or diagnostic assessments. Applications such as computer-based assessments allow more sources of data to be collected and analysed, such as the time spent by test takers on each question. This can help in the understanding of student performance more comprehensively which could be used at the classroom level to enable more targeted and timely interventions. Similarly, online marking of question papers makes available more (and certainly more accessible) data to awarding bodies for monitoring markers and evaluating their tests. Researchers and businesses may look forward to some new and innovative applications of data, as well as more refined statistical approaches to analysing complex data.

Acknowledgements

We would like to thank our colleagues Tom Benton, Nick Raikes, Sylvia Green and Frances Wilson for their advice.

References

BBC (2013). *The age of big data: BBC Horizon.* Retrieved from http://www.youtube.com/watch?v=CO2mGny6fFs

BIG (2014). *Big data public private forum.* Retrieved from http://big-project.eu

Beyer, M. A., & Laney, D. (2012). *The importance of 'Big Data': A definition* (Gartner Report G00235055). Retrieved from https://www.gartner.com/doc/2057415?ref=clientFriendlyURL

Bradbury, D. (2013, June). Effective social media analytics. *The Guardian*. Retrieved from http://www.theguardian.com/technology/2013/jun/10/effective-social-media-analytics

Einav, L. & Levin, J.D. (2013). *The data revolution and economic analysis* (NBER Working Paper no. 19035). Retrieved from http://www.nber.org/papers/w19035

e-skills UK (2013). *Big data analytics. An assessment of demand for labour and skills, 2012–2017* (E-skills UK report on behalf of SAS UK). Retrieved from https://www.e-skills.com/Documents/Research/General/BigDataAnalytics_Report_Jan2013.pdf

European Commission (2014). *The EU Framework Programme for Research and Innovation.* Retrieved from http://ec.europa.eu/research/horizon2020/index_en.cfm?pg=h2020

Diaz, J. (2010). *The one hundred trillion dollars hard drive.* Retrieved from http://gizmodo.com/5557676/how-much-money-would-a-yottabyte-hard-drive-cost

Hoit, D. M. (2013). *Big data, big expectations* (Centre for Digital Education Report Q2 2013). Retrieved from http://www.centerdigitaled.com/paper/2013-Q2-Special-Report-Big-Data-Big-Expectations.html

IBM (2013). *Social media analytics: Making customer insights actionable.* Retrived from http://www-01.ibm.com/software/analytics/solutions/customer-analytics/social-media-analytics

Kalil, T. (2012). *Big data is a big deal.* Retrieved from http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

KDnuggets (2014). *Online education in analytics, big data, data mining, and data science.* Retrieved from http://www.kdnuggets.com/education/online.html

Kurmanath, K. V. (2014). Every 11th voter in Uttar Pradesh is a 'Ram'. *The Hindu Business Line.* Retrieved from http://www.thehindubusinessline.com/news/politics/big-data-throws-up-interesting-trivia-in-general-elections/article6011219.ece

Lohr, S. (2012, February 11). The age of big data. *The New York Times.* Retrieved from http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* Boston: Houghton Mifflin Harcourt.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity* (McKinsey Global Institute report). Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Nerney, C. (2013). *Universities Expanding Big Data Analytics Courses with IBM Aid.* Retrieved from http://data-informed.com/universities-expanding-big-data-analytics-courses-with-ibm-aid

Orater (2013). *List of masters courses in analytics (UK & Ireland).* Retrieved from http://www.whatisanalytics.co.uk/jm/index.php/articles/analytics-degrees/24-uk-masters-courses

Shah, S. (2012). *SAS launches academy to tackle demand for "£52,000 a year" big data specialists.* Retrieved from http://www.computing.co.uk/ctg/news/2230956/sas-launches-academy-to-tackle-demand-for-gbp52-000-a-year-big-data-specialists

Swoyer, S. (2012). *Big data – why the 3Vs just don't make sense.* Retrieved from http://tdwi.org/articles/2012/07/24/big-data-4th-v.aspx

Raconteur Media (Ed.) (2013, September 4). Big data. *The Times* [supplemental material].

Villanova University (2014). *What is big data?* Retrieved from www.villanovau.com/university-online-programs/what-is-big-data

West, D. M. (2012). *Big data for education: Data mining, data analytics, and web dashboards* (Brookings paper). Retrieved from http://www.brookings.edu/research/papers/2012/09/04-education-technology-west

Wikipedia (2014a). *Big data.* Retrieved from http://en.wikipedia.org/wiki/Big_data

Wikipedia (2014b). *Bounce rate.* Retrieved from http://en.wikipedia.org/wiki/Bounce_rate

Wikipedia (2014c). *Yotta.* Retrieved from: http:// http://en.wikipedia.org/wiki/Yotta-

Wikipedia (2014d). *Yottabyte.* Retrieved from http://en.wikipedia.org/wiki/Yottabyte

Multivariate representations of subject difficulty

Tom Bramley Research Division

Introduction

The aim of this study was to explore multidimensional ways of representing similarities and differences in the grade distributions of different A level subjects, in contrast to the more familiar unidimensional ways which are often interpreted as revealing differences in subject 'difficulty'. Of particular interest was whether an a priori scheme for classifying the subjects would be reflected in the multidimensional configurations.

Debate over whether some examination subjects are 'harder' than others has been around for a long time. Newton (2010) has previously noted the importance of distinguishing between definitions of comparability, and methods for monitoring whether it has been achieved. Statistical methods for monitoring inter-subject comparability have been both used and criticised (see Coe 2007, 2010). Where statistical methods are used, the aim is usually to produce a *single* ranking of subjects according to an indicator of difficulty. For example, the report by Coe, Searle, Barmby, Jones, and Higgins (2008) contained many tables showing the results of such rankings from research exercises carried out in the 1970s, 1980s, 1990s and 2000s using a variety of different statistical methods. Their conclusion (for A levels) was that although different methods did give slightly different results (rankings of subjects by difficulty), the differences between methods were much smaller than the differences in difficulty between subjects that the methods revealed: "the argument that the different methods do not agree is not a convincing reason to use none of them." (Coe *et al*., 2008, p.89).

Focusing on one particular (class of) statistical method, the 'Item Response Theory (IRT) approach', Bramley (2011) explored the analogy between item difficulty (which IRT methods were developed to model/measure) and subject difficulty, concluding that using IRT methods for the latter places a greater burden on the analyst to interpret the meaning of the latent trait and the difficulty parameter in the IRT model than is the case for 'normal' use of IRT. This is mainly because examined subjects at a particular level (e.g. A level) form a largely *ad hoc* collection, in contrast to the set of items on a particular examination which have been designed to assess a syllabus and cover a range of topics and difficulties with a target population of examinees in mind. In view of this, Bramley *(ibid)* suggested exploring ways of representing subject difficulty graphically, without aiming to produce a single overall ranking of subjects by difficulty.

The study reported here followed up that suggestion by applying the technique known as multidimensional scaling (MDS)[1] to data from OCR A levels taken in June 2011. The MDS results were compared with those from two unidimensional methods (the Kelly method and the Rasch method) that give a single ranking of subjects in terms of difficulty.

1. The ideas behind MDS have been developed independently by different researchers in different places and consequently there is a variety of terminology in use.

Classifying A level subjects

The A level subjects were classified in advance into categories in order to see if the location of subjects in the unidimensional rankings or the MDS representations corresponded to these a priori classifications. There are obviously many different ways in which A level subjects could be categorised, all of which would be to some extent arbitrary. For example, the list of 'academic disciplines' (not A levels) on Wikipedia[2] has the following high-level groupings:

- Humanities (e.g. History, Philosophy, Performing Arts)
- Social sciences (e.g. Economics, Psychology, Anthropology)
- Natural sciences (e.g. Physics, Chemistry)
- Formal sciences (e.g. Computer sciences, Mathematics, Logic)
- Professions and applied sciences (e.g. Agriculture, Law, Engineering).

The problem with the above list is that it is more appropriate for university disciplines than A level subjects. Languages would only appear indirectly as 'Linguistics' or 'Literature' within the humanities, whereas they seem to form a more definite category of A level subject.

Taking a Facet Theory approach (e.g. Borg & Shye, 1995) to producing a categorisation scheme would require identifying a rule or rules by which a given A level could be unambiguously allocated to a category. Following discussion with colleagues of various categorisations currently in use, and given an aim to have some fairly uncontroversial and intuitive categories, the categorisation in Table 1 below was used for this research.

The STEM classification seemed fairly self-explanatory, even though no rule was created. Problem cases were Geology (classified as STEM), Psychology (classified as a Humanity) and Applied Science (classified as Applied).

2. http://en.wikipedia.org/wiki/List_of_academic_disciplines (Accessed 13/03/14).
3. Science, Technology, Engineering and Mathematics – a grouping often used in media reporting.

Table 1: Classification of A level subjects into categories

Category	*Rule*	*Examples*
STEM3		Maths, Physics, Computing
Humanities	Knowledge, skills & understanding expressed mainly through extended writing	English Literature, Classics, Media Studies, Psychology.
Languages	Require learning some of the vocabulary and grammar of a second language.	Latin, French, Spanish, Turkish.
Expressive	Knowledge, skills & understanding expressed mainly through performances or artefacts	Music, Design and Technology, Art and Design, Performing Arts.
Applied	Knowledge, skills & understanding lead more directly to jobs or job-related further courses.	Accounting, Health & Social Care, Applied ICT, Law.

The list of subjects in the Humanities category seemed reasonable enough, although the classification rule itself would not be good enough to unambiguously allocate subjects to the category.

The Languages category was also fairly straightforward, although it requires assuming that the first language of A level examinees is English (which was classified as a Humanity for this research). For some language A levels (e.g. Turkish) it seems possible that a significant proportion of native speakers may take the A level, but this does not so much cast doubt on the validity of the classification rule, but on the validity of inferences made about the relative difficulty of some language A levels using statistical methods.

The Expressive categorisation was more problematic in that Design and Technology could perhaps also fit in the STEM or Applied categories, and that in some cases it is perhaps doubtful whether knowledge, skills and understanding are expressed *mainly* through performances and artefacts (as opposed to through written responses).

The Applied category was relatively straightforward on the assumption that subjects with the word 'applied' in their specification title are indeed intended to lead more to jobs or job-related further study than to academic study, as per the classification rule for this category.

Unidimensional representations

Kelly method

The Kelly method (Kelly, 1971) is a relatively straightforward way of deriving rankings of subjects by difficulty. It is used by the SQA to obtain rankings of Scottish Highers. The method is described in technical detail by Coe (2007). Basically, the output of the method is a difficulty rating for each subject which can be interpreted as the adjustment that should be made to the (numerical[4]) grades in each subject in order that, on average, examinees achieve the same average adjusted grade in their other subjects that they achieve in any particular subject. A positive value therefore indicates a more difficult subject (defined by this method as one in which examinees on average obtained lower grades than in the other subjects they took).

The analysis used a sub-set of 33 of the OCR A level specifications from the June 2011 examination session. For subjects with more than one specification, the one with the larger entry was retained. Specifications with fewer than 400 examinees taking at least one other OCR A level were dropped, with the exception of German and Spanish, which were retained so that the category of Languages would be better represented. Table 2 shows the Kelly difficulty ratings of these 33 subjects, colour coded by higher-level category. The change in rank position (out of 33) from 2010 to 2011 is also included.

Inspection of Table 2 shows that Kelly difficulty rating was related to category, with (in general) STEM subjects and Languages being more difficult, Expressive and Applied subjects being easier, and Humanities generally in the middle, with the exceptions of General Studies and Critical Thinking being more difficult, and Sociology and Media Studies being easier. A plausible explanation for the relative difficulty of General Studies is motivation – if examinees do not try as hard or prepare as well for this exam then it will appear harder. Similarly Critical Thinking may suffer from both motivation effects, and a lack of teaching time and teaching experience (see Black, 2009). The stability of the ranking from 2010 to 2011, as shown by the fact that no subject changed by more than three places in the overall ranking, is indirect evidence of within-subject standard maintaining from year to year.

4. Letter grades are converted to numbers on an interval scale: A*=6, A=5, ... E=1, U=0.

Table 2: Difficulty ratings (Kelly method) of 33 OCR A level specifications in June 2011

	Category	*Assessment Name*	*Difficulty from 2010*	*Change*
1	STEM	Further Mathematics	0.95	=
2	Humanities	Critical Thinking	0.74	=
3	Humanities	General Studies	0.60	+1
4	STEM	Physics A	0.49	-1
5	STEM	Chemistry A	0.46	=
6	STEM	Biology	0.23	=
7	Languages	Classics: Latin	0.19	+1
8	Languages	French	0.18	+2
9	Expressive	Music	0.12	+3
10	Languages	German	0.11	-3
11	STEM	Computing	0.06	-2
12	STEM	Mathematics	0.04	-1
13	Languages	Spanish	0.02	=
14	Applied	Applied ICT	-0.02	=
15	Humanities	Economics	-0.09	=
16	Humanities	History A	-0.18	+2
17	Humanities	Government And Politics	-0.21	-1
18	Humanities	Classics: Classical Civilisation	-0.25	-1
19	Humanities	Geography	-0.27	+1
20	Humanities	Psychology	-0.31	=
21	Humanities	English Literature	-0.33	-1
22	Humanities	Religious Studies	-0.35	-3
23	Applied	Physical Education	-0.43	+2
24	STEM	Geology	-0.48	-1
25	Applied	Law	-0.49	-1
26	Applied	Business Studies	-0.62	+1
27	Expressive	Performance Studies	-0.65	+2
28	Applied	Health And Social Care	-0.66	-2
29	Expressive	Design And Technology: Product Design	-0.67	-1
30	Humanities	Sociology	-0.85	=
31	Expressive	Art And Design: Fine Art	-0.95	=
32	Humanities	Media Studies	-1.01	=
33	Expressive	Art And Design: Photography – Lens And Light-Based Media	-1.37	=

Rasch method

The Rasch method for comparing subject difficulty is also described in Coe (2007) and Bramley (2011). The Rasch model characterises persons and items (here, A level specifications) by a single number that can be taken as representing their location on the overall construct that is being measured by the items. In this case, the overall construct has to be interpreted as something like 'general academic ability' (Coe, 2010).

The Rasch Partial Credit model (PCM) (Masters, 1982) was fitted to the A level data, which instead of the usual examinee × item matrix contained examinees on the rows but A level specifications in the columns, with the data being the numerical grade obtained by the examinee in that specification. The matrix was large and contained mostly missing data (as examinees took at most five A levels). The data was analysed with the FACETS program (Linacre, 1987).

Although the input is identical to that for the Kelly analysis, the Rasch analysis is more complex in that an explicit model is fitted, and parameters are estimated for the thresholds between each grade category. There is therefore no single difficulty of an item estimated with the PCM, although it is conventionally taken as the mean of the threshold parameters. An interpretation of this mean value is that it is the ability level at which obtaining a grade in the bottom (U) or top (A*) categories is equally likely (Linacre, 2005). Higher values therefore indicate more difficult subjects, but the logit (log odds) scale is less readily interpretable than the Kelly output (which is in terms of numerical grades).

As Coe *et al.* (2008) found, the Kelly and Rasch results were very similar. The correlation was 0.90, which rose to 0.96 when outliers were excluded.[5] Unlike the Kelly method, the Rasch method also produces an indication of how well each person and item (here, A level subject) has fit the model. An 'overfitting' item or person is one whose observed responses conform more closely to the model than expected, given its probabilistic nature, whereas an 'underfitting' or 'misfitting' item or person is one whose observed responses confirm less well to the model than expected. A 2-dimensional representation of the Rasch results can thus include both difficulty and fit[6], as shown in Figure 1 for the 33 subjects.

Figure 1 shows that as well as being more difficult, STEM subjects tended to overfit the Rasch model. The Languages, and the subjects classified as Expressive tended to underfit (misfit) the model. Interestingly the two most difficult Humanities subjects (General Studies and Critical Thinking) also had large values for the misfit indicator, supporting the earlier conjecture that factors other than general academic ability might have affected the observed grades in these subjects. The Humanities and Applied subjects generally seemed to fit the model reasonably well.

5. These outliers were specifications containing grade categories with no examinees in them (usually U or A*). The Rasch analysis 'collapses' such empty categories when they occur, thus changing the meaning of some of the parameters and hence of their average value.

6. The fit statistic shown in Figure 1 is the infit-z statistic output from FACETS. Negative values indicate overfit, positive values misfit.

Multidimensional representations

Although no one denies that the results of unidimensional representations such as those of the Kelly or Rasch methods have produced stable outcomes over a long period of time, there is much more disagreement over the interpretation, utility and implications of such results. Detailed discussion can be found in Coe (2007) and Newton (2010). The main purpose of the present study was to explore whether anything might be gained from setting aside the potentially inflammatory search for a single ranking of subjects by difficulty and looking for other ways to summarise or characterise the same underlying data (i.e. the grades obtained by OCR A level examinees in each specification). The technique explored was multidimensional scaling (MDS).

MDS is actually a set of techniques that have the common aim of representing indices of similarity or dissimilarity between a set of objects as a spatial configuration of points, where the points represent the objects, and the distances between points in the configuration reflect relationships between the indices of similarity or dissimilarity. See Kruskal and Wish (1978) or Borg and Groenen (2005) for detailed explanations of MDS concepts, formulas, applications and issues. There are many choices that have to be made when carrying out an MDS analysis, including:

- the function relating similarities/dissimilarities to distances in the MDS configuration;
- which index of similarity or dissimilarity to use;
- the dimensionality of the configuration;
- whether to give equal weight to all the data, or more weight to some points and less to others.

For all the analyses, a non-metric (i.e. ordinal) function was specified – this imposes the fewest constraints on the analysis. The aim of non-metric MDS is to preserve rank-order relationships as far as possible in the MDS configuration. For example, if (according to the index of

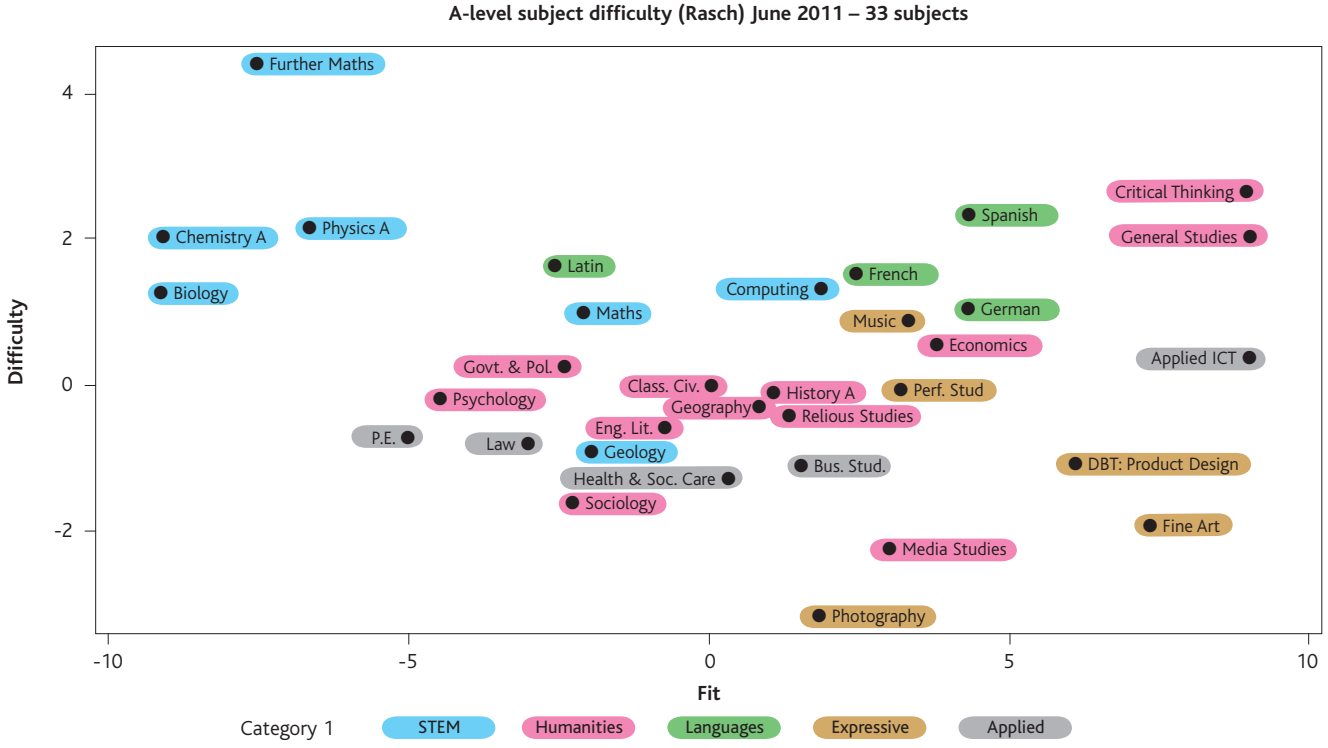


Figure 1: Plot of Rasch average difficulty v fit.

similarity used) subject P is more similar to subject Q than subject R is to subject S, then in the MDS configuration, P will be as close or closer[7] to Q than R is to S.

Three indices of similarity between pairs of subjects were explored here. First, the absolute (unsigned) difference between mean numerical grade obtained by common examinees. The (signed) difference between mean grades is familiar from subject-pairs analyses (e.g. Forrest & Smith, 1972), the idea being that if a group of examinees obtains (on average) a higher grade in subject P than in subject Q, then subject P is less difficult than subject Q. For the analysis here, however, only the size of the difference between each pair of subjects was preserved, not the direction of the difference.

The second index was the proportion of common examinees obtaining exactly the same grade in subject P as subject Q (denoted here as P_0). Clearly, the higher this index the more similar it can be argued the two subjects are – but it does not address difficulty *per se* because it does not take into account what grades were obtained by common examinees who did *not* get the same grade. It is in principle possible for two pairs of subjects (PQ and RS) to have the same value for P_0, but for the majority of common examinees in one (say PQ) who did *not* get the same grade in P as Q to get a better grade in P than Q, whereas in the other (RS) for common examinees who did not get the same grade to be equally distributed among those who had got a better grade in R than in S and those who had got a better grade in S than in R.

The third index of similarity was Guttman's coefficient of monotonicity μ_2 (Guttman, 1977). This is essentially an ordinal correlation coefficient, ranging from -1 to +1. The formula is given in the appendix. It takes its maximum value of +1 when an increase in one variable is always associated with an increase (or no change) in the other. As with P_0, it does not address difficulty – two subjects could have a perfect monotonic correlation between the grades of common examinees, but the grades obtained in subject P might be systematically higher (or lower) than in subject Q. The μ_2 coefficient has been the index of similarity favoured by many practitioners of Facet Theory because it requires no assumptions of interval-scale measurement or of linear relationships.

Solutions for 1 to 4 dimensions were investigated in each case, to gain a feel for how much information was being lost by reducing the dimensionality. It seemed sensible to give more weight in the analyses to indices of similarity from pairs of subjects with large numbers of common examinees, on the assumption that common examinees from such subject pairs were more likely to be representative of the general examinees in those subjects, and the view that it was in general more important to give weight to the larger-entry subjects. The software used to run the analyses was the PROC MDS procedure in SAS 9.2. The default options were used[8].

1. Similarity in mean grade of common examinees

The two-dimensional MDS solution had a value for the 'stress' (badness of fit) statistic just under 0.20. The value of 0.20 is given by some sources as a rule of thumb for an acceptable or adequate fit, although most sources emphasise that (as with any complex statistical method), rules of thumb are often misleading as stress can be affected by a number of factors, such as the number of points being represented and error in the proximities. However, there is agreement that the main purpose in exploratory MDS is to arrive at an interpretable visual representation, which means that in practice usually only 2- and 3-dimensional solutions are considered. For the other indices of similarity (see later) the 2-D stress was above 0.2, so 3-D representations were considered for those.

It can be seen from Figure 2 that the location of points along Dimension 1 happened to correspond closely to the ordering of difficulty

7. This is the 'weak monotone' function most commonly used (Borg & Groenen, 2005, p.40).

8. See http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mds_sect004.htm for a description of these default options.

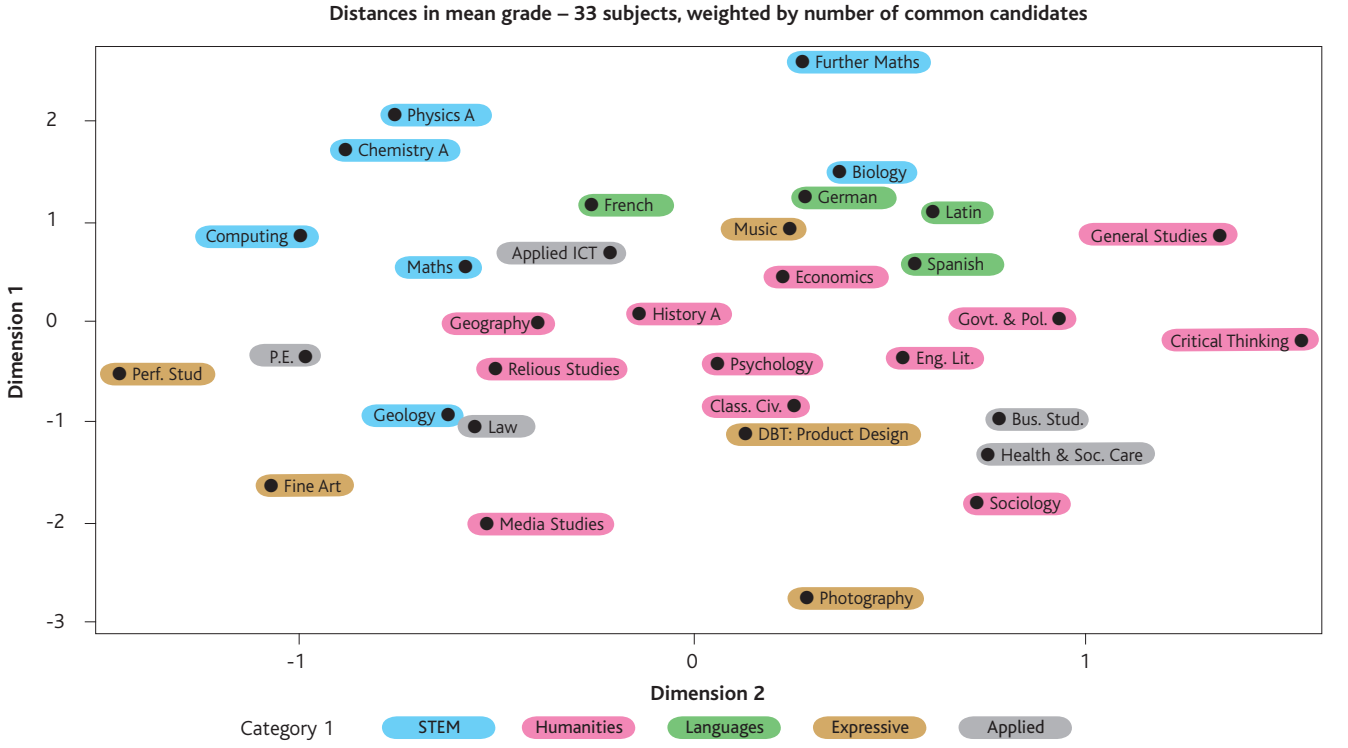


Figure 2: Two-dimensional non-metric MDS representation of differences in mean grade.

from the Kelly and Rasch methods ($r \approx 0.9$). Dimension 2 however did not correlate highly with the fit to the Rasch model ($r \approx 0.2$), as can be seen by comparing Figure 2 with Figure 1. Of course, the axes in an MDS solution have no intrinsic meaning – it is the distances between points that are relevant. (In other words, the configuration in Figure 2 could be rotated or reflected without affecting the fit of the solution). Nonetheless, it is still reasonable to look for any interpretable directions across the configuration so the relationship with difficulty is interesting, particularly since it emerged without 'telling' the software which subject in each pair had the higher mean grade. The STEM subjects, the Languages and to a lesser extent the Humanities do seem to group together in Figure 2, suggesting that within these groupings, differences in mean grade of common examinees were more similar than across groupings. There is no obvious pattern for the subjects classified as Applied, but for the Expressive subjects there is a tendency for them to be on the edge of the configuration, suggesting greater differences between these subjects and the others.

2. Similarity in percentage of common examinees with the same grade

The 2-D solution had a stress value of around 0.24, but the 3-D solution had a value around 0.16, suggesting that three dimensions were needed to adequately preserve the relationships between similarity of P_0 values. The 3-D representation is shown in Figure 3 below.

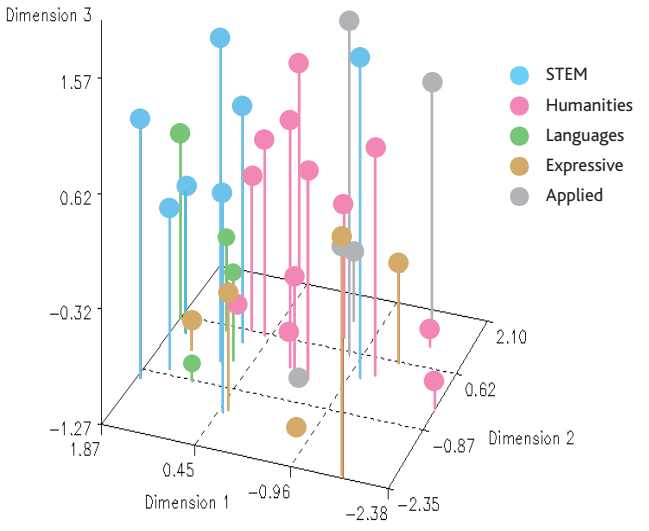


Figure 3: Three-dimensional non-metric MDS representation of differences in P_0

In Figure 3 the STEM (except Geology) and Language subjects seem to group well, and the Humanities reasonably well. The Expressive and Applied are less clearly grouped but still closer (by eye) than a random allocation of points. Interpreting static 3-D representations on a 2-D surface is not easy, so rotating the graph (possible with most modern graphics software) can make it easier to look for patterns.

3. Similarity of coefficient of monotonicity between grades of common examinees

The 2-D solution had a stress value of around 0.26, but the 3-D solution had a value around 0.18, suggesting that three dimensions were needed to adequately preserve the relationships between similarity of μ_2 values. There was some discernible clustering by group, but not quite as clear-cut as for the P_0 similarity index depicted in Figure 3.

Discussion

The MDS analyses have shown that 2- or 3-D representations of aspects of the raw data (the A level grades obtained by OCR examinees) do highlight groupings of the subjects in terms of categories that can be identified prior to analysis. Although there were differences in the patterns across the MDS representations using different indices of similarity, at a broad level the same findings were observed – that is, STEM subjects, Languages and Humanities clustering together fairly well in the representations, Expressive and Applied subjects less well.

Did increasing the dimensionality beyond the usual one (interpreted as 'difficulty') yield new insights? Unfortunately the difficulties in interpretation remained, and this is an intrinsic feature of the data at hand: examinees choose a small and very non-random subset of the possible subjects. Twenty seven pairs of subjects had no common examinees, and 167 pairs (of the 528) had fewer than ten. Only seven pairs had more than 1,000 common examinees, and these all involved STEM subjects and General Studies. Clearly it is impossible to create the 'ideal' situation where all examinees take all subjects. We can therefore never know whether some pairs of subjects would have higher (or lower) indices of similarity if more examinees had taken both of them.

The MDS methods do not of themselves permit an interpretation of the dimensions of the configuration – it is the distances between the points that should be interpreted. Nevertheless, it can be hard to resist the temptation to look for a 'difficulty' dimension, given the stable Kelly and Rasch findings. It was interesting that one of the dimensions in all three of the MDS analyses seemed to be fairly closely related to unidimensional difficulty, given that only the first of the three indices of similarity was directly related to difficulty. However, the input for calculating all three indices of similarity was essentially the same – the 7×7 cross-table of grade in subject X against grade in subject Y with cells of the table containing the number of examinees containing the corresponding pair of grades in the two subjects, as shown in the example in Table 3.

The first index of similarity, absolute difference of mean grade, only uses the information in the margins of the table: abs [(257×6 + 472×5 ... + 112×1) – (240×6 + 455×5 ... + 92×1)]/1763.

The second index, P_0, the proportion of examinees achieving the same grade in both, only uses the information in the shaded blue top-left to bottom-right diagonal and the overall number of common examinees: (161 + 269 +... + 14)/1763. The third index, Guttman's μ2, takes into account the frequencies in each cell of the table, as shown in the second formula in the appendix. Although the different indices therefore use different aspects of the table, they are not independent. For example, if Physics were graded more leniently, more examinees would move into the cells above and to the right of the shaded diagonal. This would increase the mean grade difference and decrease P_0 (assuming that more examinees would move out of the shaded diagonal than into it). Larger differences in difficulty are therefore likely to correspond to lower values of P_0, and hence it is perhaps not surprising that one direction in the MDS configurations correlated well with unidimensional difficulty.

Future work could verify that the Rasch and Kelly results continue to show a stable pattern, and explore whether the 2- and 3-D MDS configurations also show stability. If there is a clearly identifiable 'background of stability' this could prompt investigations of any subjects that appear to be moving against the stable background – for example this might signify changing entry patterns, or changing grading standards.

Table 3: Cross-tabulation of grades obtained by examinees taking both Physics and Chemistry in June 2011

		Chemistry							
		A (6)*	*A (5)*	*B (4)*	*C (3)*	*D (2)*	*E (1)*	*U (0)*	*Total*
Physics	*A* (6)*	161	86	9	1	0	0	0	257
	A (5)	74	269	115	12	2	0	0	472
	B (4)	5	89	198	71	14	2	0	379
	C (3)	0	10	95	123	51	12	0	291
	D (2)	0	1	15	72	100	22	8	218
	E (1)	0	0	3	14	39	42	14	112
	U (0)	0	0	0	1	5	14	14	34
	Total	240	455	435	294	211	92	36	1763

However, the index of similarity for subjects with small entries is always likely to be unstable and therefore the relative positioning of such subjects is likely to fluctuate. Another extension of this work could be to try other categorisations of subjects to see if there are some that lead to cleaner/sharper delineations of regions in the resulting spatial representations. To stay within the spirit of Facet Theory there would ideally need to be a rule or principle by which the categorisations could be applied.

In conclusion, the MDS representations of A level subjects according to various indices of similarity derived from the joint grade distributions of common examinees are interesting, but perhaps have too many difficulties attached to their interpretation to be worth pursuing. It may ultimately be easier to interpret trends and patterns in the indices of similarity directly, perhaps via consideration of cross-tabulations of pairs of subjects like those shown in Table 3 above. If a visual representation of a large number of subjects is desired then in my opinion the 2-D plot of the Rasch results (Figure 1) is the most informative, because it both allows a specific interpretation of 'difficulty', but also clearly shows the caveats in terms of the large differences in fit – which also are systematically related to subject groupings.

References

Amar, R. (2005). *HUDAP mathematics. 3rd edition*. Jerusalem: The Hebrew University of Jerusalem Computation Authority.

Black, B. (2009). *Introducing a new subject and its assessment in schools: the challenges of introducing Critical Thinking AS/A level in the UK*. Paper presented at the Association of Educational Assessment-Europe Conference, 7th November 2009, Malta.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: theory and applications*. New York: Springer Verlag.

Borg, I., & Shye, S. (1995). *Facet Theory: form and content*. Thousand Oaks: CA: SAGE.

Bramley, T. (2011). Subject difficulty – the analogy with question difficulty. *Research Matters: A Cambridge Assessment Publication, Special Issue 2: comparability*, 27–33.

Coe, R. (2007). Common examinee methods. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. (pp.331–367). London: Qualifications and Curriculum Authority.

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education, 25*(3), 271–284.

Coe, R., Searle, J., Barmby, P., Jones, K., & Higgins, S. (2008). *Relative difficulty of examinations in different subjects*. Durham: CEM Centre, Durham University.

Forrest, G. M., & Smith, G. A. (1972). *Standards in subjects at the Ordinary level of the GCE, June 1971*. Occasional Publication 34. Manchester: Joint Matriculation Board.

Guttman, L. (1977). What is not what in statistics. *The Statistician, 26*(2), 81–107.

Kelly, A. (1971). The relative standards of subject examinations. *Research Intelligence, 1*(2), 34–38.

Kruskal, J., & Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications.

Linacre, J. M. (1987). FACETS (Version 3.67.1): www.winsteps.com.

Linacre, J. M. (2005). The partial credit model and the one-item rating scale model. *Rasch Measurement Transactions, 19*(1), 1009.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174.

Newton, P. E. (2010). Contrasting conceptions of comparability. *Research Papers in Education, 25*(3), 285–292.

Appendix: Guttman's coefficient of monotonicity, μ_2

$$\mu_2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)(y_i - y_j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j||y_i - y_j|}$$

where x_i is the numerical grade of examinee i on exam X, y_i is the numerical grade of examinee i on exam Y, and N is the total number of common examinees.

or, from an $n \times m$ cross-tabulation of frequencies on X and Y:

$$\mu_2 = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{m-1}\sum_{l=k+1}^{m}(f_{ki}f_{lj} - f_{li}f_{kj})(\xi_i - \xi_j)(\psi_k - \psi_l)}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{m-1}\sum_{l=k+1}^{m}(f_{ki}f_{lj} + f_{li}f_{kj})|\xi_i - \xi_j||\psi_k - \psi_l|}$$

where ξ_i is the numerical grade of the ith category of exam X, ψ_k is the numerical grade of the kth category of exam Y, and f_{ki} is the number of examinees obtaining grade category k on exam X and grade category i on exam Y.

The above formulas are taken from the reference manuals for the HUDAP software package (Amar, 2005).

Calculating the reliability of complex qualifications

Tom Benton Research Division

Introduction

In order for qualifications to be meaningful they typically need to cover a greater range of curriculum material than could reasonably be assessed within a single paper. For this reason, all current GCSE and A level qualifications consist of multiple assessment components. Candidate achievement across these different elements is then combined in order to determine the final grade they will be awarded.

Estimating the reliability of qualifications that are examined through a composite of multiple assessments creates some challenges as any estimate of reliability must adequately account for the different amounts of weight given to different components. Some possible approaches to this issue are discussed by He (2009). However, a bigger problem arises when candidates have multiple options regarding which assessments will count towards their overall qualification grade. Such a situation could arise due to candidates working towards the same qualification being able to choose between:

- Different tiers
- Different papers covering different optional topics
- Different examination sessions for individual components as was possible in unitised assessment schemes

Previous work examining how the reliability of qualifications with multiple possible routes may be estimated, such as that by Bramley and Dhawan (2013), have addressed this issue by simply focussing on the most common set of options chosen by candidates to achieve a given qualification. The aim of this article is to demonstrate a relatively simple, and highly intuitive method of calculating reliability for such qualifications that includes the results of candidates across all possible routes. This method is exemplified for a very complicated qualification to show the power of the method in circumstances where it would not be feasible to derive estimates of composite reliability for each possible route.

The Qualification

This article focusses on OCR Mathematics A level specification 7890 and the candidates that certificated for this qualification in June 2012. To be awarded an A level, candidates needed to complete four compulsory units (Core Mathematics 1 to 4) and two out of a possible six optional units (two in each of Mechanics, Probability and Statistics and Decision Mathematics). For their optional papers they could either take both papers within the same optional subject area (e.g. both Mechanics papers), or the first paper in two different subject areas (e.g. Mechanics 1 and Decision Mathematics 1). To make matters more complicated they had the option to take a version of these papers within any of four examination sessions (January 2011, June 2011, January 2012 and June 2012)[1].

1. In theory candidates could also take any of these units prior to 2011 but this was rare for those candidates that completed the A level in 2012.

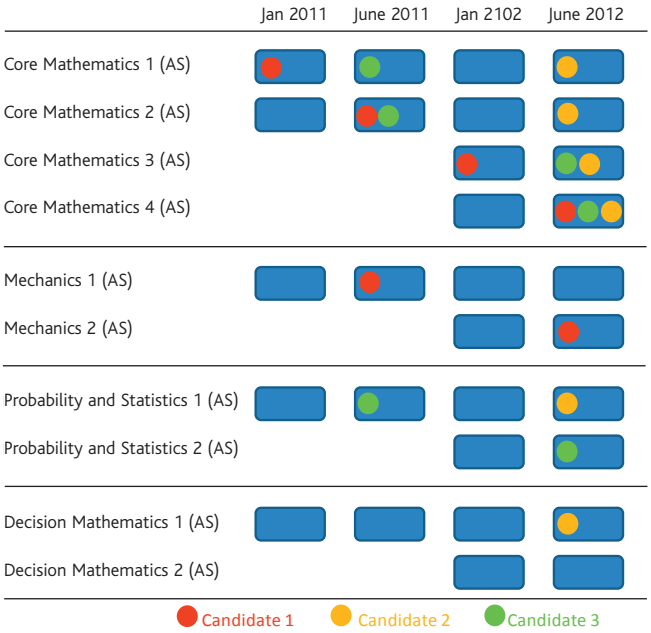


Figure 1: Possible routes through OCR Mathematics A level specification 7890 (assuming all A2 units taken in 2012)

Figure 1 illustrates some possible routes through this qualification. The 30 rectangles represent 30 of the papers available to candidates within this qualification. Note that there were no common questions across the 30 papers. The circles illustrate three different possible combinations of papers that would lead to completion of the A level. Candidate 1 takes the A level in a progressive modular fashion; taking one core unit in each available session and one optional unit (in Mechanics) each June. In contrast, candidate 2 takes a fully linear approach taking all core units and both optional units, this time split across Probability and Statistics and Decision Mathematics, in June 2012. Another option is illustrated by candidate 3; a modular approach but limited to using the June examination sessions[2].

As can be seen from Figure 1, there were an enormous number of options available to candidates. Even if we assume that all of the (more challenging) A2 units were taken towards the end of the course (that is, in 2012 rather than 2011), and that at least one of these A2 units was taken in June 2012 itself, there remains a total of 3,648 possible different combinations of papers that would have led to a Mathematics A level. Note that if, instead, all units had been required to be taken in a linear fashion in June 2012 then the number of possible routes would have reduced substantially, but would still have left six possible combinations of papers leading to the same qualification.

Given that each of the possible routes through the A level will lead to candidates being awarded the same qualification at the same time, it is

2. Note that for the purposes of this article resits are not relevant. For the purposes of this article we need only consider a candidates best performance within any element of the A level. This means that for each candidate we can restrict ourselves to exactly six examination scores.

of interest to calculate the overall reliability of the qualification. Whilst numerous techniques exist for evaluating the reliability of any one of the 30 papers listed in Figure 1 individually, there is little consensus regarding how the reliability of the qualification as a whole should be calculated.

In order to calculate reliability we first need to define what it means. In some senses this is a general problem with any reliability estimate with usual decisions including whether we are interested in reliability in terms of which questions are included – that is, how much difference the exact choice of questions within an assessment has upon the performance of candidates – or marking reliability – that is, how much difference it would make if the same set of responses from each candidate were marked by a different marker. For the purposes of this article we are only interested in reliability relating to the choice of questions[3]. In addition to the usual decisions over the definition of reliability there are some that are particular to the problem in hand. Specifically, we need to determine whether we are interested in:

- how much difference it would make if different questions had been used within each paper but that each candidate's route remained constant,
- or how much difference it would make if candidates had chosen a different route through the same qualification and answered different questions as a result.

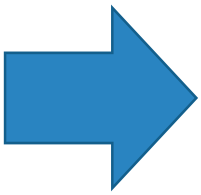
In this article we will concern ourselves with the former of these. That is, we are interested in evaluating how much difference it would make to results if each candidate's route through the qualification remained constant but a different set of questions were included in each of the 30 papers. Putting it another way, we wish to calculate what percentage of the variance in candidates' scores is attributable to their underlying mathematical ability as would be demonstrated if they were able to answer an infinitely large number of questions covering the skills assessed by their chosen route through the qualification.

The Idea

In order to evaluate the reliability of our qualification we will make use of the method of split-halves. Given that the usual definition of assessment reliability is "the consistency of…measurements when the testing procedure is repeated on a population of individuals" (AERA & NCME, 1999), the most intuitive way we might seek to measure reliability is to get candidates to take two versions of the same test and compare their scores. However, this would require candidates to spend additional (and possibly unnecessary) time taking a second version of the same test. To circumvent this issue the split-half procedure instead splits a single question paper into two halves, and then explores the extent to which test scores are 'repeated' from one half of the question in a test to another[4]. If all candidates tend to have similar scores across both halves of the test then we infer that the exact choice of questions has little impact on achievement as the set of questions in one half give a similar result to the entirely different set in the other. Thus, we can be confident that had we written another version of the test and got candidates to take that instead, their results would still be largely unaffected. Conversely, a massive difference between scores on different halves would indicate that candidates' performances were highly dependent upon the precise choice of questions, so that another version of the test may have led to very different results. The formulae used to convert comparisons of scores in different halves into an overall reliability coefficient rely on the correlation (or covariance) of scores between halves, and have been in existence for more than 70 years (see Rulon, 1939).

A simple example of how a test might be split into halves is shown in Figure 2. This figure is based upon the scores for one particular candidate taking the Core Mathematics 1 paper in June 2012. In this case their total score on 10 questions out of 72 is split into two total scores each based on 5 questions and out of 36. Although, this particular candidate has raw scores that are similar across the two halves, a full calculation of reliability would require separate scores on each half to be calculated for each candidate and an estimation of the correlation (or covariance) between the two.

Question	Score
1 (3 marks)	3
2 (5 marks)	3
3 (5 marks)	4
4 (6 marks)	6
5 (6 marks)	4
6 (7 marks)	5
7 (6 marks)	4
8 (8 marks)	7
9 (11 marks)	8
10 (15 marks)	15
Total (out of 72)	59



	Half 1	Half 2
		3
	3	
		4
	6	
	4	
		5
		4
	7	
	8	
		15
Totals (out of 36)	28	31

Figure 2: Example of split half scores for one candidate taking Core Mathematics 1 in June 2012

The great advantage of the split halves technique in our scenario is that it automatically handles the issue of multiple routes through a qualification. Note that all of the possible routes through the Mathematics A level, as defined by Figure 1, require candidates to take exactly six assessments. Thus, if we were to split all 30 assessments into halves thus creating 30 half 1s and 30 half 2s, regardless of a candidate's route through the qualification, they will have scores from exactly six half 1s and six half 2s. Thus if we add up all 'half 1' scores to make one total and all the 'half 2' scores to make another, we can produce two 'half A level' scores for each candidate. These scores can then be compared in the usual way to estimate reliability for the A level as a whole. In applying this technique we are only examining the *reliability* of the scores, that is, the extent to which the achieved scores would be replicable if different questions were used in each paper. The question of whether all possible routes through a qualification are equally valid is not addressed. Furthermore, the overall reliability coefficient generated in this way will essentially provide an average level of reliability across all the possible routes. It does not examine whether particular routes provide a more reliable final score than others.

Note that the technique suggested here could equally well be used to examine the reliability of scores comprised of results in different subjects. For example, we could theoretically apply a similar method to examine the reliability of candidates' UCAS scores that combine A level performance across numerous subjects and are used for university

3. Although as discussed by Benton (2013b), because each question must be marked, it is likely that such estimates will also account for a proportion of marking unreliability.
4. Technically we tend to be interested in changes in standardised scores rather than raw scores. That is, the change in each candidate's score after accounting for any changes in the overall mean and standard deviation of scores across all candidates. This value is of more interest as such overall changes to the score distribution are likely to be accounted for within the process of grade awarding in any case.

applications in the UK. This would address the question of the extent to which applicants' UCAS scores are dependent upon the precise set of questions included in their particular examinations. This would not address the issue of whether all UCAS scores are equally valid predictors of university performance or whether all subjects provide equally reliable scores; it would simply provide an average reliability coefficient across the different subject choices chosen by candidates.

Which split half?

As can be seen from Figure 2, there are numerous possibilities for how we should split a single test into two parts (from now on referred to as 'halves' even though they may not be of equal size). In fact, if we imagine that question 1 is always in half 2, then each other question is either in the same half as question 1 or the opposite side. Thus there are 2^9=512 ways to split this test into halves minus the one split with all questions on the same side. In Figure 2 we have focussed on ensuring that the same number of items and the same number of marks are available in each half. However, although this is intuitively appealing, ensuring similar coverage in terms of the curriculum content and skills required by each half is probably more important.

If we make the (reasonable) assumption that scores on questions measuring the same skills are likely to have stronger associations, then we can encourage each half to measure similar skills by looking for the split that maximises the association between scores on one half and scores on the other. Maximising the association, as measured by covariance rather than by correlation, is advantageous in that it will encourage each half to have a similar score distribution. This approach has been adopted by numerous authors and the resulting reliability coefficient is sometimes referred to as Guttman's λ_4 (after Guttman 1945, see Callender & Osburn 1977; Ten Berg & Socan 2004).

Compared to Cronbach's alpha, Guttman's λ_4 is less likely to underestimate the reliability of a test (Ten Berg & Socan, 2004). On the other hand, there is a danger that, in small samples or with very large numbers of available items, it may grossly overestimate reliability (Ten Berg & Socan, 2004). That is, because it focusses on finding the best split half, it may overestimate the likely similarity between candidates' scores on two real parallel versions of a test. However, this issue was investigated further by Benton (2013a) and is unlikely to be a concern in our scenario as all of the 30 papers investigated contained ten questions or fewer and all but three had sample sizes numbering in the thousands.

Finding the best split half

As explored by Benton (2013a), there are several possible algorithms for identifying the best split half. For the purposes of this study we used the 'start-then-improve' algorithm. As suggested by the name, this algorithm begins with an initial split of the items into two halves (such as based upon an odd and even split) and then examines whether swapping any pair of items from opposite sides will improve the strength of association between the two sides. In essence this means that items that are found to be more strongly associated with the overall score on their own half than the overall score on the opposite half are likely to be swapped across. Once a swap has been made, the algorithm looks for further swaps that may improve the association between scores on opposite sides. This continues until there are no remaining swaps that will improve this association any further.

An example of how this algorithm works in practice is shown in Table 1. Initially the questions are split such that all the odd numbered questions are in half 1 and all the even numbered questions are in half 2[5]. The top section of the Table 1 (labelled 'step 1') examines the improvement in the covariance between scores on the two halves that would result from swapping any question in half 1 with any other question in half 2. For example, swapping question 1 to half 2 and question 2 the other way would increase the covariance between halves by 1.27. Note that, questions cannot swap with themselves and that questions cannot swap from a half if they are not already included in that half. This leads to the regular pattern of 0s in the matrix. Note that, the algorithm also considers swapping any question to the opposite half without moving another question in the opposite direction. This possibility is explored in the last row and last column within each step in Table 1.

All swaps that would lead to a positive change are highlighted in blue and the swap leading to the greatest improvement (item 1 swapping with item 8) is highlighted in green. Once this is done, we can then recalculate the improvement in covariance from any subsequent swaps ('step 2'). In fact, only one swap (item 5 with item 4) leads to any improvement. Once these items are swapped, there are no possible improvements from further swaps ('step 3'). This means the final split has questions 3, 4, 7, 8 and 9 in half 1 and questions 1, 2, 5, 6 and 10 in half 2.

Results

The algorithm described above was applied to each of the 30 papers detailed in Figure 1. The half scores on each paper were rescaled so that, for each candidate, the total of their scores on the two halves equalled their total UMS score[6] for each paper as a whole rather than their total raw score[7]. This means that, for each candidate, the total of their 12 half-paper scores equalled their total UMS score for the A level as a whole – the score used to determine their final grade.

Rather than simply comparing the total of the scores on all the first halves with the total of the scores on all the second halves, we applied a best split of best splits method to ensure both halves are representative of a full A level. Having applied this method, we finally have scores on two 'half A levels' each comprising of total scores across a mixture of half 1 and half 2 scores from different units so as to maximise the association.

Figure 3 compares the scores on each half A level for a random sample of 1000 candidates. As can be seen, there is a very strong relationship between the two scores with the majority of candidates displaying close agreement. Table 2 displays the mean and standard deviation of scores for the whole cohort on each half. As can be seen, the distribution of UMS scores is fairly similar on each half.

The reliability of Mathematics A level as a whole is estimated via the association between the two halves. Overall, there was a very strong correlation between halves (0.928). This can be combined with the Spearman–Brown formula to generate an overall reliability estimate of 0.963. An almost identical reliability coefficient can be generated based upon the covariance between the two halves and using the formula of Rulon (1939); that is, the usual formula for Guttman's λ_4.

5. In practice, more than one initial starting split is used in order to ensure that the optimal split is identified.
6. Uniform Mark Scale. See AQA (2009) and Gray and Shaw (2009) for details.
7. In order to achieve this, the each candidate's total UMS score was divided between the two halves according to the proportion of their total raw score that was achieved on each half.

Table 1: Example of the algorithm used to find the optimal split for Core Mathematics 1 in June 2012

			Question to swap from half 2										
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*None*
	Step 1	*1*	0	1.27	0	1.80	0	1.94	0	2.19	0	-1.64	-0.84
		2	0	0	0	0	0	0	0	0	0	0	0
		3	0	-1.06	0	-0.62	0	1.17	0	1.20	0	1.25	-5.36
		4	0	0	0	0	0	0	0	0	0	0	0
		5	0	-0.72	0	0.03	0	1.46	0	1.60	0	0.63	-4.56
		6	0	0	0	0	0	0	0	0	0	0	0
		7	0	-2.32	0	-1.65	0	0.80	0	0.58	0	1.38	-7.26
		8	0	0	0	0	0	0	0	0	0	0	0
		9	0	-3.90	0	-3.42	0	-0.49	0	-0.50	0	2.12	-9.82
		10	0	0	0	0	0	0	0	0	0	0	0
		None	0	1.31	0	1.89	0	1.48	0	1.85	0	-3.18	0
	Step 2	*1*	0	0	0	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0	0	0	0	0
		3	-0.99	-0.22	0	-0.23	0	-0.78	0	0	0	-5.38	-1.67
Question		*4*	0	0	0	0	0	0	0	0	0	0	0
to swap		*5*	-0.59	-0.13	0	0.17	0	-0.75	0	0	0	-6.25	-1.13
from		*6*	0	0	0	0	0	0	0	0	0	0	0
half 1		*7*	-1.61	-0.42	0	-0.19	0	-0.09	0	0	0	-4.18	-2.50
		8	-2.19	-0.92	0	-0.39	0	-0.25	0	0	0	-3.82	-3.03
		9	-2.69	-0.90	0	-0.87	0	-0.28	0	0	0	-2.34	-3.97
		10	0	0	0	0	0	0	0	0	0	0	0
		None	-0.33	-1.55	0	-1.41	0	-4.17	0	0	0	-13.50	0
	Step 3	*1*	0	0	0	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0	0	0	0	0
		3	-1.16	-0.23	0	0	-0.40	-0.68	0	0	0	-5.17	-1.96
		4	-0.76	-0.30	0	0	-0.17	-0.91	0	0	0	-6.42	-1.30
		5	0	0	0	0	0	0	0	0	0	0	0
		6	0	0	0	0	0	0	0	0	0	0	0
		7	-2.05	-0.69	0	0	-0.36	-0.26	0	0	0	-4.23	-3.06
		8	-2.32	-0.89	0	0	-0.56	-0.12	0	0	0	-3.58	-3.28
		9	-3.10	-1.14	0	0	-1.04	-0.42	0	0	0	-2.37	-4.49
		10	0	0	0	0	0	0	0	0	0	0	0
		None	-0.22	-1.26	0	0	-1.58	-3.78	0	0	0	-12.99	0

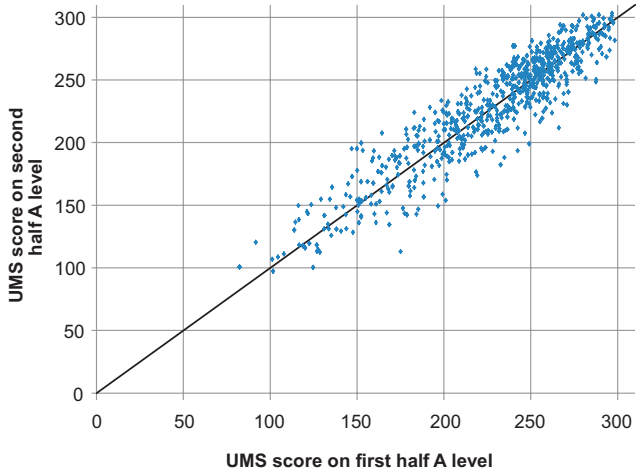


Figure 3: The relationship between scores on overall split halves of a Mathematics A level (n=1,000)

Table 2: Descriptive statistics for each "half A-level" (n=11,771)

Half	*Mean score (UMS)*	*Standard Deviation*
First	225.3	45.0
Second	226.5	45.7

Summary

This article has demonstrated how a method of split halves can be used to calculate the reliability of a complex qualification. In contrast to some approaches to this problem, based upon item response or classical test theory, the recommended approach does not start with a pre-conceived model for the way in which scores on different items will relate to one another. Rather, the underlying model is implicitly built up through the search for the most appropriate splits. This allows us to directly examine the extent to which scores, possibly representing skills across multiple domains, remain consistent across different sets of questions. This same method could be used to estimate reliability for any qualification where multiple routes are possible, including qualifications with options with regard to topic choices.

In the case of Mathematics A level, the analysis reveals an extremely high level of reliability with almost 97% of the variance in scores attributable to the underlying mathematical ability of candidates as would be demonstrated if they were able to answer an infinitely large number of questions covering the skills assessed by their chosen route through the qualification. This implies that the impact of the exact selection of questions seen by any candidate is extremely small.

Any internal estimate of reliability requires some assumptions. In particular, the split-half approach recommended in this paper assumes that the skills measured by one half are equivalent to those measured

by the other. This requires that there are no particular skills that are only assessed by a single question in any exam paper, as a single question can by definition only occur in one half. Thus, although our approach is intended to maximise the similarity between halves, we cannot be certain that the two halves of any given paper measure exactly the same set of skills. In this technical sense, the reliabilities derived via this method may be viewed as a lower bound on the true level of reliability. Having said this, as demonstrated by Benton (2013b), genuine alternative versions of the same test may also be less 'parallel' than would be desirable in a technical sense. In this way, from a practical perspective, it may be more reasonable to view the estimates as accurate, but slightly optimistic.

One limitation of the suggested method is that it only works if all units, taken within a qualification can be split into parts. This is not universally the case, for example, if one unit of the qualification comprises of a single, non-dividable mark for coursework. However, provided such elements only comprise a minority of the qualification, it will still be possible to provide a reasonably accurate estimate of reliability by adding the score from this non-dividable element to one of the two 'half A level' scores derived for the remainder of the qualification as described above. This approach would provide a reliability estimate at least as good as the classical test theory composite reliability approach suggested (amongst other approaches) by Bramley and Dhawan (2013).

As stated earlier, whilst the method suggested here estimates an overall reliability coefficient, it does not investigate whether all routes provide equally valid, or even equally reliable test scores. However, users of test scores such as employers, or university admissions, are unlikely to be aware of the route an applicant has taken through a qualification and will only see their final result. From their point of view, the ability to quantify the general level of reliability for a qualification, and in the case of Mathematics A level verify an extremely high level of reliability, may be important, even if it does not necessarily apply to every possible route. Alongside this is the ongoing duty of qualification providers to ensure that all of the individual assessments underlying the different routes through a qualification are themselves individually reliable.

References

American Educational Research Association and National Council on Measurement in Education, (1999). *Standards for Educational and Psychological Testing.* Washington, D C: AERA.

AQA. (2009). Uniform Marks in GCE, GCSE and Functional Skills Exams and Points in the Diploma. Retrieved from http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF.

Benton, T. (2013a) *An empirical assessment of Guttman's Lambda 4 reliability coefficient.* Paper presented at the 78th Annual Meeting of the Psychometric Society, July 2013. Available from: http://www.cambridgeassessment.org.uk/Images/141299-an-empirical-assessment-of-guttman-s-lambda-4-reliability-coefficient.pdf.

Benton, T. (2013b). Exploring equivalent forms reliability using a key stage 2 reading test, *Research Papers in Education, 28*(1), 57–74.

Bramley, T. and Dhawan, V. (2013). Problems in estimating composite reliability of 'unitised' assessments, *Research Papers in Education, 28*(1), 43–56.

Callender, J. and Osburn, H.G. (1977). A method for maximizing split-half reliability coefficients, *Educational and Psychological Measurement, 37*, 819–825.

Gray, E. and Shaw, S. (2009). De-mystifying the role of the Uniform Mark in assessment practice: concepts, confusions and challenges. *Research Matters: A Cambridge Assessment Publication 7:* 32–37.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255–282.

He, Q. (2009). *Estimating the Reliability of Composite Scores.* Coventry: Ofqual. Available from http://ofqual.gov.uk/documents/estimating-reliability-composite-scores/.

Rulon, P. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9*, 99–103.

Ten Berge, J., and Socan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69*, 613–625.

An intra-board comparison at syllabus level based on outcomes of rank-ordering exercises at component level

Louis Yim Cambridge International Examinations (The author is currently at the Singapore Examinations and Assessment Board, Singapore.)

Introduction

Ensuring the equivalence of standards of similar qualifications across different awarding bodies or across time within the same awarding body has been a salient area of research in educational assessment in England for some time. For the former, the rationale behind this research is that a number of examination boards in England offer public examinations which lead to the same qualifications, i.e. GCE A level and GCSE. Although each examination syllabus must conform to general qualifications criteria approved by the examinations regulator[1], and also to a common core of subject content, the syllabuses may differ between boards in other respects. A crucial question of whether it is equally difficult to obtain a given grade in a particular examination with one board than with another arises. In fact, this issue is not limited to England alone, but extends to other countries where candidates sit examinations which are claimed to be equivalent qualifications to the GCE A level and GCSE. For examinations taken within the same awarding body but at different times, the issue of equivalence of standards is more commonly known and is about maintaining the same standard for a particular examination between different administrations (e.g. in different years) within an awarding body. For example, the standard of a given grade in a particular examination from three years ago should be comparable to the same grade in an examination a year later within an awarding body; or the standard of a given grade between two administrations (e.g. in different time zones) from the same examination session within an awarding body.

Rank-ordering is one of many comparability methodologies, and has been used relatively effectively to compare standards quantitatively between two exam sessions at component level[2] (Bramley, 2005; Bramley, 2007). Such a method has been modified to measure the equivalence of standards at syllabus level, based on examiners' holistic evaluation of scripts from prescribed components for each syllabus. Several studies (Yim, Shaw & Lewis, 2008; Yim & Shaw, 2009) have been conducted to demonstrate its feasibility and capability. The method has been used in both inter-board (Yim & Forster, 2010) and intra-board (Yim, 2012) studies pertaining to Cambridge International Examinations' (CIE) time zone question papers administered within the same exam session. Results so far have shown that the rank-ordering method could, to a large extent, produce comparable results when conducted repeatedly[3]. The qualitative feedback from questionnaire responses, on the other hand, revealed that some expert judges lacked confidence in their final rank-order judgements because the method's large cognitive demand requires them to retain the script information from several candidates holistically before making rank-order decisions. Interestingly, the quantitative results supported their judgements (Yim, 2012).

This paper describes a variant comparability methodology which uses the rank-ordering method at component level to derive results at syllabus level for intra-board comparison. In other words, instead of judges holding several components' information about each candidate in their minds and making a holistic evaluation of individual candidates during comparison, judges only rank-order candidates' performance within each prescribed component. The final rank order at syllabus level of each judge is derived based on his/her component level's rank orders. This variant methodology could enhance judges' experience during the exercise, as well as generate quantitative evidence of comparison for each prescribed component in order to inform threshold adjustment at component level during grading, in addition to the syllabus level only evidence from the holistic approach. This piece of information should provide an improvement in terms of clarity for grading advice, compared to that for the syllabus level only methodology.

The rationale behind conducting research at syllabus level is that quantitative results can generally help inform CIE's grading decisions in terms of threshold adjustment of an entire option/syllabus. The materials used in this study were question papers, mark schemes and syllabus specifications. Real candidates' component scripts with the same scheme of assessment and subject content from the same examination session within the same examination board were used in this study. These were then evaluated by external consultants (or judges) to generate rankings of candidates' scripts for each component. The rank-order data for each component were analysed using the multifacet Rasch modelling technique (Linacre, 1987). The outputs (or 'measures') from each component were combined by a weighted average method to generate the overall measure at syllabus level. The difference in standards between candidates' scripts at component as well as syllabus levels was deduced from the graphs. The methodology, the research outcome, and judges' feedback are described in detail below.

Background to comparability exercises

In this context, comparability is concerned with the application of the same standard across different examinations (Newton, 2007). The purpose of inter-board comparability studies is to compare standards across different examination boards. In making this comparison, it is important to distinguish between *content standards* and *performance standards*: "Content standards refer to the curriculum (or syllabus/specification) and what examinees are expected to know and to be able to do ... performance standards communicate how well examinees are expected to perform in relation to the content standards" (Hambleton, 2001). In fact, a more precise definition of comparability is paramount since many different aspects of qualifications can be compared, such as the demand of the curriculum, similarity of content materials,

1. The Office of Qualifications and Examinations Regulation, England.
2. In CIE, an exam *syllabus* usually comprises several ***components*** which assess different areas of skills/competencies in order to cover the subject knowledge to be assessed. For example: *Component 1:* Algebra, *Component 2:* Calculus, and so on, in a Maths *syllabus*. A component level comparability means, say, only comparing *Component 1s* between 2010 and 2013 exam sessions.
3. The same set of scripts from prescribed components within each concerned syllabus was used in two separate research studies, i.e. inter- and intra-board comparisons.

difficulty experienced by candidates, demand of assessment materials, perceived quality of candidate outcome based on scripts and standards of attainment, etc.

One way to compare performance standards across assessments from different boards (or across parallel assessments from the same board) is to ask experts to compare pairs of scripts from each assessment and make judgements about which one demonstrates better quality. Such exercises address the question: "Which syllabuses' grade boundary scripts[4] are perceived by expert judges to be of better quality (after allowing for slight differences in syllabus content, question paper and mark scheme difficulty)?"

One way of analysing the data from these paired comparison judgements is by Thurstone's model (case 5) for comparative judgements (Thurstone, 1927). For a discussion of how Thurstone's method has been applied in the context of examination comparability, see Bramley (2007). For recent applications of the method see Yim, Shaw and Lewis (2008), and Yim and Shaw (2009).

The main advantage of this approach is that the use of candidates' scripts provides explicit evidence of the knowledge, understanding and skills of examinees. As such, direct comparison of performance standards can be achieved. For inter-board comparisons it should be noted that it is only possible to compare performance standards if the content standards across the examination boards are similar enough for the different assessments to be considered to be measuring the same construct (underlying trait). If the question papers, mark schemes and syllabus specifications are very different, examiners will be expected to make judgements about the relative performance standards in a context of possible differences in content standards. The outcome of such an exercise would be rendered less reliable due to disparate schemes of assessment and syllabus contents.

In practice, the nature of the scripts (objects) being compared is such that the scripts take a long time to read, and paired comparisons are unlikely to be independent, because of the repeated use of shared scripts. Hence examiners might already have the knowledge of either or both of the scripts before the paired comparisons, which violates the assumption of local independence between paired judgements. Therefore instead of asking judges to make paired comparisons, it is less time-consuming to ask them to put sets of scripts into rank-order of perceived quality. It is then possible to extract paired comparison data from the rank-order in the form of '1 beats 2', '2 beats 3', '1 beats 3' and so on (Bramley, 2007). These extracted paired comparisons are not statistically independent, because they are constrained by the ranking, but as explained above even genuine paired judgements would arguably not be independent either. In other words, a rank-ordering method is a time-saving variant of the paired comparison method for comparing performance standards. Such comparison exercises draw heavily on the expertise of senior examiners, and their ability to judge the quality of examinees' work, taking into account the demand placed upon examinees by the individual syllabuses/specifications, question papers and mark schemes.

Method

This study rank-ordered each prescribed intra-board component individually at component level using the same procedures as Yim and Forster (2010) with respect to the algorithm for selecting real candidates, the pack design, the instructions given to expert judges, and the data analysis method. Each judge's rank-orders for each prescribed component were then fed into the FACETS software (Linacre, 1987) to generate the outcome for the multifacet Rasch analysis, which would then be presented in graphical form for standards' comparison at component level. The outcome (or 'measure') of each component was then standardised by linear scaling such that they could be combined with each other in association with the weighting factor of each component specified in the syllabus specification, i.e. standardised weighted average, to yield the measure at syllabus level. The advantage of this approach is that the amount of script information that judges hold cognitively before making a rank-order decision is reduced, which is likely to improve on the accuracy of the rank-order results, enhance judges' ranking experience, and help boost their confidence in the exercise. Furthermore, the weighting factor of each component is applied during the generation of the weighted average at syllabus level. This is in contrast to the method used in a holistic evaluation approach, in which the application of a weighting factor could be less rigorous. Quantitative results in terms of differences in standards at component level can be generated in addition to those at syllabus level to inform grade boundary adjustment during awarding meetings[5] if there is a need to align standards with another assessment option (or exam board in the case of inter-board comparison).

The materials used in this project were question papers, mark schemes, syllabus specification and real candidates' scripts from the examination board. The first assessment is referred to as 'Option AA' and the second as 'Option BB' in this article. Each option has the same three components, namely, multiple choice (Component 1), structured questions (Component 2) and analysis and critical evaluation (Component 3). Thirty-four (or 17 from each assessment option) exact 'flat' profiles of real candidates' scripts at grade boundaries, A, B, C, D and E, and their intermediate grade boundaries at 2/3 and 1/3 of a grade above each grade, and 1/3 and 2/3 of a grade below each grade for both assessments were selected. A candidate with an exact 'flat' profile on a three-component assessment could be a candidate who achieves a mark exactly at the grade boundary of, say, B[6] at syllabus level with all three components also being at a mark exactly at the grade boundary of B; a candidate with an uneven profile could achieve a mark at the grade boundary of B at syllabus level, but with uneven grades at component level, for example, a mark at well above grade A in Component 1, a mark at the boundary of grade B in Component 2 and a mark at the middle of grade C in Component 3. The latter is more common/authentic in examination practice. The use of the exact 'flat' profile is to indicate to judges that a clear-cut standard across component level, for example, all components at the boundary of grade B, will lead to the same syllabus grade level, that is, grade B.

As a result of using the exact 'flat' candidate profile, real candidates whose script components' marks fit within ±1% of each targeted component mark at particular syllabus marks/grade levels were selected. It should be noted that the selection of real candidates' scripts meeting this criterion can only work well in an examination with a large entry, because there are enough scripts to choose from.

4. Grade boundary scripts are scripts whose marks are exactly at the grade boundaries which were set during a grading (or an awarding) meeting.

5. At awarding meetings the grade boundary locations on the raw mark scale of each component are decided.

6. There is a subtle difference between a candidate with an exact even (or 'flat') profile and one with an even profile in this discussion. The criteria of the former are a candidate with the targeted component marks at exactly the same point relative to the grade boundary; whereas the latter only requires the same grades across prescribed components (e.g. BBB) within a syllabus and no stipulation of any targeted component marks.

After selecting the real candidates' scripts, examiner markings/annotations were removed electronically via a scanner so that they did not have an influence on the rank-ordering judgements during the experts' judging process. Each candidate was then allocated into different packs of scripts in accordance with the pack design at component level. An example of the pack design layout for component 1 is illustrated in *Appendix A* for readers' reference; other components follow the same pack design layout. Each pack comprised six candidates (three from Option AA and three from Option BB). Altogether there were eight packs (A to H) for component 1; eight packs (J to Q) for component 2; and eight packs (R to Y) for component 3. The candidates and hence their scripts in each pack were randomised, coded and labelled such that the original scripts' rank-order based on marks was concealed.

The same pack design was used for each component, i.e. the same set of candidates appeared in packs A, J and R, etc. Each candidate's scripts were photocopied for each expert judge.

In each pack of six scripts for each component, two were common to the pack above and two were common to the pack below (where 'above' and 'below' refer to the rank order by total mark). The top pack had two scripts in common with the pack below and the bottom pack had two scripts in common with the pack above. This linked design allowed a common scale of 'perceived quality' to be created from the ranking judgements.

Five senior examiners (expert judges), all with marking/moderating experience of the syllabus concerned, were recruited to make judgements about the real candidates' scripts. Their task was to rank-order scripts within each pack from best (highest quality = 1) to worst (lowest quality = 6) on each component and record their outcomes in the tables provided on a record sheet. Each expert judge was asked to complete a questionnaire towards the end of the exercise for the qualitative analysis of the study.

Analysis and results

Once the rank-order data at component level were received from judges, data for each component were deconstructed into paired comparison data and then analysed using the Rasch analysis (FACETS) software to estimate the difficulty/ability of each script/candidate for each component based on the inter-relationship of examiners' rankings. It should be noted that the percentage mark at component level, instead of a raw mark, was used in the analysis in order to achieve a common scale for both Options. The FACETS outputs are given in Appendices B, C and D for component 11 vs. 12, component 21 vs. 22 and component 31 vs. 32 respectively. The separation reliability index (analogous to Cronbach's Alpha) was high in all three cases, i.e. 0.99, 0.98 and 0.97 from Appendices B, C and D respectively, showing that the variability in perceived quality among the scripts could not be attributed to chance. There are different views on what fit index is actually acceptable; McNamara (1996) suggests that the usual limits of acceptability are the mean ±0.3 (so anything between 0.7 and 1.3 will be acceptable). According to Lunz and Wright (1997:83) "Because the interpretation of fit is situationally dependent, there are no fixed levels for fit statistics acceptance or rejection." They go on to use a level of ±0.5 in their studies. Operational experience, however, would suggest lower and upper bound limits of 0.7 and 1.6 respectively for mean squares to be useful and acceptable for practical purposes; and these were used in this analysis. Fit statistics of 1.7 or greater indicate too much unpredictability in examiners' scores, while fit statistics of 0.6 or less indicate over-fit or not enough variation in examiners' scores. The fit statistics from the infit and outfit columns of the FACET outputs for scripts and judges showed a slight tendency towards over-fit in all three cases suggesting that the judges were perceiving the trait in the same way and that there was less variability in their judgements than modelled. All these scale statistics need to be treated with caution because the paired comparison analysis violates the assumption of local independence between paired judgements when derived from the rank-ordering outcome (Bramley, 2012).

The Measure column in the bottom table of each component in Appendices B, C and D indicates the ability of each candidate's script. After taking the mean and standard deviation of the Measure column of each component and standardising them to the same mean and standard deviation, the total standardised weighted average Measure at syllabus level could be obtained. This was done by combining the respective Measure of candidates' scripts of each component with the weighting factor of each component designated in the syllabus specification. The results/graph of the total standardised weighted average Measure obtained using the component-derived-syllabus approach can then be re-scaled to compare with those evaluated by the holistic approach in 2012 (Yim, 2012), i.e. Figure 5.

Figures 1, 2, 3 and 4 show the results of the comparability plots for the three prescribed components and that for the component-derived syllabus approach respectively. The vertical axis along the left of the figures represents the Measure (or script quality) scale in logodds units (logits). In these graphs each data point (diamond – Option AA and square – Option BB) represents a script. Each script (a data point) is positioned according to its measure. Thus performances are rank ordered with the most able candidates at the top of the axis and the least able at the bottom, that is, the scripts in the top half of the graph (above 0 logits) are judged to be of better quality than those in the bottom half (below 0 logits). The horizontal axis shows the component/overall syllabus aggregate percentage mark obtained from conventional marking of the scripts.

The two straight lines in each comparability plot shown in Figures 1, 2, 3 and 4 are linear regression lines whose equations are given in the boxes. It should be noted that the legend 'linear' in the graphs refers to the regression line, and not to a linear exam (as opposed to a modular exam). The parameter *R* is the correlation coefficient. The magnitude of *R* indicates the extent to which the two sets of measurements (Measure and Syllabus %) are linearly related. Each pair of regression lines, that is, Options AA and BB, in the four cases shares similar features such as strong correlation and similar gradient. Figures 1, 2 and 3 provide the comparison between Options AA and BB at different grade boundaries in each component; whereas Figure 4 gives an overall syllabus aggregation with the weighting factor of each component taken into consideration.

Yim (2012) used the same set of scripts and judges as the current study, but used a holistic evaluation approach at the syllabus level. By comparing the results from Yim (2012), a comparison of results from the holistic evaluation approach at syllabus level and those from the current study can be made. Figure 5 shows an intra-board comparability plot using a holistic evaluation approach (Yim, 2012). The pair of regression lines in Figures 4 and 5 shows some similar features: strong correlation, similar gradient, no reversal of position; that is, option AA regression line is consistently above Option BB. Tables 1 and 2 show a comparison of some numerical findings between the holistic approach and the component-derived-syllabus approach.

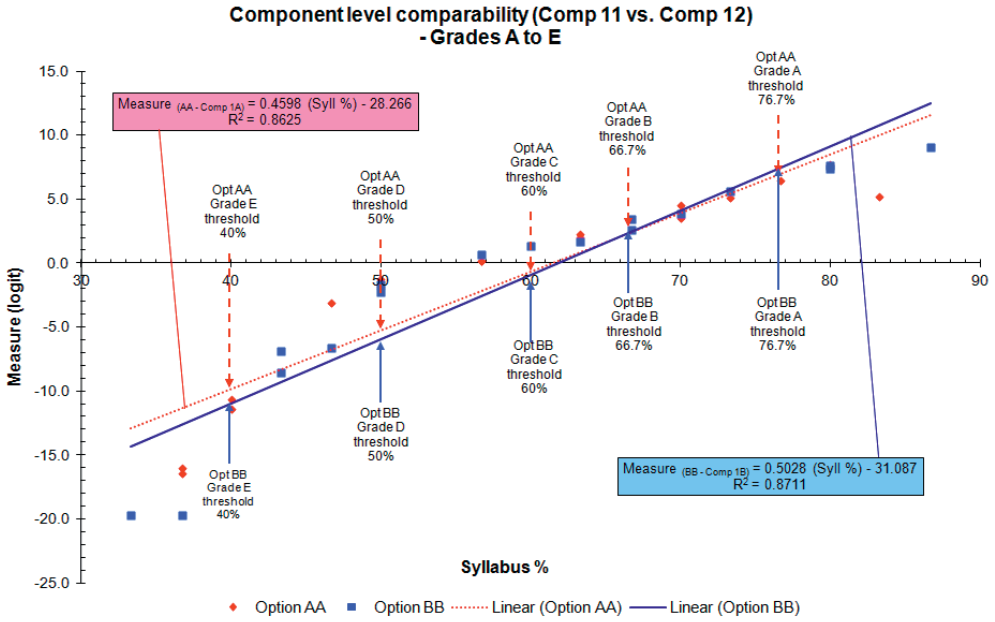


Figure 1: A comparability plot for Component 11 vs. Component 12 from grades A to E for the component-derived-syllabus approach between Options AA and BB.

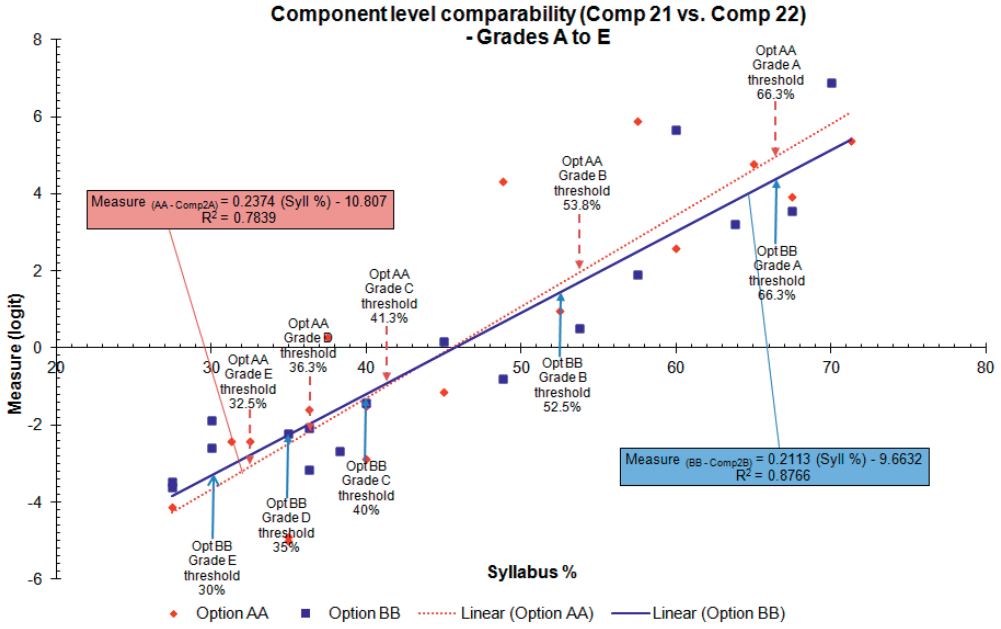


Figure 2: A comparability plot for Component 21 vs. Component 22 from grades A to E for the component-derived-syllabus approach between Options AA and BB.

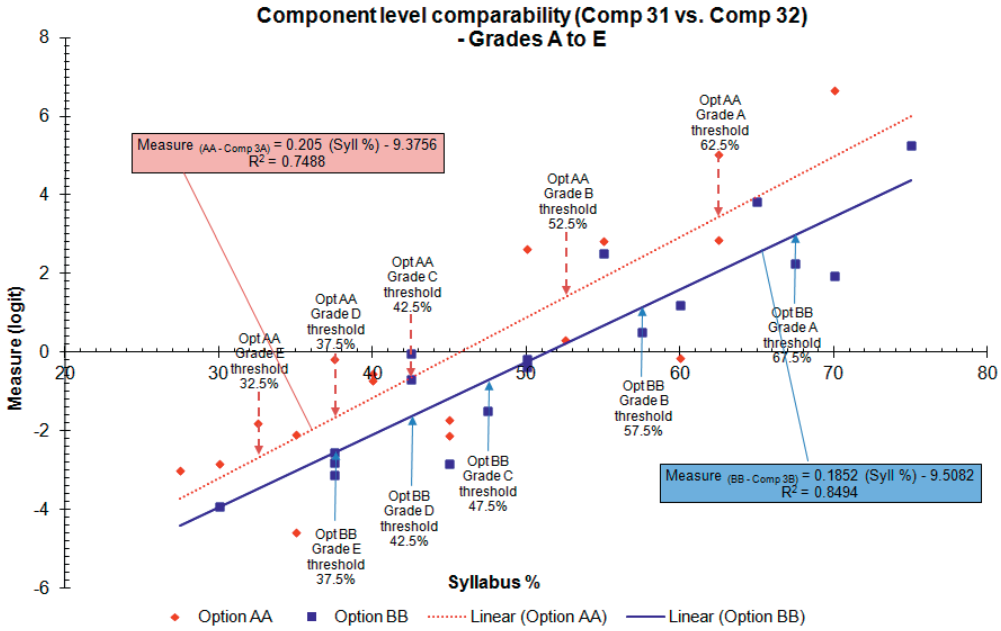


Figure 3: A comparability plot for Component 31 vs. Component 32 from grades A to E for the component-derived-syllabus approach between Options AA and BB.

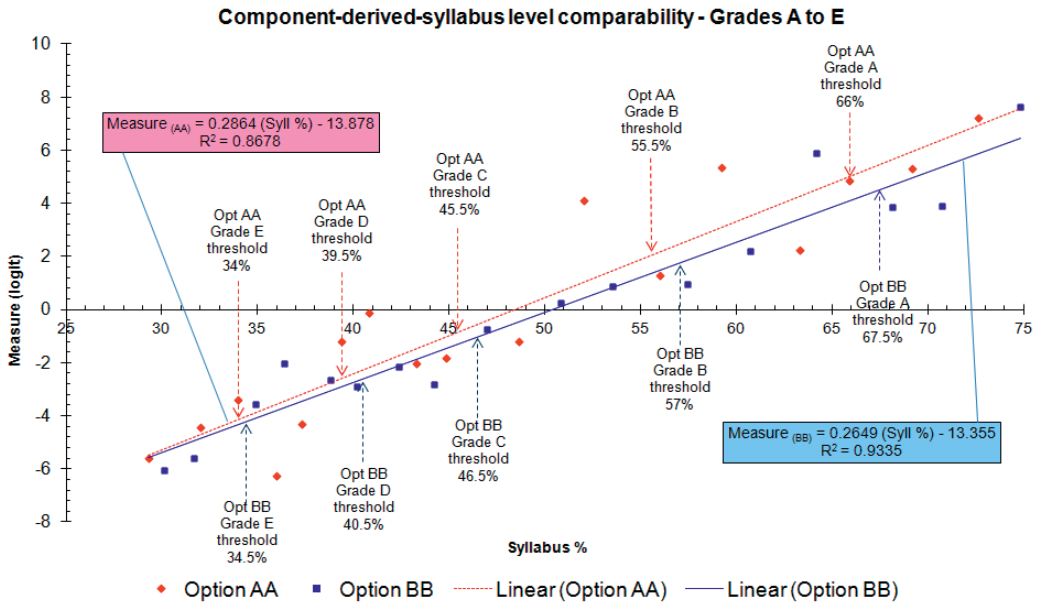


Figure 4: A comparability plot at syllabus level from grades A to E for the component-derived-syllabus approach between Options AA and BB

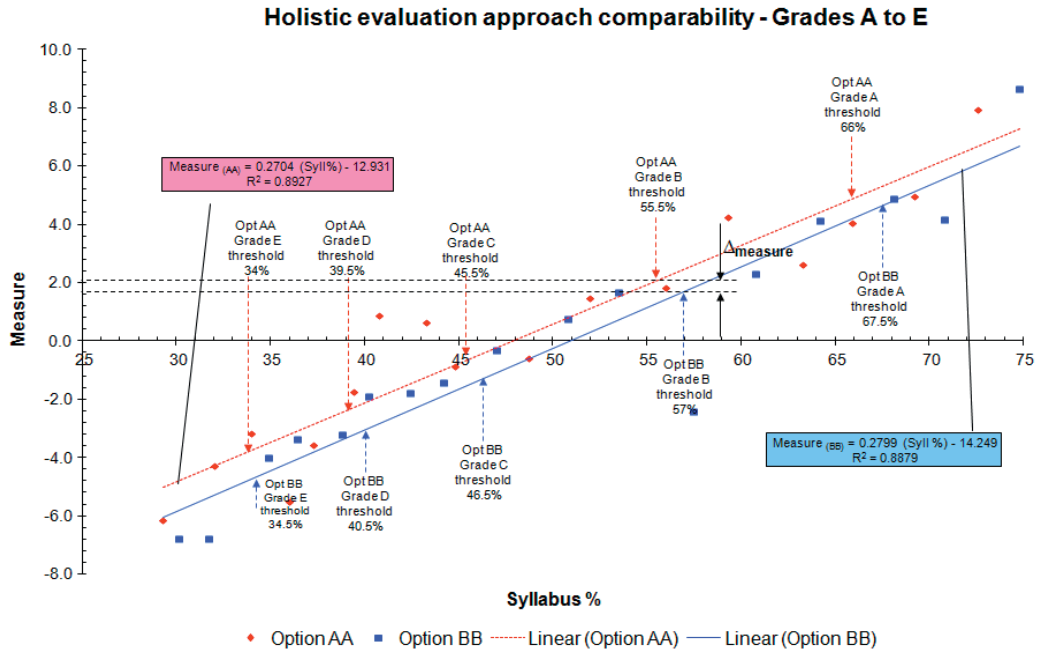


Figure 5: A comparability plot at syllabus level from grades A to E for the holistic evaluation approach between Options AA and BB (Yim, 2012).

Table 1 shows the differences in Measure (along the y-axis) between Option AA and Option BB at Grades A, B, C, D and E for both holistic evaluation and component-derived-syllabus approaches. In an ideal case the values of Δ*measure*, as shown in Figure 5, in both holistic evaluation and component-derived-syllabus approaches should be the same, but the differences in Table 1 suggest that there are disparities at all grades, albeit small, i.e. below or well below one logit. In other words, the recommendations for grade boundary adjustments at syllabus level to achieve the equivalence of standards between options are different depending on the methodology being used, which is understandable. The small differences between the two approaches at each grade are, in fact, rather encouraging as they demonstrate that the rank-ordering method could, to a certain extent, produce similar results when conducting two rank-ordering approaches.

Table 2 shows a comparison of the correlation coefficient (*R*) between 'Measure' and 'Syllabus %' for the holistic evaluation and component-

Table 1: Differences in '*Measure*' (along the y-axis) between Option AA and Option BB at Grades A, B, C, D and E for both holistic evaluation and component-derived-syllabus approaches.

Methodology	Δ*measure [logit]*				
	A	*B*	*C*	*D*	*E*
Holistic evaluation	0.27	0.37	0.61	0.66	0.86
Component-derived-syllabus	0.498	0.273	0.190	0.061	0.076

Table 2: A comparison of the correlation coefficient R between the holistic evaluation and component-derived-syllabus approaches.

Option	*Methodology*	*Correlation coefficient (R)*
AA	Component-derived-syllabus	0.93
	Holistic evaluation	0.94
BB	Component-derived-syllabus	0.97
	Holistic evaluation	0.94

derived-syllabus approaches in Options AA and BB. The correlations in both cases were very similar within the same assessment and across assessments. The strong correlations ($R \geq 0.93$) in all cases between the 'Measure' and the 'Syllabus %' show that the trait of quality as perceived by the judges was very similar to the trait of quality as rewarded by the mark scheme. It should be recalled that the only difference in terms of the research design between the previous comparability study and the current one was that in this study, the methodology of component-derived-syllabus approach was used rather than the holistic approach. Both assessments were from the same syllabus from the same examination board and assessed by the same group of judges.

Feedback from examiners

Responses on questionnaires were collected from five judges who carried out the evaluation to help understand the qualitative aspects of their rank-ordering experience relating to the overall difficulty of the task, the amount of time taken to rank order the scripts, difficulty compared with the holistic evaluation approach, what made some packs more or less difficult to rank, any differences in the task between papers, and the strategy they deployed.

Overall difficulty of the task

All five participants were senior examiners and had taken part in at least two rank-ordering exercises previously. Four of them found the overall task "fairly difficult" to execute; and one examiner found it "fairly easy". Reasons for difficulty are shown on the left-hand column in Table 3. Those from the previous holistic evaluation approach (Yim, 2012) are listed for reference. Judges tended to take an average of just under 30 minutes per pack during the evaluation as compared to between 40 and 90 minutes per pack in the holistic evaluation approach. It should be reminded that the amount of scripts between the holistic evaluation and component-derived-syllabus approaches were different, i.e. three components versus one respectively. The component-derived-syllabus approach probably took longer overall based on the number of packs being evaluated. Despite this, the judges were more confident about their rank-orders and there was generally no need to re-visit the design packs after the exercise, unlike the holistic evaluation approach. Three out of five examiners thought the length of time for the evaluation varied greatly from pack to pack when ranked by individual component.

Table 3: Overall difficulty of the task encountered by judges during the evaluation phase. Reasons from the holistic evaluation approach (Yim, 2012) are also included for reference.

Component-derived-syllabus approach	*Holistic evaluation approach*
Differences between questions in question papers from both options;	Differences between questions in question papers from both options;
Difficult to retain script information to make judgement on the rank-order;	Difficult to obtain an overview of papers with a number of parts;
Candidates' standards are very close within each pack.	Difficult to retain script information to make judgement on the rank-order;
	Candidates' standards are very close within each pack. (Yim, 2012)

Differences were also reported relating to the ease or difficulty of rank-ordering certain packs. Scripts from more able candidates were the most time-consuming to rank order although they were slightly less problematic as there was perceived to be a wider range of ability instantiated in performances. Scripts from less able candidates were more difficult to rank, and standards were perceived to be more closely grouped. Other factors included the mode of assessment. The MCQ[7] component was easier to rank compared to the written component.

All examiners concurred that the task of rank-ordering individual components was much easier compared with that of the holistic evaluation approach at syllabus level. Examiners articulated that they felt more confident at the end of the exercise with their rank order results when their focus was on the same assessment instrument rather than attempting to compare performance across a number of them. It was necessary to keep less script information 'in mind' in each pack and hence most of them were confident about their results.

All examiners felt that it was possible to carry out the judging for a pack of six candidates with one component paper, while three out of five examiners agreed that it was possible to carry out the judging for a pack of six candidates at syllabus level with three component papers holistically (Yim, 2012). It should be noted that examiners from both exercises managed to complete the research studies well, as suggested by the comparable analyses' results.

Rank-ordering strategy

Examiners were allowed to adopt their own rank-ordering strategy during the evaluation phase though they were not allowed to re-mark the scripts. A variety of strategies were identified as follows:

- Identification of common and indicative questions across question papers to evaluate candidates' ability.
- Identification of questions attempted by less able students: based on examiners' experience, some questions can act as an indicator to distinguish between able and less able candidates.
- Identification of the quality of answers given, e.g. correct terminology, accuracy of diagrams.

Overall judgement of depth and accuracy of answers

No examiner indicated a change of approach as the rank order task became increasingly more familiar. Three out of five examiners commented that they employed the same strategy as for the holistic evaluation approach exercise that they completed a year ago.

Examiners were uncertain as to whether more or less time on each script made any difference to the final rank order. However, in the main, they believed that a reduction or extension in the time taken to undertake the exercise would have little impact on the outcome.

Conclusions

A new component-derived-syllabus rank-ordering approach for intra-board comparability study has been reported in this paper. The aim of this approach is to enhance judges' experience and the quality of results from the evaluation exercise, and to generate quantitative evidence of comparison at component level for grading purposes, in addition to the usual practice of acquiring evidence only at syllabus level. The results showed that the component-derived-syllabus recommendations for grade boundary adjustments at syllabus level were close to the findings recommended by the holistic evaluation approach under the same

7. By manually circling individual candidates' answers on the multiple choice question papers based on their answer strings (original m.c. responses), judges evaluated candidates' answers in relation to each question to rank their performances within each pack design.

boundary conditions (Yim, 2012). The small differences between the two approaches at each grade boundary are, in fact, rather encouraging as they, demonstrate comparable results even though different rank-ordering approaches were used.

In the current study the correlations between perceived quality and aggregate mark were very similar across the component-derived-syllabus approach and holistic evaluation approach. The implication of this finding is that the use of different rank-ordering approaches does not affect how the trait of quality is perceived. This contradicts the initial hypothesis that the application of a weighting factor to individual components could improve the correlation. The *prima facie* evidence of the current study suggests that there is no advantage in terms of using either type of approach in relation to the internal quality of the scale produced (separation reliability and fit), or its correlation with an external variable (aggregate Syllabus % mark). The qualitative feedback from all expert judges suggests that the component-derived-syllabus approach was made much easier by rank-ordering scripts by component rather than by the holistic evaluation approach. They felt confident in carrying out the tasks as well as their rank-order judgements.

Limitations of the study

If judges see the same two scripts in a consecutive design pack, there may be a memory effect which could affect the rank-order results during the evaluation of each component. In the light of this, an evaluation procedure of scrutinising alternate design packs was used. In other words, judges were strongly recommended to evaluate design packs according to the designated sequence: A→C→E→G→B→D→F→H in the instructions; and they were also reminded to complete the evaluation of a design pack fully before moving on the next one. Although this should, to a large extent, minimise the impact of the memory effect, it could not totally eliminate the possibility of memory effects. Since the same full set of scripts was presented to the same judges a year ago (but with different design pack arrangements) and the previous study used the holistic evaluation approach, there is a small chance that some judges could have remembered certain scripts. However, the rank-order data required this time was very different from that in the earlier study, so the impact should therefore be minimal.

Acknowledgements

The author would like to thank Tom Bramley for the discussion of the component-derived-syllabus level comparability methodology.

References

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgement. *Journal of Applied Measurement, 6*(2), 202–223.

Bramley, T. (2007). Chapter 7. Paired comparison methods. In Newton, P., Baird, J., Goldstein, H., Patrick, H. & Tymms, P. (Eds.), *Techniques for monitoring the comparability of examination standards.* (pp.246–294). London: Qualifications and Curriculum Authority.

Bramley, T. (2012). The effect of manipulating features of examinees' scripts on their perceived quality. *Research Matters: A Cambridge Assessment Publication*,*13*, 18–26.

Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In: G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives*. (pp.89–116). Mahwah, NJ: Lawrence Erlbaum Associates.

Linacre, J.M. (2008). FACETS Rasch measurement computer program. Chicago: Winsteps.com.

Lunz, M. E., & Wright, B. D. (1997). Latent trait models for performance examinations. In Jürgen Rost & Rolf Langeheine (Eds), *Applications of latent trait and latent class models in the social sciences*. http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/ltlc.htm.

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

Newton, P (2007). Chapter 1. Contextualising the comparability of examination standards. In Newton, P., Baird, J., Goldstein, G., Patrick, H & Tymms, P (Eds.), *Techniques for monitoring the comparability of examination standards.* London: Qualifications and Curriculum Authority.

Thurston, L.L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 273–286. Chapter 3 in L.L. Thurston (1959), *The measure of values.* Chicago, Illinois: University of Chicago Press.

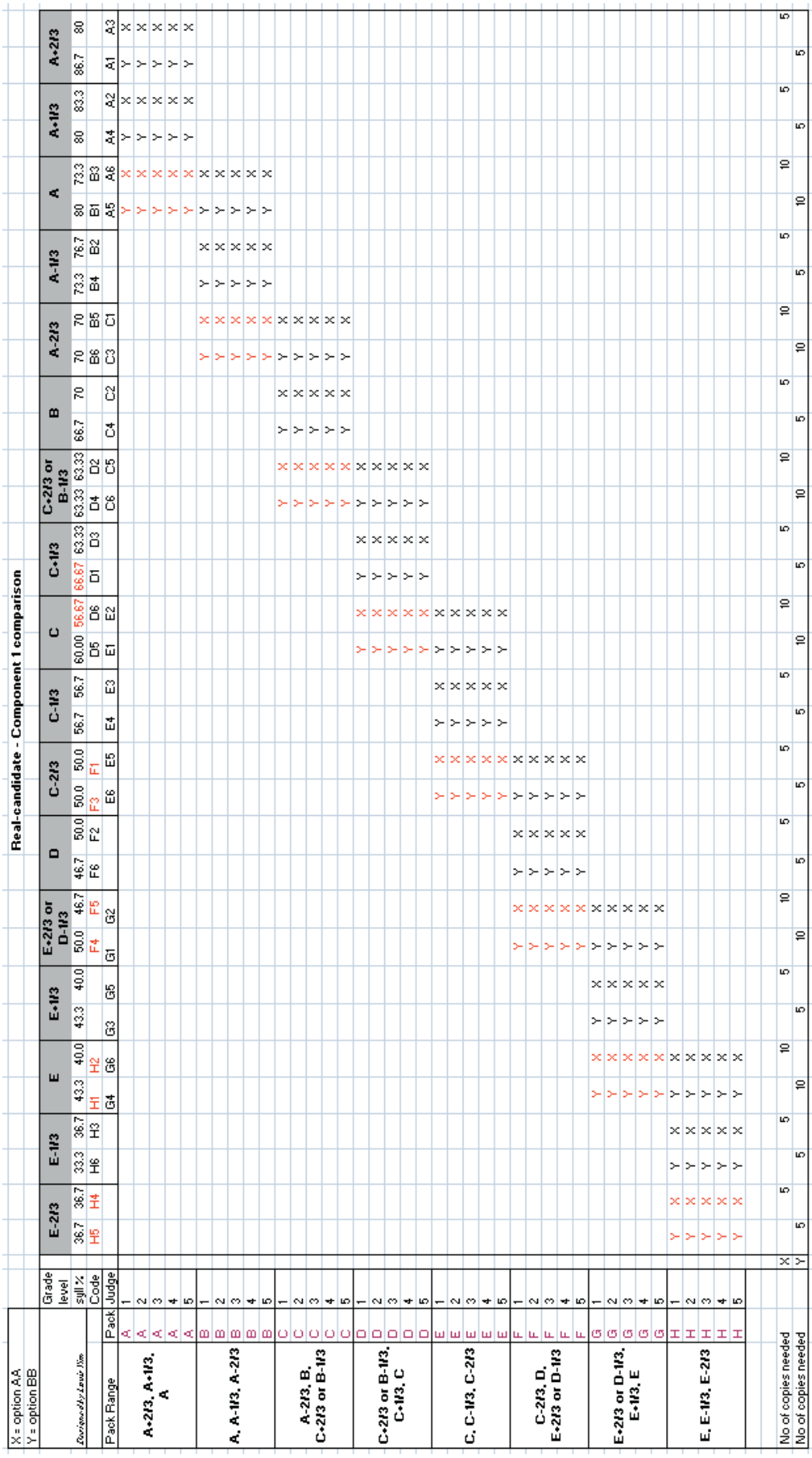
Yim, L.W.K., Shaw, S.D. & Lewis, M (2008). A science comparability study between two exam boards using a rank-ordering methodology at syllabus level. *9th AEA Europe Conference Proceeding, Hisar, Bulgaria,* 6–8 Nov 2008.

Yim, L.W.K and Shaw, D. S. (2009). A comparability study using a rank-ordering methodology at syllabus level between examination boards. *35th IAEA Annual Conference Proceedings, Brisbane, Australia,* 13–18 September 2009.

Yim, L.W.K and Forster, M. (2010). A comparison between the effect of using pseudo candidates' scripts and real candidates' scripts in a rank-ordering comparability methodology at syllabus level. *36th IAEA Annual Conference Proceedings (2010) – Assessment for the future generations, Bangkok, Thailand,* 22–27 August 2010.

Yim, L.W.K. (2012). An Intra-board comparison of the effect of using pseudo candidates' scripts and real candidates' scripts in a rank-ordering exercise at syllabus level. *Research Matters: A Cambridge Assessment Publication*, *14*, 2–9.

Appendix A – A pack design layout for Component 1. Components 2 and 3 follow the same pack design.



Appendix B – FACETS output

Component 11 vs. Component 12

Table 7.1.1 Judge Measurement Report (arranged by mN)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Judge
60	120	.5	.50	.00	.26	1.74	4.7	1.80	1.0	-.11	.50	.62	1 CHS
60	120	.5	.50	.00	.26	.67	-2.8	.43	.4	1.51	.68	.62	2 TC
60	120	.5	.50	.00	.26	.87	-.9	1.08	.7	1.13	.63	.62	3 PC
60	120	.5	.50	.00	.26	.96	-.3	.76	.6	1.09	.63	.62	4 NB
60	120	.5	.50	.00	.26	.68	-2.7	.44	.4	1.49	.67	.62	5 GM
60.0	120.0	.5	.50	.00	.26	.98	-.4	.90	.6		.62		Mean (Count: 5)
.0	.0	.0	.00	.00	.00	.40	2.8	.51	.2		.07		S.D. (Population)
.0	.0	.0	.00	.00	.00	.44	3.1	.57	.3		.07		S.D. (Sample)

Model, Populn: RMSE .26 Adj (True) S.D. .00 Separation .00 Reliability 1.00
Model, Sample: RMSE .26 Adj (True) S.D. .00 Separation .00 Reliability .80
Model, Fixed (all same) chi-square: .0 d.f.: 4 significance (probability): 1.00

Table 7.3.1 Script Measurement Report (arranged by mN)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Script
12.5	25	.5	1.00	9.04	.64	1.07	.3	2.31	1.3	.80	.71	.76	18 t2_A1 (mark 86.7)
12.5	25	.5	1.00	7.73	.50	1.42	1.6	1.86	1.6	.23	.39	.60	2 t1_A3 (mark 80)
25	50	.5	1.00	7.61	.41	.68	-1.5	.53	-.9	1.34	.79	.71	20 t2_A5B1 (mark 80)
12.5	25	.5	1.00	7.35	.50	.89	-.3	.84	-.3	1.18	.66	.60	19 t2_A4 (mark 80)
12.5	25	.5	1.00	6.44	.55	.67	-1.0	.81	-.2	1.30	.77	.67	4 t1_B2 (mark 76.7)
12.5	25	.5	1.00	5.62	.51	.71	-1.0	.63	-1.0	1.39	.76	.62	21 t2_B4 (mark 73.3)
12.5	25	.5	.99	5.19	.63	1.59	1.4	2.95	1.9	.36	.60	.76	1 t1_A2 (mark 83.3)
25	50	.5	.99	5.10	.40	.94	-.2	1.01	.1	1.04	.70	.69	3 t1_A6B3 (mark 73.3)
12.5	25	.5	.99	4.51	.52	1.07	.3	1.28	.7	.85	.59	.64	6 t1_C2 (mark 70)
25	50	.5	.98	3.88	.36	1.03	.2	1.07	.2	.93	.59	.61	22 t2_B6C3 (mark 70)
25	50	.5	.97	3.49	.37	.99	.0	.81	-.3	1.05	.63	.62	5 t1_B5C1 (mark 70)
12.5	25	.5	.97	3.44	.46	.89	-.5	.87	-.4	1.25	.57	.48	23 t2_C4 (mark 66.7)
12.5	25	.5	.93	2.55	.46	1.01	.1	.91	-.1	1.03	.50	.49	25 t2_D1 (mark 66.7)
12.5	25	.5	.90	2.23	.44	.90	-.6	.81	-.7	1.39	.52	.41	8 t1_D3 (mark 63.3)
25	50	.5	.88	1.97	.33	.84	-1.1	.76	-1.0	1.34	.63	.53	7 t1_C5D2 (mark 63.3)
25	50	.5	.85	1.70	.34	1.11	.7	1.31	1.1	.75	.48	.55	24 t2_C6D4 (mark 63.3)
25	50	.5	.80	1.37	.34	1.16	1.0	1.18	.6	.70	.47	.55	26 t2_D5E1 (mark 60)
12.5	25	.5	.66	.66	.49	1.05	.2	.93	.0	.95	.55	.56	27 t2_E4 (mark 56.7)
25	50	.5	.54	.17	.36	.82	-.9	.69	-1.1	1.26	.72	.62	9 t1_D6E2 (mark 56.7)
12.5	25	.5	.53	.13	.47	.98	.0	.89	-.2	1.07	.55	.53	10 t1_E3 (mark 56.7)
25	50	.5	.22	-1.24	.37	1.19	.9	1.33	1.2	.70	.54	.60	11 t1_E5F1 (mark 50)
25	50	.5	.16	-1.68	.44	1.02	.1	.57	2.5	.98	.71	.71	29 t2_F4G1 (mark 50)
12.5	25	.5	.14	-1.85	.48	1.00	.0	1.00	.0	1.00	.52	.51	12 t1_F2 (mark 50)
25	50	.5	.09	-2.29	.40	.85	-.7	.59	-.9	1.26	.73	.68	28 t2_E6F3 (mark 50)
25	50	.5	.04	-3.14	.49	.99	.0	.59	1.1	1.03	.75	.74	13 t1_F5G2 (mark 46.7)
25	50	.5	.00	-6.92	.96	.74	-.1	.16	2.6	1.16	.86	.85	32 t2_G4H1 (mark 43.3)
12.5	25	.5	.00	-8.56	.83	1.15	.4	.61	.8	.95	.81	.82	31 t2_G3 (mark 43.3)
25	50	.5	.00	-10.69	.73	1.12	.4	.37	1.7	.98	.87	.87	15 t1_G6H2 (mark 40)
12.5	25	.5	.00	-11.38	.83	.86	-.1	.23	4.1	1.21	.82	.81	14 t1_G5 (mark 40)
12.5	25	.5	.00	-16.01	.90	.98	.0	.33	4.6	1.15	.83	.82	16 t1_H3 (mark 36.7)
12.5	25	.5	.00	-16.42	.91	.99	.0	.33	5.7	1.13	.81	.79	17 t1_H4 (mark 36.7)
12.5	25	.5	.00	(-6.67	1.84)	Minimum					.00	.00	30 t2_F6 (mark 46.7)
12.5	25	.5	.00	(-19.70	1.85)	Minimum					.00	.00	33 t2_H5 (mark 36.7)
12.5	25	.5	.00	(-19.70	1.85)	Minimum					.00	.00	34 t2_H6 (mark 33.3)
17.6	35.3	.5	.55	-1.35	.65	.99	.0	.92	.8		.60		Mean (Count: 34)
6.2	12.3	.0	.44	7.92	.41	.19	.7	.58	1.7		.22		S.D. (Population)
6.2	12.5	.0	.44	8.04	.42	.20	.7	.59	1.7		.23		S.D. (Sample)

With extremes, Model, Populn: RMSE .77 Adj (True) S.D. 7.88 Separation 10.30 Reliability .99
With extremes, Model, Sample: RMSE .77 Adj (True) S.D. 8.00 Separation 10.45 Reliability .99
Without extremes, Model, Populn: RMSE .56 Adj (True) S.D. 6.63 Separation 11.85 Reliability .99
Without extremes, Model, Sample: RMSE .56 Adj (True) S.D. 6.74 Separation 12.05 Reliability .99
With extremes, Model, Fixed (all same) chi-square: 3201.1 d.f.: 33 significance (probability): .00
With extremes, Model, Random (normal) chi-square: 32.4 d.f.: 32 significance (probability): .44

Appendix C – FACETS output

Component 21 vs. Component 22

Table 7.1.1 Judge Measurement Report (arranged by mN)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Judge
60	120	.5	.50	.00	.24	.82	-1.5	.58	-.8	1.30	.69	.63	1 CHS
60	120	.5	.50	.00	.24	1.16	1.3	1.24	.6	.72	.58	.63	2 TC
60	120	.5	.50	.00	.24	1.12	1.0	.93	.0	.86	.60	.63	3 PC
60	120	.5	.50	.00	.24	1.02	.2	1.00	.1	.94	.62	.63	4 NB
60	120	.5	.50	.00	.24	.87	-1.1	.61	-.7	1.24	.68	.63	5 GM
60.0	120.0	.5	.50	.00	.24	1.00	.0	.87	-.2		.63		Mean (Count: 5)
.0	.0	.0	.00	.00	.00	.13	1.1	.25	.6		.04		S.D. (Population)
.0	.0	.0	.00	.00	.00	.15	1.3	.28	.6		.05		S.D. (Sample)

Model, Populn: RMSE .24 Adj (True) S.D. .00 Separation .00 Reliability 1.00
Model, Sample: RMSE .24 Adj (True) S.D. .00 Separation .00 Reliability .80
Model, Fixed (all same) chi-square: .0 d.f.: 4 significance (probability): 1.00

Table 7.3.1 Script Measurement Report (arranged by mN)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Script
12.5	25	.5	1.00	6.87	.75	1.01	.2	1.03	.3	.98	.83	.84	18 t2_J1 (mark 70)
25	50	.5	1.00	5.90	.53	1.04	.2	.61	.2	.99	.81	.81	4 t1_K1L3 (mark 57.5)
12.5	25	.5	1.00	5.65	.58	.99	.0	.85	.0	1.01	.72	.72	22 t2_K6 (mark 60)
12.5	25	.5	1.00	5.37	.52	.99	.0	.93	.0	1.03	.65	.63	3 t1_J5 (mark 71.3)
25	50	.5	.99	4.78	.36	.83	-.8	.94	-.1	1.19	.68	.62	2 t1_J3K3 (mark 65)
25	50	.5	.99	4.32	.81	1.10	.3	.36	.0	1.02	.92	.91	6 t1_L1M1 (mark 48.8)
12.5	25	.5	.98	3.91	.48	1.28	1.2	1.68	1.5	.35	.37	.54	1 t1_J2 (mark 67.5)
12.5	25	.5	.97	3.55	.50	.90	-.3	.77	-.3	1.20	.63	.58	19 t2_J4 (mark 67.5)
25	50	.5	.96	3.20	.38	1.04	.2	1.10	.3	.94	.64	.65	20 t2_J6K5 (mark 63.8)
12.5	25	.5	.93	2.59	.57	.70	-1.0	.41	-.9	1.38	.80	.70	5 t1_K4 (mark 60)
25	50	.5	.87	1.90	.45	1.12	.5	1.01	.2	.90	.74	.76	21 t2_K2L4 (mark 57.5)
12.5	25	.5	.73	.97	.58	1.09	.4	.73	.1	.94	.68	.69	7 t1_L5 (mark 52.5)
12.5	25	.5	.62	.50	.58	.95	.0	.60	.1	1.10	.71	.69	23 t2_L2 (mark 53.8)
25	50	.5	.57	.29	.53	.98	.0	.74	-.1	1.03	.83	.82	11 t1_N6O4 (mark 37.5)
12.5	25	.5	.54	.16	.55	1.00	.0	.85	.0	1.01	.66	.66	25 t2_M4 (mark 45)
25	50	.5	.31	-.81	.40	.93	-.3	.61	.2	1.15	.68	.66	24 t2_L6M6 (mark 48.8)
12.5	25	.5	.24	-1.14	.47	.98	-.1	.78	.2	1.12	.51	.50	9 t1_M3 (mark 45)
25	50	.5	.19	-1.42	.33	.90	-.7	.80	.0	1.26	.52	.48	26 t2_M5N5 (mark 40)
25	50	.5	.18	-1.51	.33	1.18	1.3	1.13	.4	.53	.41	.48	8 t1_M2N1 (mark 40)
12.5	25	.5	.17	-1.59	.52	.80	-.7	.63	-.6	1.31	.70	.62	12 t1_O1 (mark 36.3)
12.5	25	.5	.13	-1.87	.51	.91	-.3	.63	-.5	1.25	.66	.61	31 t2_P4 (mark 30)
25	50	.5	.11	-2.09	.34	1.19	1.3	1.53	1.7	.49	.40	.53	28 t2_N4O3 (mark 36.3)
25	50	.5	.10	-2.21	.35	.88	-.8	.64	-1.1	1.30	.65	.58	30 t2_O5P2 (mark 35)
12.5	25	.5	.08	-2.42	.45	.95	-.2	.91	-.3	1.15	.50	.45	16 t1_Q2 (mark 31.3)
25	50	.5	.08	-2.42	.33	.97	-.2	.90	-.3	1.09	.53	.50	15 t1_P3Q4 (mark 32.5)
25	50	.5	.07	-2.59	.32	.99	.0	1.00	.0	1.02	.49	.48	32 t2_P6Q5 (mark 30)
12.5	25	.5	.06	-2.69	.48	.93	-.2	.78	-.4	1.19	.59	.54	27 t2_N3 (mark 38.3)
12.5	25	.5	.05	-2.88	.49	1.05	.2	.92	.0	.95	.56	.58	10 t1_N2 (mark 40)
12.5	25	.5	.04	-3.16	.54	1.45	1.3	1.29	.6	.54	.52	.66	29 t2_O2 (mark 36.3)
12.5	25	.5	.03	-3.48	.43	.86	-.8	.86	-.7	1.44	.53	.39	33 t2_Q1 (mark 27.5)
12.5	25	.5	.03	-3.63	.44	1.05	.3	1.06	.3	.86	.37	.42	34 t2_Q3 (mark 27.5)
12.5	25	.5	.02	-4.13	.48	1.05	.2	1.09	.3	.91	.52	.55	17 t1_Q6 (mark 27.5)
12.5	25	.5	.01	-4.91	.63	.85	-.2	.66	-.4	1.15	.81	.77	14 t1_P1 (mark 35)
25	50	.5	.01	-5.01	.52	1.03	.2	.78	.0	1.01	.82	.82	13 t1_O6P5 (mark 35)
Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Script
17.6	35.3	.5	.44	.00	.49	1.00	.0	.87	.0		.63		Mean (Count: 34)
6.2	12.3	.0	.40	3.39	.11	.14	.6	.27	.6		.14		S.D. (Population)
6.2	12.5	.0	.41	3.44	.11	.14	.6	.28	.6		.14		S.D. (Sample)

Model, Populn: RMSE .50 Adj (True) S.D. 3.35 Separation 6.72 Reliability .98
Model, Sample: RMSE .50 Adj (True) S.D. 3.41 Separation 6.83 Reliability .98
Model, Fixed (all same) chi-square: 1567.3 d.f.: 33 significance (probability): .00
Model, Random (normal) chi-square: 32.3 d.f.: 32 significance (probability): .45

Appendix D – FACETS output

Component 31 vs. Component 32

Table 7.1.1 Judge Measurement Report (arranged by mN)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Judge
60	120	.5	.50	.00	.24	1.14	1.2	1.14	.5	.79	.56	.61	1 CHS
60	120	.5	.50	.00	.24	.88	-1.0	.84	-.4	1.17	.65	.61	2 TC
60	120	.5	.50	.00	.24	1.04	.4	1.00	.0	.93	.59	.61	3 PC
60	120	.5	.50	.00	.24	.86	-1.2	.72	-.9	1.23	.67	.61	4 NB
60	120	.5	.50	.00	.24	1.07	.6	.92	-.1	.93	.60	.61	5 GM
60.0	120.0	.5	.50	.00	.24	1.00	.0	.92	-.2		.61		Mean (Count: 5)
.0	.0	.0	.00	.00	.00	.11	1.0	.14	.5		.04		S.D. (Population)
.0	.0	.0	.00	.00	.00	.12	1.1	.16	.6		.05		S.D. (Sample)

Model, Populn: RMSE .24 Adj (True) S.D. .00 Separation .00 Reliability 1.00
Model, Sample: RMSE .24 Adj (True) S.D. .00 Separation .00 Reliability .80
Model, Fixed (all same) chi-square: .0 d.f.: 4 significance (probability): 1.00

Table 7.3.1 Script Measurement Report (arranged by mN)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Script
12.5	25	.5	1.00	6.67	.77	.96	.1	.45	.0	1.08	.85	.83	2 t1_R2 (mark 70)
12.5	25	.5	.99	5.27	.60	.78	-.5	.45	-.8	1.27	.81	.74	18 t2_R3 (mark 75)
12.5	25	.5	.99	5.03	.59	1.37	1.0	1.71	1.1	.61	.62	.73	1 t1_R1 (mark 62.5)
12.5	25	.5	.98	3.83	.57	.95	.0	.67	-.2	1.10	.71	.69	22 t2_S3 (mark 65)
25	50	.5	.95	2.86	.37	.84	-.8	.68	-.7	1.25	.69	.63	3 t1_R5S5 (mark 62.5)
25	50	.5	.94	2.83	.35	1.10	.6	1.16	.6	.81	.53	.58	4 t1_S4T3 (mark 55)
25	50	.5	.93	2.63	.41	1.03	.2	.76	-.1	1.00	.70	.70	6 t1_T4U6 (mark 50)
25	50	.5	.93	2.52	.40	1.01	.1	.87	.0	1.01	.69	.69	23 t2_T2U3 (mark 55)
25	50	.5	.91	2.25	.37	1.07	.4	.90	.0	.93	.61	.63	19 t2_R4S2 (mark 67.5)
12.5	25	.5	.88	1.95	.61	1.01	.1	1.19	.5	.90	.73	.74	20 t2_R6 (mark 70)
25	50	.5	.77	1.20	.36	.84	-.7	.78	-.8	1.22	.69	.60	21 t2_S1T1 (mark 60)
12.5	25	.5	.62	.50	.52	1.27	1.0	1.51	1.2	.58	.49	.63	24 t2_T6 (mark 57.5)
12.5	25	.5	.57	.30	.54	.90	-.2	.93	.0	1.10	.70	.67	7 t1_T5 (mark 52.5)
12.5	25	.5	.49	-.03	.75	.95	.1	.63	.0	1.06	.84	.82	31 t2_X1 (mark 42.5)
12.5	25	.5	.46	-.16	.76	.98	.1	.68	-.1	1.04	.85	.84	5 t1_S6 (mark 60)
12.5	25	.5	.46	-.17	.49	1.00	.0	1.08	.3	.97	.57	.58	12 t1_W5 (mark 37.5)
25	50	.5	.46	-.17	.38	.87	-.5	.74	-.7	1.18	.71	.65	26 t2_U2V3 (mark 50)
12.5	25	.5	.41	-.37	.56	1.19	.6	1.39	.7	.76	.63	.70	25 t2_U1 (mark 50)
25	50	.5	.37	-.54	.33	1.13	.9	1.22	1.0	.70	.41	.51	11 t1_V6W4 (mark 40)
12.5	25	.5	.33	-.69	.45	.94	-.2	.87	-.4	1.17	.53	.47	30 t2_W3 (mark 42.5)
12.5	25	.5	.33	-.72	.46	.82	-.9	.74	-.9	1.43	.62	.49	10 t1_V5 (mark 40)
25	50	.5	.18	-1.51	.32	1.00	.0	1.00	.0	1.00	.46	.46	27 t2_V1W1 (mark 47.5)
12.5	25	.5	.15	-1.73	.59	.68	-1.0	.41	-.3	1.40	.78	.71	9 t1_U5 (mark 45)
25	50	.5	.14	-1.80	.35	1.01	.1	.99	.0	.99	.57	.57	14 t1_X4Y4 (mark 32.5)
25	50	.5	.11	-2.10	.34	1.04	.2	.97	.0	.96	.53	.54	13 t1_W6X6 (mark 35)
25	50	.5	.11	-2.13	.39	1.12	.6	.95	.1	.86	.61	.64	8 t1_U4V4 (mark 45)
25	50	.5	.07	-2.56	.35	1.04	.3	1.08	.3	.93	.58	.60	29 t2_W2X3 (mark 37.5)
25	50	.5	.06	-2.81	.32	.99	.0	.93	-.2	1.05	.47	.46	32 t2_X2Y2 (mark 37.5)
12.5	25	.5	.06	-2.84	.57	.95	.0	.87	-.1	1.06	.73	.70	28 t2_V2 (mark 45)
12.5	25	.5	.05	-2.85	.42	.86	-1.1	.84	-1.1	1.75	.52	.32	17 t1_Y6 (mark 30)
12.5	25	.5	.05	-3.00	.42	1.11	.9	1.14	.9	.37	.15	.31	16 t1_Y3 (mark 27.5)
12.5	25	.5	.04	-3.14	.42	.95	-.3	.95	-.3	1.25	.39	.33	34 t2_Y5 (mark 37.5)
12.5	25	.5	.02	-3.93	.47	1.04	.2	1.11	.4	.91	.50	.53	33 t2_Y1 (mark 30)
12.5	25	.5	.01	-4.60	.75	.95	.1	.63	.0	1.06	.84	.82	15 t1_X5 (mark 35)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Script
17.6	35.3	.5	.47	.00	.48	.99	.0	.92	.0		.62		Mean (Count: 34)
6.2	12.3	.0	.36	2.75	.13	.13	.6	.28	.6		.15		S.D. (Population)
6.2	12.5	.0	.37	2.79	.14	.14	.6	.29	.6		.15		S.D. (Sample)

Model, Populn: RMSE .50 Adj (True) S.D. 2.71 Separation 5.42 Reliability .97
Model, Sample: RMSE .50 Adj (True) S.D. 2.75 Separation 5.50 Reliability .97
Model, Fixed (all same) chi-square: 1068.2 d.f.: 33 significance (probability): .00
Model, Random (normal) chi-square: 31.9 d.f.: 32 significance (probability): .47

Statistical Reports

Tim Gill Research Division

The on-going '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. These reports, produced using national-level examination data, are available on the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports.

The most recent additions to this series are:

- Statistics Report Series No.69: Progression from GCSE to AS and A level, 2012

Additionally the following reports have been revised, to more accurately reflect the true levels of uptake and provision of GCSEs and A levels in England:

- Statistics Report Series No.34 (revised): Provision of GCSE subjects 2010
- Statistics Report Series No.43 (revised): Provision of GCSE subjects 2011
- Statistics Report Series No.55 (revised): Uptake of GCE A level subjects 2012
- Statistics Report Series No.56 (revised): Provision of GCSE subjects 2012

Research News

Jessica Munro Research Division

Society for Research into Higher Education (SRHE)

The 2013 SRHE conference was held in Newport, Wales in December. The conference explored global trends and transformations in Higher Education. Frances Wilson presented a paper entitled Aspiring to bridge the gap between A-level and HE: A study of assessments and additional support lessons.

Annual Meeting of the National Council on Measurement in Education (NCME)

The 2014 NCME Annual Meeting took place in Philadelphia, United States from 2-6 April. Colleagues from the Research Division, CIE and the Institute of Education, University of London presented the following papers: Tom Bramley, Anthony Dawson and Paul Newton: *On the limits of linking: experiences from England*. Paul Newton and Stuart Shaw: *Do We Need to Use the Term 'Validity'?*

British Congress of Mathematics Education (BCME)

The eighth BCME conference was held in Nottingham in June. OCR was a headline sponsor for the event, which focussed on enabling greater collaboration between researchers and classroom teachers. Frances Wilson presented a paper on Research informing the new maths GCSE: The development of teaching resources in times of curriculum change.

The International Association for Educational Assessment (IAEA)

The IAEA annual conference allows researchers and assessment professionals from around the world to share their expertise and exchange ideas. The 40th annual conference was held in Singapore in May, and explored the theme of Assessment Innovations for the 21st Century. Simon Lebus, Group Chief Executive, and Michael O'Sullivan, Chief Executive at CIE, attended alongside colleagues from the Research Division, Cambridge English and CIE. The following papers were presented by colleagues from across Cambridge Assessment: Tim Oates: *Textbooks count: The relationship between textbooks, assessment and the curriculum*. Sylvia Green: *Models of internal, school based assessment: challenges and possibilities*. Tom Benton: *Comparing the reliability of standard maintaining via examiner judgement to statistical approaches*. Helen Eccles: *The Cambridge Approach to 21st Century skills: definitions, development, and dilemmas for assessment.* Phineas Hodson: *Practical validation: organisational approaches to large-scale evaluation and continuous improvement.* Isabel Nisbet: *What is meant by 'rigour' in examinations?* Isabel Nisbet and Paul Newton: *Validity – an approach for the 21st century and what this might mean for national assessment systems across the world.* Nick Saville: *Learning Oriented Assessment - a systemic view of assessment within educational context.* Nick Saville: *Investigating the impact of language tests in their educational context.*

Publications

The following articles and books have been published since Issue 17 of *Research Matters*:

Benton, T. (2014). Using meta-regression to explore moderating effects in surveys of international achievement. *Practical Assessment, Research & Evaluation*, *19*(3). Retrieved from: http://pareonline.net/pdf/v19n3.pdf

Child, S.F.J., Theakston, A., & Pika, S. (in press). How do modelled gestures influence preschool children's spontaneous gesture production? Social versus semantic influence. *Gesture*, *14*(1).

Crisp, V. & Green, S. (2013). Teacher views on the effects of the change from coursework to controlled assessment in GCSEs. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19*(8), 680–699.doi: 10.1080/13803611.2013.840244

Newton, P., & Shaw, S. (2014). *Validity in Educational and Psychological Assessment*. London: SAGE.

Cambridge Assessment
CONFERENCE 2014

CAMBRIDGE ASSESSMENT

INTERNATIONAL EDUCATION

interpretation
importance
impact

There is no doubt that internationally-focused education is rising up the agenda of governments worldwide. But what exactly do we mean by an international education? How best can we prepare students for an increasingly interconnected world?

The seventh Cambridge Assessment Conference will welcome over 140 education experts from across the UK and overseas to scrutinise the challenges and opportunities that education without borders creates. A must-attend event for professionals involved in the shaping and delivery of international education at school and policy levels.

CONFIRMED SPEAKERS

Isabel Nisbet Executive Director, A Level Content Advisory Board
Jeremy Hodgen Professor of Mathematics Education, King's College London
Sunny Varkey Founder and Executive Chairman, GEMS Education Group
David Graddol Director, The English Company
Dr Stephen Spurr Headmaster, Westminster School
David Barrs and **Jill Martin** Headteachers, Anglo European School
Marc Tucker President and CEO, National Center on Education and the Economy, USA
Dr Karin Zimmer Researcher, German Institute for International Educational Research

15 October 2014 | Downing College | Cambridge

BOOK YOUR PLACE: www.cambridgeassessment.org.uk/conference2014

CONTENTS : Issue 18 Summer 2014

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

Tel: 01223 552666
Fax: 01223 552700
Email: ResearchProgrammes@cambridgeassessment.org.uk
http://www.cambridgeassessment.org.uk

ISSN: 1755–6031