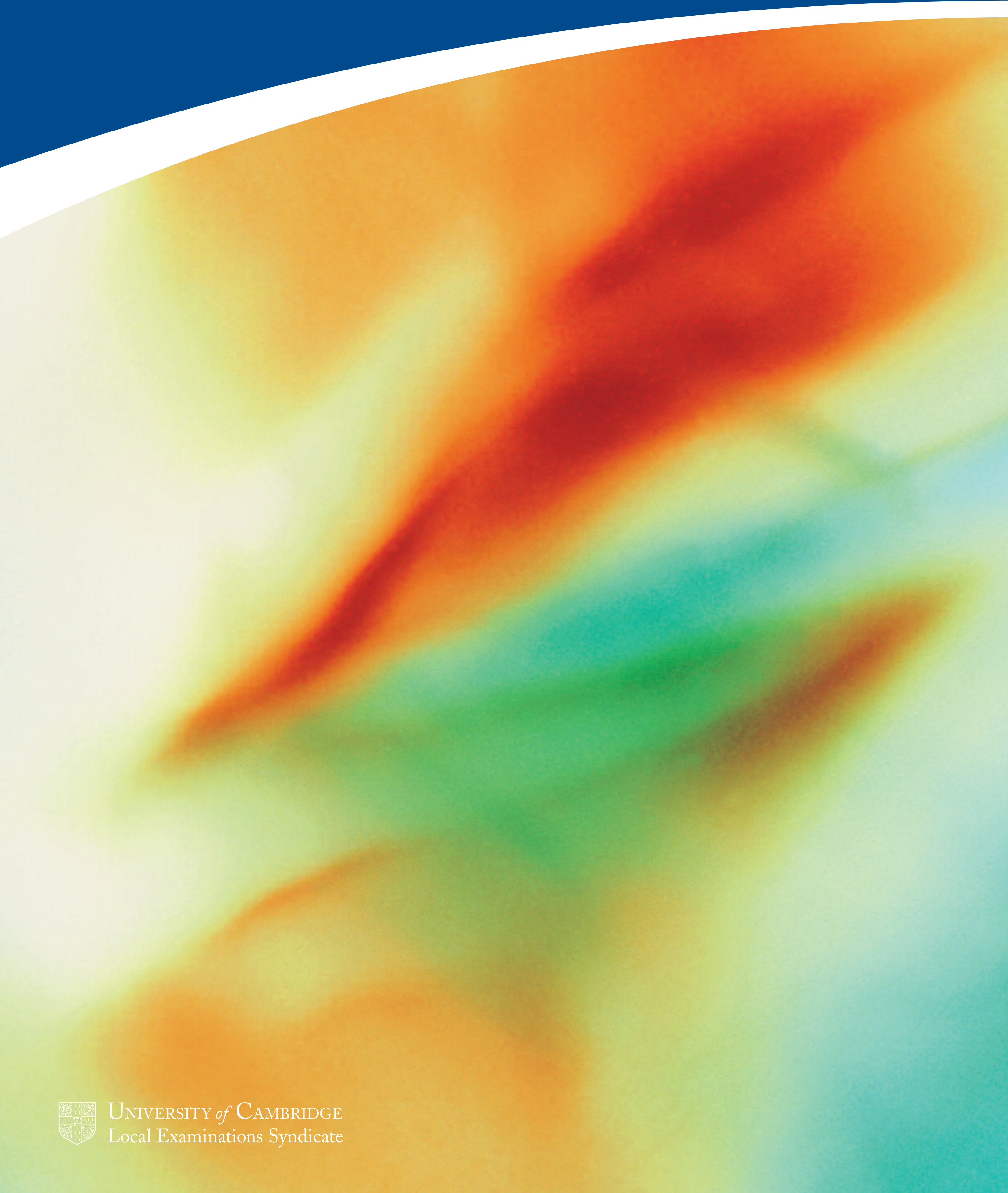
Issue 22 Summer 2016

Research Matters

CAMBRIDGE ASSESSMENT



UNIVERSITY *of* CAMBRIDGE
Local Examinations Syndicate

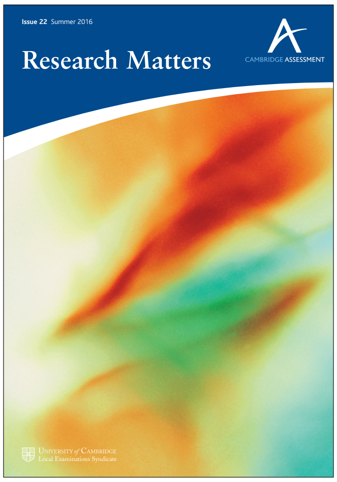
CAMBRIDGE ASSESSMENT

Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.

Citation

Dunn, K., Darlington, E., & Benton, T. 2016). Revisiting the topics taught as part of an OCR History qualification. *Research Matters: A Cambridge Assessment publication, 22*, 2–8.

All details are correct at the time of going to print in August 2016



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

The full issue of *Research Matters 22* and all previous issues are available from our website: www.cambridgeassessment.org.uk/our-research

Research Matters : 22

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

The research which underpins and consolidates the quality and operation of our assessments remains fundamental to our programme of work, but we continue to combine this with contextual analysis regarding learning, participation and policy. These matters have high prominence in this edition of *Research Matters*, with two articles of particular significance, despite relatively modest titles. Elliot's article reporting her incisive work on identifying the highest performing systems brings some much-needed sophistication to the contemporary research scene associated with high-performing jurisdictions. Following the Programme for International Student Assessment (PISA) 2000 the 'done thing' was to buy a ticket to Finland. It seemed obvious. In fact, doing this was steeped in naiveté, as my own *Finnish Fairy Stories* (Oates, 2015) and Salhgren's excellent *Real Finnish Lessons* (Centre for Policy Studies, 2015) now have made abundantly clear. There is much to learn from Finland, if the right questions are asked, and the right 'guided enquiry' undertaken. Elliott's scrutiny of indicators does not undermine sound transnational comparative analysis; it supports sophisticated understanding of the assets and deficits of specific sets of national education arrangements. It does not ignore complexity, but indicates how an appreciation of the complexity of system performance can further the insights gained from comparative analysis. Meanwhile, back in our domestic system, the issue of 'all subjects being at the same level of difficulty' has emerged again in policy discussions, not least because of equivalence assumed in, and driven by, accountability arrangements. Work by Professor Coe, Director of the Centre for Evaluation and Monitoring, has examined the relative movement of standards in different subjects and we, at Cambridge Assessment, have argued that the pursuit of 'same difficulty' is not quite the obvious good which it might seem. Benton's analytic piece explores one dimension of this pursuit and the implications which it carries. As with all 'obvious' policy moves, this is one which needs thorough examination before rushing to immediate action.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

The first two articles in this issue report on research that investigates subject content and skills in very different contexts. The research from Dunn, Darlington and Benton follows from the work of Child, Darlington and Gill (2014, 2015). This more recent research focuses on the topics that schools intended to teach as part of their A level History qualification (first teaching from September 2015). It provides interesting insights into the potential impact of reform of the A level History specification. Also in the context of reform, Darlington and Bowyer consider the impact of changes to A level Mathematics and A level Further Mathematics as a result of the reform programme. They discuss the changes to optionality and preparedness of A level students who proceed to study Business Studies at university.

Child and Shaw discuss process and outcomes in the context of a key 21st century skill, namely, collaboration. They recognise the importance of construct definition and the challenges related to validity, reliability, comparability, and delivery in assessment. This article informs debate on issues of construct definition, and task design, as well as the challenge of group assessment.

The next two articles consider the matter of subject difficulty from different perspectives. Bramley reports on his research into the thorny issue of whether and how it could be possible to control for inter-subject comparability. He used simulated data to investigate the validity of one statistical method and highlights the problem of trying to adjust for differences in difficulty at subject level. In his article, Benton argues that differences in subject difficulty do not cause problems for school accountability, or for summarising the achievement of students at GCSE. He used data from the National Pupil Database to support his conclusion and expresses concerns about the rationale for attempting to make different subjects 'equally difficult'.

The final two articles have an international flavour. Williamson explores the challenging area of statistical moderation of school-based assessment. She outlines methods of statistical moderation that are used in jurisdictions around the world and applies them to GCSE results data. Her work illustrates that further research is needed in order to reassure stakeholders before such changes to moderation processes could be considered. In the final article, Elliott identifies a challenge for those studying international comparability. There are many different comparisons that can be used to identify high-performing jurisdictions and this makes it increasingly difficult to identify a smaller number of them since their numbers grow with the number of comparative ranking exercises that are carried out. Elliott describes a definition that has been used in research at Cambridge Assessment to identify the highest performing jurisdictions. Although she identifies limitations to her approach, it provides an interesting and pragmatic definition to those wanting to identify a small number of the highest performers.

Sylvia Green *Director, Research Division*

Revisiting the topics taught as part of an OCR A level History qualification

Karen Dunn British Council, **Ellie Darlington** and **Tom Benton** Research Division (The study was completed when the first named author was based in the Research Division)

Introduction

Given the introduction of a broader range of options in the Oxford, Cambridge and RSA (OCR) new General Certificate of Education (GCE) Advanced level (A level) History, this article follows on from a previous analysis of A level History options based on the previous specification for OCR History (Specification A) (Child, Darlington, & Gill, 2014, 2015). That research relied on OCR History centres responding to requests for participation in an online survey. However, OCR's introduction of an online 'specification creator' tool for centres has provided quantitative information about the topics which schools intend to teach their students as part of their A level.

OCR introduced a redeveloped specification, History H505, for first teaching in September 2015. It aimed to provide more 'stretch and challenge' for students, and requires that students study topics which cover a chronological range of at least 200 years. It comprises three units: *British History* (13 possible topics), *Non-British History* (24 topics), and *Historical Themes* (21 topics).

Teachers have the option to select a combination of topics across the three units rather than selecting a particular 'route' through the course based on time period (Medieval/Early Modern/Modern) as per past specifications. There are two provisos to the new approach:

1. They must meet the Government's requirements for 200-year minimum coverage (Department for Education [DfE], 2014);
2. They do not include prohibited unit combinations.

Of the 6,552 possible combinations of topics under the 3 units, 338 are prohibited owing to non-compliance with the 200-year rule, and an additional 144 combinations are prohibited because of an overlap in content. This leaves 6,070 permitted topic combinations across the 3 units – a vast range of options.

Aims

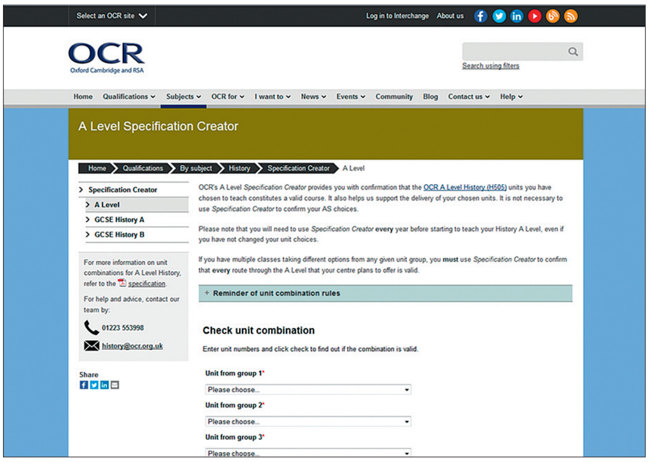
As with the previous study, we sought to establish what the common topic choices and combinations are.

It was intended that the analysis could enable some comparisons to be made with the previous findings (Child et al., 2014). However, since the structure of the options available to schools has changed considerably, it is not possible to make direct links. It must also be noted that the mode of collecting the information is quite different: in the previous study the teachers were canvassed about their choices post hoc, whilst this study used data collected before the teaching of the course had begun. There was also a much stronger motivation for teachers to provide the information analysed in this study, since the internet tool used to collect the information was also informing them about the viability of their topic choices.

Method

The A level Specification Creator tool

Information regarding A level History options was collected from 438 schools using OCR's Specification Creator[1]. Schools considering teaching an OCR History A level course are recommended to use this tool in order to ensure that their choice of options from the three groups of A level History units fits the minimum 200-year requirement set out by the DfE (2014).



A Level Specification Creator

Home > Qualifications > By subject > History > Specification Creator > A Level

- Specification Creator
  - A Level
  - GCSE History A
  - GCSE History B

For more information on unit combinations for A Level History, refer to the specification.

For help and advice, contact our team by:

☎ 01223 553998

✉ history@ocr.org.uk

Share

OCR's A Level *Specification Creator* provides you with confirmation that the OCR A Level History (H505) units you have chosen to teach constitutes a valid course. It also helps us support the delivery of your chosen units. It is not necessary to use *Specification Creator* to confirm your AS choices.

Please note that you will need to use *Specification Creator* **every** year before starting to teach your History A Level, even if you have not changed your unit choices.

If you have multiple classes taking different options from any given unit group, you **must** use *Specification Creator* to confirm that **every** route through the A Level that your centre plans to offer is valid.

+ Reminder of unit combination rules

Check unit combination

Enter unit numbers and click check to find out if the combination is valid.

Unit from group 1*

Please choose...

Unit from group 2*

Please choose...

Unit from group 3*

Please choose...

Figure 1: OCR's A level Specification Creator tool

It is useful to highlight a number of caveats regarding the data that are used for this study. There was nothing preventing a number of different teachers from the same school entering a range of unit choices for their students. For example, some individual classes could study certain topic combinations, and other classes different ones depending on individual teacher specialisms. Additionally, it is possible that teachers would enter a number of possible options into the specification creator, just to check their viability, but only in reality be offering a single route through the course to their students. Teachers may indeed even choose a different awarding body altogether before their students take their examinations.

There will be no definitive information about the number of centres that follow OCR's A level History course for 2017 until the candidates are entered for the final examinations. However, it is assumed unlikely that the teachers will have had motivation to enter misleading information

[1] http://www.ocr.org.uk/qualifications/by-subject/history/specification-creator/a-level-specification-creator/

into the Specification Creator; therefore the information collected in this manner is believed to be a reasonably accurate reflection of their intentions.

In order to tackle such issues and to ensure that the data analysed represented a situation as close to reality as possible, some data cleaning took place. Duplicate entries from the same schools were removed. Sometimes one user accidentally submitted the same options twice. On other occasions, multiple teachers from the same school had used the Specification Creator and entered the same information as each other. In instances when multiple submissions were made from one school, the school was contacted[2] to ascertain whether all of the options specified in the submissions were being pursued.

Information from the Specification Creator tool was extracted in October 2015, once schools had begun teaching to the specification.

National databases

Following the extraction and cleaning of the information from the Specification Creator database, centre number information was used to gain insight into the centres represented in the dataset. Summary information about the schools (e.g., school sector, number of A level candidates) was retrieved from the National Centre Number database, and information on school attainment was calculated using National Pupil Database (NPD) information.

Data collected

School sector: Information from 438 schools was collected. The majority (69.6 per cent) were state schools, 24.9 per cent were from the independent sector, and the remaining schools categorised either as 'other' (e.g., hospital schools) or the relevant data was missing.

School attainment: Mean A level scores in 2015 across all subjects and all awarding bodies were calculated by assigning a number to each A level grade (A*=6, A=5, B=4, etc.) and taking the mean of all A levels taken by all of the students at the school.

Following the methodology employed by Child et al. (2014), participating centres were then divided into 'High', 'Medium' and 'Low' groups based upon the relative attainment of the centres included in this dataset. That is, it did not link to any external measure/benchmark of attainment. The number of centres in each category and some descriptive information about the attainment are given in Table 1.

Table 1: A level score distributions within each attainment category[3]

Attainment	*A level scores*				*No. of participating centres (%)*
	Minimum	*Maximum*	*Mean*	*SD*	
Low	1.67	3.22	2.9130	.25959	134 (32.9)
Medium	3.23	3.68	3.4315	.12334	136 (33.4)
High	3.69	5.27	4.1460	.36921	137 (33.7)

As an indication of how to interpret this data, in 2015 the average A level grade awarded was equivalent to a C+ (DfE, 2015). The data in Table 1 approximates to a 'low-attaining' centre as having an average result of a D+, a 'medium' centre of a C and a 'high-attaining' centre a B- in A levels.

Cohort size: Schools were asked to give an estimate of the cohort size for each of the routes through the course they were teaching. Table 2 shows that the majority of participants reported that their cohorts would be at most 20 students large, with a significant minority reporting cohorts of between 21 and 40 students.

Table 2: Cohort estimates

Cohort estimate	*No. of schools (%)*
0–20	272 (53.8)
21–40	151 (29.8)
41–60	49 (9.7)
61–80	15 (3.0)
81–99	12 (2.4)
100+	7 (1.4)
Total	**506 (100.0)**

Multiple routes: Fifty-six schools indicated that they would be offering several routes through the course. That is, different classes within the same school may have been studying different topics from each other. Of these, nine schools were offering more than two routes (with a maximum of five routes offered by two schools).

Some cautions

Naturally the information about each of the schools summarised above does not represent entirely independent variables. Schools with a greater number than 1,000 total A level entries across all subjects are exclusively within the state sector. The majority of schools in the low and medium attainment groups are from the state sector; with independent schools dominating those represented in the high attainment group (see Table 3).

Table 3: Attainment categories by school sector

Attainment category	*School type*				*Total*
	Independent	*State*	*Other*	*Missing*	
Low	7	119	5	3	134
Medium	13	119	0	4	136
High	83	52	1	1	137
Total	**103**	**290**	**6**	**8**	**407**

Additionally, owing to the fact that there are a large number of different topic choices and therefore combinations, a number of the possible topics and combinations will not be offered by any OCR centres this year. In fact, it would be impossible for all combinations to be offered as there are 6,070 permitted combinations but far fewer schools which teach A level History.

[2] This was only done if the user had indicated on the Specification Creator that they were willing to be contacted.

[3] The frequency data excludes that for 31 centres where attainment categories could not be categorised because the relevant data was not available.

Table 4: Most popular topics for Unit 1: *British History*

Overall rank	*Unit*	*Topic*	*Period*			*No. of schools offering topic (%)*
			Medieval	*Early Modern*	*Modern*	
4	Y106	England 1485–1558: The Early Tudors		●		98 (19.4)
5	Y113	Britain 1930–1997			●	96 (19.0)
8	Y107	England 1547–1603: The Later Tudors		●		71 (14.0)
9	Y108	The Early Stuarts and the Origins of the Civil War 1603–1660		●		67 (13.2)
10	Y105	England 1445–1509: Lancastrians, Yorkists and Henry VII	●			47 (9.3)
11=	Y112	Britain 1900–1951			●	44 (8.7)
19=	Y102	Anglo-Saxon England and the Norman Conquest 1035–1107	●			27 (5.3)
21=	Y110	From Pitt to Peel: Britain 1783–1853			●	26 (5.1)
30=	Y109*	The Making of Georgian Britain 1678–c.1760		●		9 (1.8)
32=	Y111	Liberals, Conservatives and the Rise of Labour 1846–1918			●	8 (1.6)
40=	Y101*	Alfred and the Making of England 871–1016	●			5 (1.0)
40=	Y103*	England 1199–1272	●			5 (1.0)
45=	Y104*	England 1377–1455	●			3 (0.6)
					Total	**506 (100.0)**

Table 5: Most popular topics for Unit 2: *Non-British History*

Overall rank	*Unit*	*Topic*	*Period*			*No. of schools offering topic (%)*
			Medieval	*Early Modern*	*Modern*	
3	Y221	Democracy and Dictatorships in Germany 1919–1963			●	108 (21.3)
6	Y219	Russia 1894–1941			●	74 (14.6)
7=	Y213	The French Revolution and the rule of Napoleon 1774–1815			●	72 (14.2)
11=	Y223	The Cold War in Europe 1941–1995			●	44 (8.7)
15	Y212*	The American Revolution 1740–1796		●		31 (6.1)
16	Y203	The Crusades and the Crusader States 1095–1192	●			30 (5.9)
17=	Y222	The Cold War in Asia 1945–1993			●	28 (5.5)
19=	Y216	The USA in the 19th Century: Westward expansion and Civil War 1803–c.1890			●	27 (5.3)
24	Y218	International Relations 1890–1941			●	16 (3.2)
26=	Y207	The German Reformation and the rule of Charles V 1500–1559		●		12 (2.4)
26=	Y215	Italy and Unification 1789–1896			●	12 (2.4)
26=	Y220	Italy 1896–1943			●	12 (2.4)
29	Y206	Spain 1469–1556		●		10 (2.0)
37=	Y204*	Genghis Khan and the Explosion from the Steppes c.1167–1405	●			6 (1.2)
37=	Y224*	Apartheid and Reconciliation: South African Politics 1948–1999			●	6 (1.2)
40=	Y210*	Russia 1645–1741		●		5 (1.0)
44	Y208	Philip II 1556–1598		●		4 (0.8)
45=	Y209*	African Kingdoms c.1400–c.1800: four case studies		●		3 (0.6)
50=	Y205	Exploration, Encounters and Empire 1445–1570		●		2 (0.4)
50=	Y217*	Japan 1853–1937			●	2 (0.4)
50=	Y211*	The Rise and Decline of the Mughal Empire in India 1526–1739		●		1 (0.2)
50=	Y214	France 1814–1870			●	1 (0.2)
55=	Y201*	The Rise of Islam c.550–750	●			0 (0.0)
55=	Y202*	Charlemagne 768–814	●			0 (0.0)
					Total	**506 (100.0)**

Table 6: Most popular topics for Unit 3: *Historical Themes*

Overall rank	Unit	Topic	Period			No. of schools offering topic (%)
			Medieval	Early Modern	Modern	
1	Y319	Civil Rights in the USA 1865–1992			●	149 (29.4)
2	Y318	Russia and its Rulers 1855–1964			●	133 (26.3)
13=	Y306	Rebellion and Disorder under the Tudors 1485–1603		●		38 (7.5)
13=	Y312*	Popular Culture and the Witchcraze of the 16th and 17th Centuries		●		38 (7.5)
17=	Y315	The Changing Nature of Warfare 1792–1945			●	28 (5.5)
21=	Y316	Britain and Ireland 1791–1921			●	26 (5.1)
23	Y314	The Challenge of German Nationalism 1789–1919			●	21 (4.2)
25	Y321*	The Middle East 1908–2011: Ottomans to Arab Spring			●	14 (2.8)
30=	Y317*	China and its Rulers 1839–1989			●	9 (1.8)
32=	Y311*	The Origins and Growth of the British Empire 1558–1783		●		8 (1.6)
34=	Y305	The Renaissance c.1400–c.1600	●			7 (1.4)
34=	Y307	Tudor Foreign Policy 1485–1603		●		7 (1.4)
34=	Y320*	From Colonialism to Independence: The British Empire 1857–1965			●	7 (1.4)
37=	Y302*	The Viking Age c.790–1066	●			6 (1.2)
40=	Y313	The Ascendancy of France 1610–1715		●		5 (1.0)
45=	Y303	English Government and the Church 1066–1216	●			3 (0.6)
45=	Y308	The Catholic Reformation 1492–1610		●		3 (0.6)
45=	Y310	The Development of the Nation State: France 1498–1610		●		3 (0.6)
50=	Y304*	The Church and Medieval Heresy c.1100–1437	●			1 (0.2)
55=	Y301*	The Early Anglo Saxons c.400–800	●			0 (0.0)
					Total	**506 (100.0)**

What are the most common topic choices and combinations?

Of interest are the individual topic combinations in each unit, as well as the topics that are commonly combined with one another.

The data reported in this section encompass only the responses from schools identified as state or independent schools. Data for other types of centre (e.g., A level resit colleges) or for centres for which this data was missing were excluded both in case they skewed the results, and because the previous study only analysed data from these two types of school.

Most popular topics for each unit

Tables 4–6 show the most popular topics for each unit. There is a predominance of Modern and Early Modern topics at the top of the popularity listing across all three units. None of the Medieval topics attracted more than 10% of schools.

The top choices are also heavily weighted towards topics that are the same as or extended from OCR's previous A level History specification. However, notable exceptions are *Y212: The American Revolution 1740–1796*, attracting 6.1% of the Unit 2 choices, and *Y312: Popular Culture and the Witchcraze of the 16th and 17th Centuries*, attracting 7.5% of the Unit 3 choices. In all, 29.1% of schools were intending to teach at least one topic new to the specification. This was the case for both independent schools and state schools, and across low, medium and high-attaining schools.

Topics which are new to this specification are indicated with an asterisk (*) next to the unit code in Tables 4–6 and throughout the remainder of this article.

There were some differences in the choices of certain Unit 2 topics between schools which were classified as high-, medium- and low-attaining (some of these differences are visible in Figure 2). For example, low-attaining schools were much more likely to have offered *Y221: Democracy and Dictatorships in Germany 1919–1963* ($\chi^2(2)=12.622$, $p=.002$) and *Y218: International Relations 1890–1941* ($\chi^2(2)=7.935$, $p=.019$) to their students than medium- or high-attaining schools. Similarly, *Y319: Civil Rights in the USA 1865–1993* was selected by significantly fewer higher-attaining schools than medium- and low-attaining schools ($\chi^2(2)=6.372$, $p=.041$).

Conversely, high-attaining schools were significantly more likely to offer *Y213: The French Revolution and the Rule of Napoleon 1774–1815* to their students than medium- or low-attaining schools ($\chi^2(2)=7.884$, $p=.019$). Similar can be said for *Y207: The German Reformation and the rule of Charles V 1500–1559*.

There were some differences in the choices of state and independent schools. Specifically, independent schools were significantly more likely to teach *Y203: The Crusades and the Crusader States 1095–1192* ($\chi^2(1)=4.572$, $p=.032$). Conversely, state schools were significantly more likely to offer *Y219: Russia 1894–1941* ($\chi^2(1)=4.100$, $p=.043$) as shown in Figure 3.

Many of the most popular Unit 1, 2 and 3 choices were also popular in the old specification. For example, the following topics from the current specification featured in both the top 15 topic choices identified by Child et al. (2014) based on the old specification, and the top 15 from Tables 4, 5 and 6:

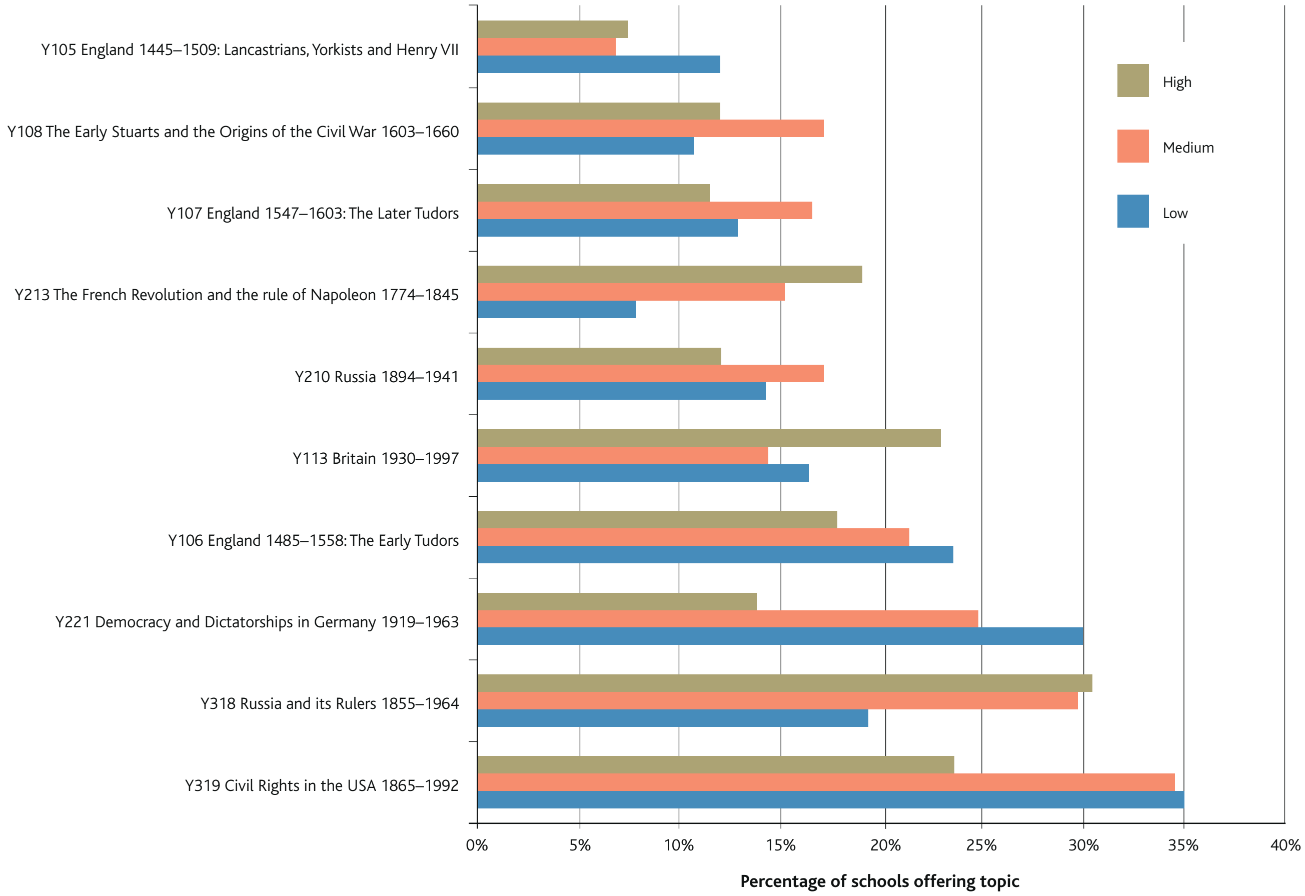


Figure 2: The 10 most popular topics by school attainment level

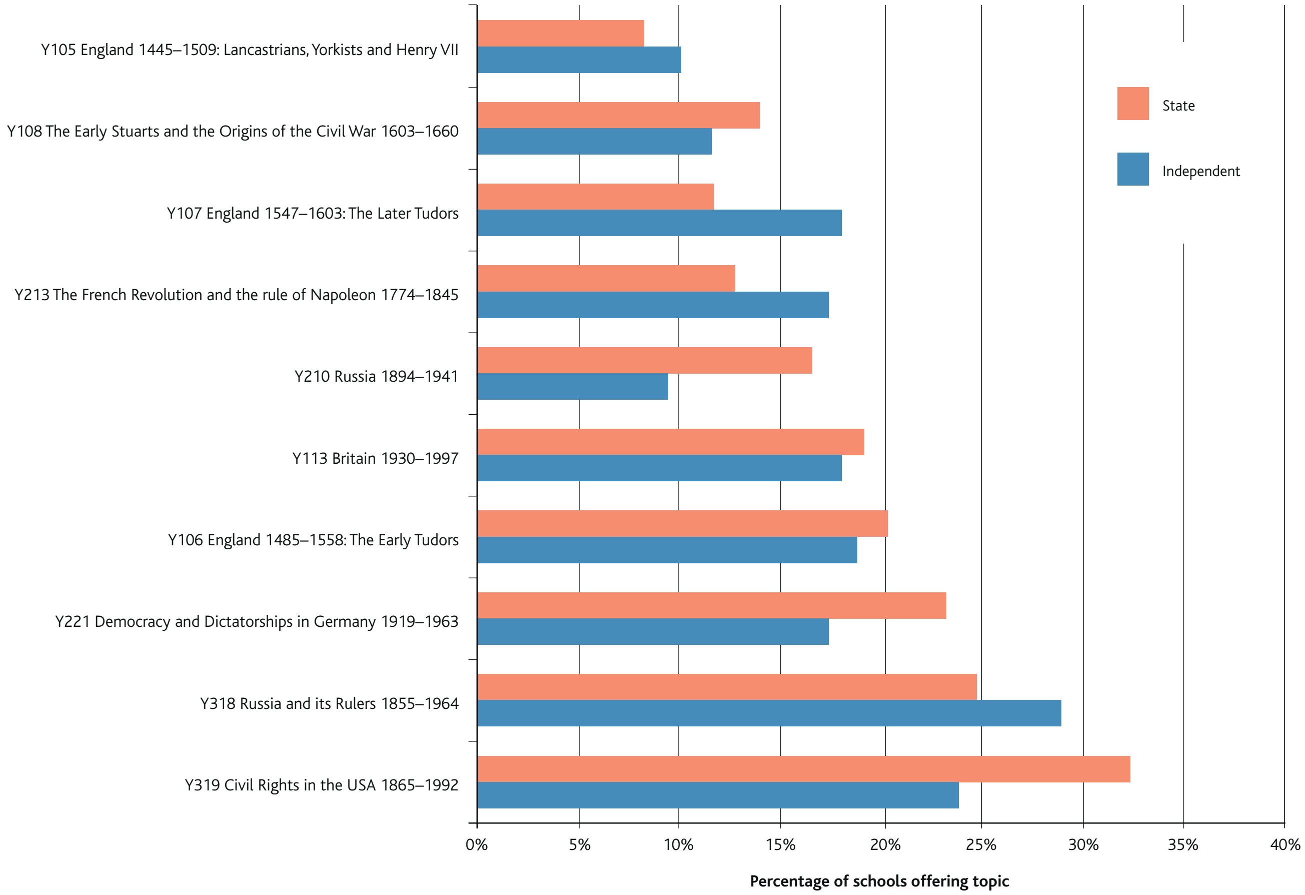


Figure 3: The 10 most popular topics by school type

- Civil Rights in the USA 1865–1992
- Russia and its Rulers 1855–1964[4]
- Democracy and Dictatorships in Germany 1919–1963
- England 1485–1558: The Early Tudors and/or England 1547–1603: The Later Tudors[5]
- Russia 1894–1941[6]
- The French Revolution and the rule of Napoleon 1774–1815[7]

None of the significant differences between school sector or attainment level matched the differences identified in the previous study. This may largely be due to the fact that the topics currently available for study have undergone considerable restructuring, and the methodology employed in data collection is very different between that reported here and that used by Child et al. (2014). The 2014 study examined the topic choices of 90 schools; however, the Specification Creator has enabled the collection of data from 438 schools for the present study, meaning that the findings are more representative of all schools offering OCR's A level History qualification.

Most popular topic combinations

There are a huge number of possible combinations of units (see Figure 4), which mean that even the most popular 'three-way' combination reported in this study only represented 3 percent of all routes.
This means that with regard to three-way combinations, schools are not herding towards a single route through the course and are taking advantage of the wide-ranging options available in combining topics to meet their individual needs. The five most popular combinations are given in Table 7. Note that all of the topics are from Modern periods, with the exception of Y106 and Y107.

[4] The old specification's topic was Russian Dictatorship 1855–1992.

[5] The old specification's topic was Rebellion and Disorder under the Tudors 1485–1603.

[6] The old specification's topic was From Autocracy to Communism: Russia 1894–1941.

[7] The old specification's topic was The Origins and Course of the French Revolution 1774–1795.

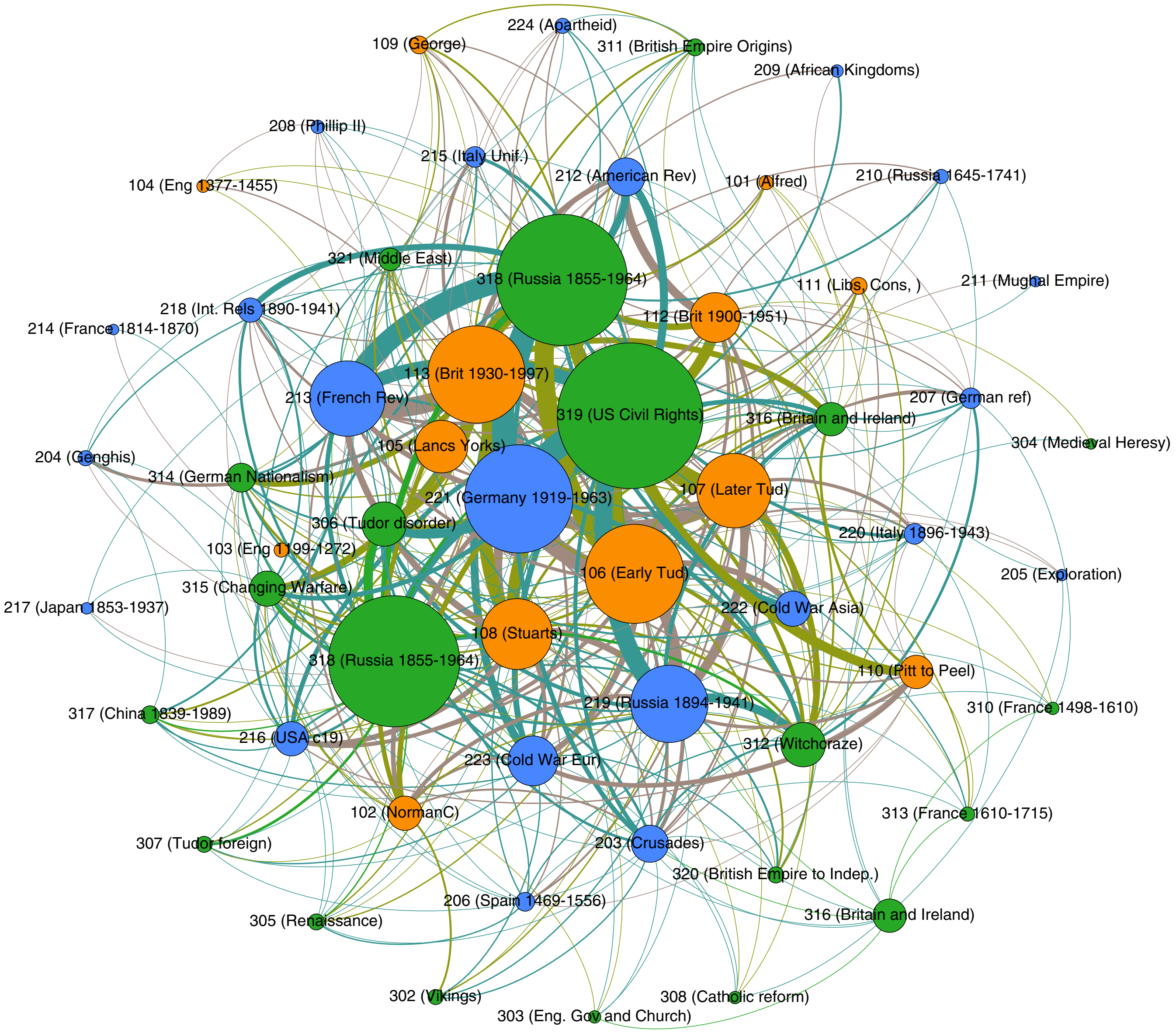


Figure 4: Network diagram showing the most popular topics and topic combinations

Table 7: Top 5 most popular three-way unit combinations

Rank	*Unit 1 choice*		*Unit 2 choice*		*Unit 3 choice*		*No. of schools (%)*
1	Y113	Britain 1930–1997	Y213	The French Revolution and the rule of Napoleon 1774–1815	Y318	Russia and its Rulers 1855–1964	15 (3.0)
2=	Y106	England 1485–1558: The Early Tudors	Y221	Democracy and Dictatorships in Germany 1919–1963	Y319	Civil Rights in the USA 1865–1992	12 (2.4)
2=	Y106	England 1485–1558: The Early Tudors	Y221	Democracy and Dictatorships in Germany 1919–1963	Y318	Russia and its Rulers 1855–1964	12 (2.4)
4	Y106	England 1485–1558: The Early Tudors	Y219	Russia 1894–1941	Y319	Civil Rights in the USA 1865–1992	9 (1.8)
5	Y107	England 1547–1603: The Later Tudors	Y221	Democracy and Dictatorships in Germany 1919–1963	Y318	Russia and its Rulers 1855–1964	8 (1.6)

Figure 4 highlights the vast number of different topic combinations that are possible through this A level History course. The colour coding shows Unit 1 topics in orange, Unit 2 topics in blue and Unit 3 topics in green. The node size is relative to the number of schools which reported to be offering that topic, with the thickness of the interconnecting lines indicating the popularity of that particular combination of topics.

Limitations

The key limitation of this research is the fact that it relies upon teachers' plans for teaching History A level for assessment in 2017 not changing and being accurate. The Specification Creator is a relatively casual instrument, and the teachers are by no means tied to the intentions stated. Feedback from the selection of teachers emailed as part of the research does indicate that schools largely entered topic choices that reflected their teaching intention as of summer 2015. Equally, there is little motivation for teachers to input misleading information at the outset; however situations at individual schools may change prior to the date of the final examination. This data therefore must be treated as speculative, rather than a definite representation of the final choices made.

Key observations

The common topic choices under each individual unit do not produce any unexpected findings. There is a predominance of Modern and Early Modern topics at the top of the popularity listing across all three units. None of the Medieval topics attract more than 10 per cent of schools. This is in line with the findings relating to the old specification (Child et al., 2014).

The most popular topics in British History relate to the Tudors and Stuarts, as well as very Modern History (e.g., *Y113: Britain 1930–1997*). In European history, the most popular topics concern Russia in the early 20th century and Germany's leadership in the first half of the 20th century. Worldwide, topics based on events in the United States are the most popular.

Whilst teachers have selected a range of topics within each unit, the most popular topics in each of the units are those that were popular in the old specification. The topics schools choose to teach as part of A level History have been found to be primarily based on teacher expertise and resource availability (Child et al., 2015). Therefore, it is unsurprising that topics which were introduced for the first time in the new specification do not feature in the top choices. Nonetheless, *Y312: Popular Culture and the Witchcraze of the 16th and 17th Centuries*, and *Y212: The American Revolution 1740–1796* were both brand new topics for this specification and both feature in the top 15 most popular units across the whole specification, indicating that there is appetite amongst schools for something new to teach.

References

Child, S., Darlington, E., & Gill, T. (2014). An analysis of the unit and topic choices made in an OCR A level History course. *Research Matters: A Cambridge Assessment Publication*, *18*, 2–9.

Child, S., Darlington, E., & Gill, T. (2015). A level History choices: Which factors motivate teachers' unit and topic choices? *Research Matters: A Cambridge Assessment Publication*: *19*, 2–6.

Department for Education. (2014). *GCE AS and A level subject content for history*. Retrieved from https://www.gov.uk/government/publications/gce-as-and-a-level-for-history

Department for Education. (2015). *School and College Performance Tables: 2014 Performance Tables Archive*. Retrieved from http://www.education.gov.uk/schools/performance/2014/index.html

Accounting for students' mathematical preparedness for Finance and Business degrees

Ellie Darlington and **Jessica Bowyer** Research Division

Introduction

Background: A level reforms

A wide-ranging reform programme of General Certificate of Education (GCE) Advanced level (A level) in Mathematics and Further Mathematics is currently underway, with new qualifications due for first teaching in England in 2017. All A levels are moving from a modular to a linear system, requiring students to take their examinations at the end of the two-year A level course, rather than throughout as is currently the case. Furthermore, the optionality currently available in the choice of A level Mathematics units will cease, with the content of this qualification becoming 100 per cent prescribed, whilst Further Mathematics will have 50 per cent prescribed content. Although this will assist in reducing the variability in students' mathematical backgrounds when they begin university study, the Applied Mathematics content (currently available in Statistics, Mechanics and Decision Mathematics topics) that students are able to study will therefore be reduced.

These two qualifications prepare students for the workplace or undergraduate study in a range of STEM (Science, Technology, Engineering and Mathematics), Medicine and Social Science subjects. Consequently, the reforms will have implications for a large number of new undergraduates. This article reports on part of a large-scale study of over 4,000 undergraduates and 30 lecturers of these subjects regarding their perceptions of the existing A levels as preparation for the mathematical components of their degrees, as well as their motivations for, and experiences of, studying Further Mathematics (Darlington & Bowyer, 2016).

Business Studies is a broad field of study. Indeed, MacFarlane (1997) argues that this is "an eclectic, multi-disciplinary area" and that there is "no singular concept of 'Business Studies'" (p.7). Therefore, this study sought to ascertain the views of students of a discipline whereby mathematical skills are highly important, yet rarely demanded of prospective applicants at A level before commencing university study.

Undergraduate Business Studies

In the United Kingdom (UK), the number of students studying full-time for undergraduate degrees in the area of Accounting, Business and Management has been steadily increasing since the early 2000s. Data from the Higher Education Statistics Agency (HESA) (2016) show that the proportion of all UK undergraduates studying for these courses increased consistently throughout this time, last year comprising 13.8 per cent of the UK's undergraduate student body[1].

At the school level, there were 26,745 A level Business Studies candidates in 2014 (3.2% of all A levels), a number which has been decreasing since a peak of 36,834 students (4.9% of all A levels) in 2001 (Joint Council for Qualifications [JCQ], 2015). This change may be in response to the fact that students have begun to opt for more traditional subjects, such as those recommended by the Russell Group (2013).

Mathematics in Business Studies

Mathematics requirements for undergraduate courses in the UK

A study by the Higher Education Academy (HEA) regarding the mathematical preparedness of undergraduate students of Business and Management degrees (Cottee, Relph, & Robins, 2014) found that, of the 131 English universities offering Business and Management courses, for 2013 entry:

- 41% did not specify a Mathematics requirement;
- 40% required a grade C at General Certificate of Secondary Education (GCSE);
- 16% required a grade B at GCSE;
- 2% required a grade A at GCSE; and
- only 1% required A level Mathematics.

However, although Mathematics requirements for entry to study Business-related degrees are varied and reasonably low, more than a quarter of new undergraduates studying Business in 2011 had A level Mathematics (Vidal Rodeiro & Sutch, 2013, p.17). Furthermore, in 2011, 9.8 per cent of A level Mathematics students went on to study Business and Administrative Studies (BAS) degrees at university (Vidal Rodeiro, 2012, p.5). The most popular A level subjects amongst these students are given in Table 1. Of these 2.7 per cent had taken A level Further Mathematics (Vidal Rodeiro & Sutch, 2013, p.16).

Table 1: Top 10 most popular A level subjects amongst Business and Administrative Studies students (Vidal Rodeiro & Sutch, 2013)

Rank	*Subject*	*Proportion of students (%)*
1	Business Studies	38.0
2	Mathematics	27.3
3	Economics	17.9
4	Psychology	15.1
5	General Studies	13.0
6	History	11.2
7	Geography	10.4
8	English Literature	9.6
9=	Media Studies	9.5
9=	Sociology	9.5

[1] HESA produces data according to the University and Colleges Admissions Service (UCAS) subject classification codes (JACS). The subject area which most closely matches the subject of this article is 'Business and Administrative Studies'.

The HEA's (2014) study into the mathematical preparedness of Business Studies undergraduates also interviewed lecturers and undergraduates. Interviews with lecturers of these courses revealed that they did not feel that a grade C in GCSE Mathematics was an adequate entry requirement. However, pressures to recruit sufficient students for the course meant that lecturers did not believe it was possible to raise the requirements. Furthermore, only 87% of students knew that there would be quantitative elements to their degree, and 26% reported that they encountered more Mathematics than they had expected. Nearly a quarter reported that they found quantitative methods (QM) to be different to what they had expected, and 20% described themselves as "someone who struggles with quantitative methods" (Cottee et al., 2014, p.25). This is despite the same study indicating that the majority of degree programmes in the area of Business and Management have compulsory QM courses in the first year. In the United States, the picture is different – a large (*N*=684) American study found that students majoring in Business Studies were generally positive about their experience of the statistical elements of their course, more so than students of other Social Sciences (Griffith, Adams, Gu, Hart, & Nichols-Whitehead, 2012).

The minimal Mathematics requirements for Finance, Business and Management (FBM) courses are therefore intriguing when contrasted with the Quality Assurance Agency for Higher Education (QAA)'s benchmark statement for 'general Business and Management' degrees. The QAA specifies that graduates of these degrees must conduct "effective problem solving and decision making using appropriate quantitative and qualitative skills including identifying, formulating and solving business problems". Students should also develop "numeracy and quantitative skills including data analysis, interpretation and extrapolation" (QAA, 2007b, p.3). One might question whether a student should be able to demonstrate a capability in these areas before beginning their course, or whether universities are expected to teach these areas to students from scratch. The low Mathematics entry requirements suggest that universities either teach this content to their students, or expect that GCSE Mathematics is sufficient to equip students.

Similarly, the benchmark statement for Accountancy degrees (QAA, 2007a) states that graduates must have "numeracy skills, including the ability to manipulate financial and other numerical data and to appreciate statistical concepts at an appropriate level" (p.3). This makes specific reference to the development of statistical skills, something which Levine (1992; cited by Parker, Pettijohn, & Keillor, 1999) found when researching the topics taught in quantitative courses for undergraduate Business Studies students. The five most commonly covered topics were estimation and hypothesis testing, probability distribution, linear regression and correlation, descriptive statistics and tables and charts. Furthermore, Dunham (2002) claims that fundamental mathematical ideas in Finance include compound interest, present and future values, options pricing, debt repayment and cash flow.

A study of the most commonly taught mathematical topics in the top 50 business schools in the United States revealed all of these topics to be embedded in Statistics. Additionally, a study of 25 lecturers and heads of departments in UK universities that offer Business and Management degrees found that the areas of Mathematics taught most frequently included descriptive statistics, correlation and regression, graphical representation of data, the use of Excel, probability, algebraic manipulation, time series and forecasting, fractions, percentages and decimals, and calculus (Cottee et al., 2014). Again, the basis for these topics (excluding calculus) is in Statistics.

Impact of Mathematical backgrounds on performance

Empirical research into the impact of school Mathematics performance in undergraduate FBM degrees is mixed.

Surprisingly, a study by Rowbottom (2013) on a sample of 430 students at a Russell Group university, where 56.5 per cent of students had A level Mathematics, found no relationship between 'pre-university numeracy' and performance at any point in their Accounting degree. Similarly, Gammie, Jones, and Robertson-Millar (2003) found that prior performance in secondary Mathematics examinations in Scotland had no significant impact on the performance of a sample of 79 Accounting and Finance students at Robert Gordon University. A very small longitudinal study (*N*=39) by Bartlett, Peel, and Pendlebury (1993) found that those with A level Mathematics did not significantly outperform those without in Accounting examinations at a UK university.

However, Guney (2009) found that students with better GCSE and A level Mathematics grades performed better in Accountancy, although performance at GCSE was more indicative of future performance than at A level. The data suggested that it might be more important for admissions tutors to ask for high GCSE Mathematics grades than to ask that students have taken A level Mathematics. In the United States, Brookshire and Palocsay (2005) found that amongst 310 students, overall school performance had a greater impact on Business Studies students' performance than did their Mathematics performance alone, although this did have a positive impact. Additionally, Keef (1988) found that, in a New Zealand university, prior attainment and exposure to Mathematics had only a negligible effect on students' performance in Accounting.

Nevertheless, it has been found that stronger mathematical backgrounds have a positive impact on the performance of Business Studies, Accounting and Finance students in Hong Kong (Gul & Fong, 1993), Iran (Zandi, Shahabi, & Bagheri, 2012), the United States (Gist, Goedde, & Ward, 1996), Australia (Alcock, Cockcroft, & Finn, 2008), Canada (Standing, 2006), and Malaysia (Tho, 1994). Furthermore, Koh and Koh (1999) found that a Mathematics background based on achievement in International A level Mathematics grades had a significant impact on the performance of 526 students of Accountancy in Singapore. Indeed, Keef (1988) argues that Mathematics is a vital part of a Business undergraduate's education in the UK.

Many of the studies referenced in this review are rather old. The issue regarding mathematical preparedness of undergraduate FBM students is an issue that appears not to have been addressed for many years. The recent drive to promote STEM subjects has resulted in increased interest in this area (e.g., Cottee, 2014), though there are not a lot of publications. Education systems at the secondary and tertiary level change constantly and the nature of FBM and related disciplines have evolved over the last decades. Hence, caution should be taken when interpreting the outcomes of the research outlined in this section.

Changes to A level Mathematics and Further Mathematics

The research on which this article is based, summarised in Darlington and Bowyer (2016), was conducted in response to the forthcoming changes to A level Mathematics and Further Mathematics from 2017 (Department for Education [DfE], 2013). The nature of the reforms planned meant that the perspectives of current undergraduates regarding the current A levels were sought in order to inform the development of the new specifications, as well as to consider the implications of the reforms for universities and prospective students. It is therefore important to set the scene for this research in terms of outlining the content and structure of A level Mathematics and Further Mathematics.

AS and A level Mathematics

Presently, A level Mathematics comprises four compulsory Core Pure Mathematics units of equal weighting, with two Applied Mathematics units. These units may be chosen from the following strands:

1. Mechanics;
2. Statistics; and
3. Decision Mathematics.

It is not necessarily the case that students will be able to take the units that they want to. Restrictions on resources and timetabling within their schools and colleges may mean that they are given a restricted choice, if at all.

Within each of these strands are between two and five sequential units, depending on the particular strand and awarding body. The more advanced units (e.g., Mechanics 3 and above) can only be studied as part of Advanced Subsidiary (AS) or A level Further Mathematics.

Students may study either two units from the same strand (e.g., Statistics 1 and Statistics 2) or one from two different strands (e.g., Mechanics 1 and Decision Mathematics 1). Hence, there are six[2] possible routes through A level Mathematics.

At AS level, students must take two compulsory Core Pure Mathematics units and one applied unit (Mechanics 1, Statistics 1 or Decision Mathematics 1).

The reformed qualification will see the removal of optionality in the applied units. Students will all study a mixture of Statistics and Mechanics material (though not necessarily the same as the content of the current Statistics 1 and Mechanics 1 units), after the A level Content Advisory Board recommended the removal of Decision Mathematics from A level Mathematics (ALCAB, 2014).

AS and A level Further Mathematics

A level Further Mathematics comprises two compulsory Further Pure Mathematics units, plus four optional units. At AS level, students must take Further Pure Mathematics 1 and two optional units.

The optional units can be selected from any of the three standard Applied Mathematics strands offered within A level Mathematics or from an additional two Further Pure Mathematics units. There are therefore a large number of possible routes through Further Mathematics[3].

Method

A large number of different degree titles fall under the area of Business Studies, most of which require a level of mathematical competency. Hence, all universities which offered degrees in the area of Business Studies and Finance (including Accounting) were contacted, requesting their participation in the study. Relevant departments were asked to pass on the details of an online questionnaire aimed at students who fulfilled two criteria:

1. They must have been in their second year of study or above, in order that they could reflect on their experiences so far; and
2. They must have taken at least AS level Mathematics, and this must have been taken no earlier than 2006 (when the qualification underwent restructuring).

Those who took International A levels were not permitted to take part, as the structure and content of those qualifications are different to the domestic qualifications.

The questionnaire surveyed students regarding:

- their mathematical background;
- their current studies;
- their perceptions of the A level(s);
- the factors which motivated them to take Further Mathematics (if applicable); and
- their experience of Further Mathematics (if applicable).

The questionnaire comprised a mixture of multiple choice questions, closed questions and open-ended questions. It was developed by the authors and an A level Mathematics expert, before being piloted by three recent graduates of mathematically-demanding degrees. Small changes were made in response to the piloting. The questionnaire was made available in an online format, and was open for responses between September and December 2014.

Results

Sample

After data cleaning, a total of 104 responses were retained. It was considered inappropriate to conduct statistical testing for differences between groups in responses to the questionnaire due to the small sample size.

- **Institution of study:** Participants in the online questionnaire came from 25 different universities. There was an average of 4.1 participants per university (SD=3.1). Of the universities attended by participants, 89.3% attended universities in England, 4.9% in Scotland, 3.9% in Wales and 1.9% in Northern Ireland.
- **Degree programme:** Only 2.0% of participants were studying for undergraduate Master's degrees, with the remainder studying for Bachelor's degrees. Participants studied for 31 different specific degree courses, which have been simplified in this report into five degree areas (see Table 2). Most (55.8%) were studying for joint honours degrees within FBM, although some studied combinations with Law or Economics.
- **Year of study:** There was a mixture of participants currently in their second (60%), third (33%) and fourth (7%) years of study.

[2] (1) M1+M2; (2) S1+S2; (3) D1+D2; (4) M1+D1; (5) M1+S1; (6) D1+S1.

[3] Students are not allowed to take units as part of Further Mathematics that they have already taken as part of A level Mathematics.

Table 2: Students' degrees (includes joint honours)

Degree area	*No.*	*Proportion of participants (%)*
Accounting	44	42.3
Management	18	17.3
Business Economics	17	16.3
Finance	16	15.4
Business	9	8.7
Total	**104**	**100**

Participants' academic performance

A level performance: Participants had a mixture of backgrounds in A and AS level Mathematics and Further Mathematics. A quarter had taken both Mathematics and Further Mathematics to A level (see Table 3).

Table 3: Participants' A level qualifications

Mathematics qualification(s)	*No.*	*Proportion of participants (%)*
AS level Mathematics only	8	7.7
A level Mathematics only	61	58.7
A level Mathematics + AS level Further Mathematics	9	8.7
A level Mathematics + A level Further Mathematics	26	25.0
Total	**104**	**100.1***

*Due to rounding

In both subjects, most participants achieved an A or A* grade, which is disproportionate to the proportions of students who achieve these grades nationally (see Figures 1 and 2). Furthermore, the 77 per cent of participants who achieved an A* or A is a much higher proportion than the 32 per cent of all undergraduate FBM students who achieved the same grades in A level Mathematics in 2011[4]. This overrepresentation of high-achievers was taken into consideration throughout the analysis, although the majority of FBM students who took A level Further Mathematics in 2011 did have an A or A*.

Most participants were awarded their final Mathematics or Further Mathematics A or AS level in either 2012 or 2013 (42.4 per cent in each), with 1 student in 2006, 3 in 2010 and 11 in 2011.

A level Mathematics units: Participants were asked which optional units they studied as part of A and AS level Mathematics and Further Mathematics. The data suggest that it was most common for students to study a mixture of different areas of Applied Mathematics rather than specialising in one particular area (see Figure 3). It was more common for participants to have taken more Statistics units than Mechanics, with 60 participants indicating they had studied at least one Mechanics unit, and 89 indicating they had taken at least one Statistics unit.

University results: Students were asked about their performance in their previous year's examinations, where applicable (see Figure 4). Most participants were awarded Upper Second-class degree honours (usually a result of 60–69%), with small numbers achieving a Third-class degree result, and two students failing their examinations.

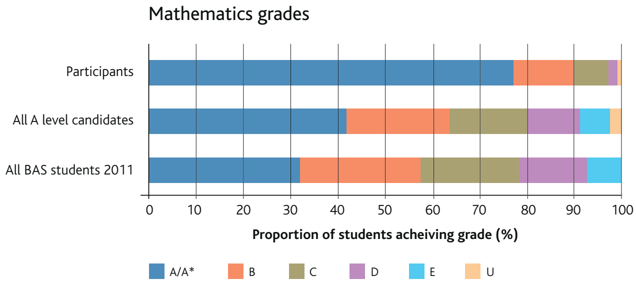


Figure 1: Participants' AS or A level Mathematics grades

Additional data from JCQ (2015) and Vidal Rodeiro[5] (2012).

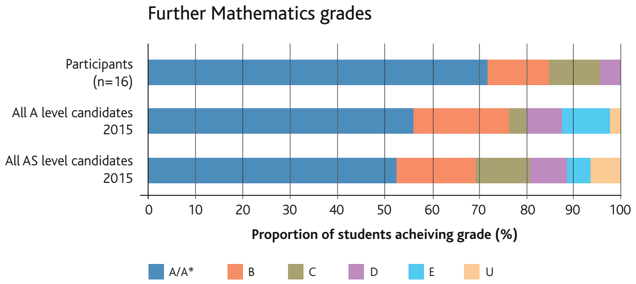


Figure 2: Participants' AS or A level Further Mathematics grades

Additional data from JCQ (2015).

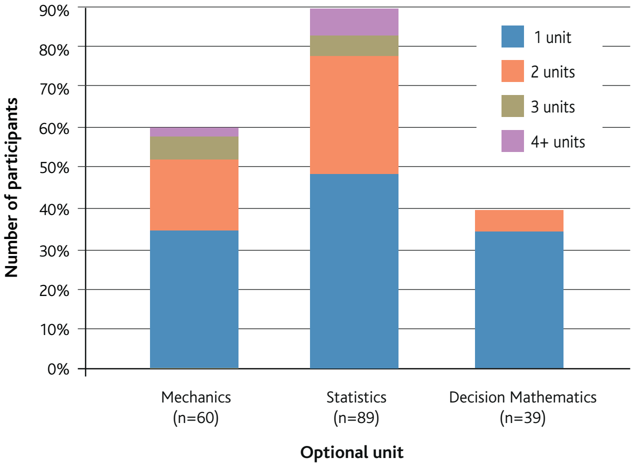


Figure 3: Optional units studied

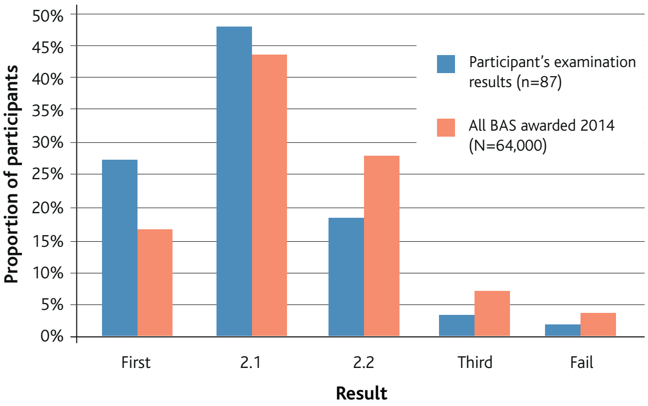


Figure 4: Previous year's examination results

Additional data from HESA (2015).

[4] 2011 is the most recent year for which this type of data is available.

[5] 'Business and Administrative Studies' is the most relevant grouping of student available in her study to this sample.

Figure 4 shows that this sample is perhaps more representative of the high-achievers; however, it should be noted that the participants may have performed better in their end-of-year examinations than they would do in their final degree examinations.

Which optional units are most helpful?

The data suggest that the most useful of the optional units for FBM undergraduates to have studied at A level are in Statistics (see Figure 5). Of the participants who took Statistics, 96.6% reported that they found it very or somewhat useful. Mechanics and Decision Mathematics were considered to be of similar utility to each other (less than 40% found them very or somewhat useful). Additionally, approximately 80% found Further Pure Mathematics units to have been useful, although of lower utility than Statistics.

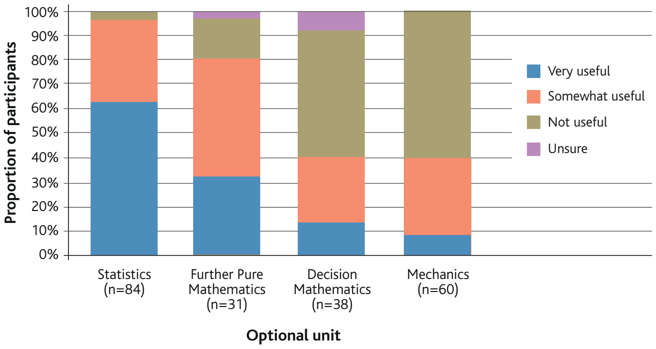


Figure 5: Participants' views of the utility of optional units

What motivates students to take Further Mathematics?

Approximately 34% of participants had taken Further Mathematics to AS or A level, meaning that their motivations for doing so could be investigated.

When asked to indicate factors which motivated their decision to study Further Mathematics from a list of 15 statements[6] (see Figure 6), it emerged that the participants had mainly been influenced by three main areas in their decision to study Further Mathematics:

- **An enjoyment of Mathematics:** 87.9% of participants reported that they were influenced 'a lot' by an enjoyment of school Mathematics. Only one participant reported that this did not influence their decision to study Further Mathematics.
- **Perceived utility:** Not only did 68.8% of participants report that they were heavily influenced by the utility of Further Mathematics, but 87.9% reported that they were influenced to some extent by the consideration of studying for a Mathematics or Mathematics-related degree at university.
- **Fit with other A levels:** Most participants (81.8%) reported that Further Mathematics fitting well with their other subject choices had some influence on their decision.

The data suggest that very few students were strongly influenced by what their peers were studying and their school Mathematics department's results, and that there was no strong parental influence.

What are students' experiences of studying Further Mathematics?

Students who studied Further Mathematics were asked to describe their experiences of studying it. Their responses were largely positive (see Table 4).

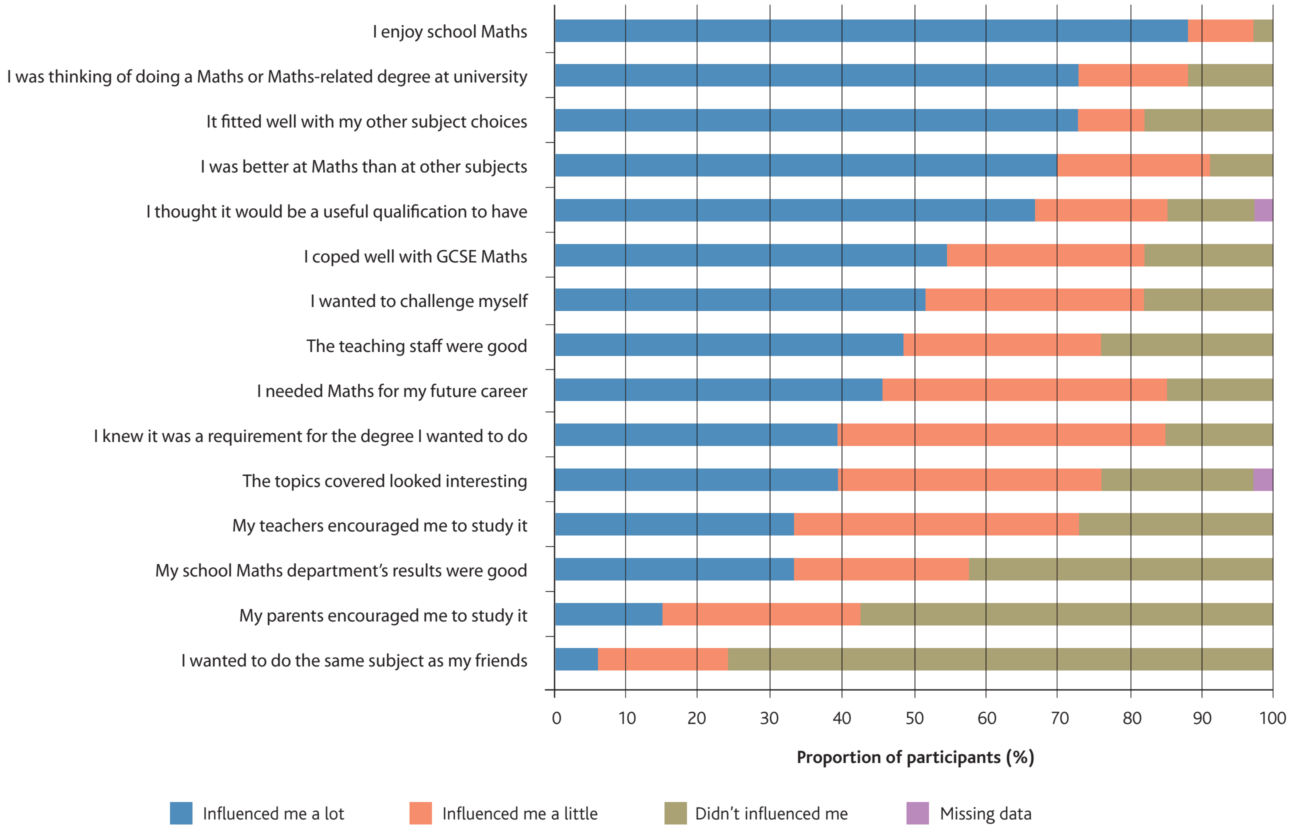


Figure 6: Participants' motivations for studying Further Mathematics (n=33)

[6] These statements were taken from a study by the Qualifications and Curriculum Authority (2006) which examined student participation in A level Mathematics, but are applied here in the context of Further Mathematics.

Table 4: Participants' experiences of studying Further Mathematics

Statement	*Strongly agree*	*Agree*	*Neither agree nor disagree*	*Disagree*	*Strongly disagree*
	Number of participants (%)				
I am glad I took Further Maths	20 (60.6%)	10 (30.3%)	1 (3.0%)	2 (6.1%)	0 (0.0%)
I took Further Maths because I was thinking of doing a Maths or Maths-related degree at university	17 (51.5%)	8 (24.2%)	4 (12.1%)	3 (9.1%)	1 (3.0%)
I found Further Maths challenging	15 (45.5%)	11 (33.3%)	7 (21.2%)	0 (0.0%)	0 (0.0%)
I enjoyed Further Maths	13 (39.4%)	15 (45.5%)	3 (9.1%)	2 (6.1%)	0 (0.0%)
Further Maths was my most difficult A level	11 (34.4%)	7 (21.9%)	6 (18.8%)	7 (21.9%)	1 (3.1%)
In my first year at university, we were taught material that I had learned in Further Maths	10 (30.3%)	10 (30.3%)	3 (9.1%)	5 (15.2%)	5 (15.2%)
Most people on my university course studied Further Maths	3 (9.4%)	5 (15.6%)	6 (18.8%)	13 (40.6%)	5 (15.6%)

Only a quarter of participants reported that they thought that most people studying their university course had taken Further Mathematics, although 60.6% agreed that they had covered material that they had learned in Further Mathematics during their first year at university.

This overlap suggests that there may be benefits to studying Further Mathematics in addition to Mathematics in order to ease the transition into the mathematical element of FBM degrees. However, only 39.4 % of participants reported that they strongly agreed that, 'Studying Maths and Further Maths was sufficient preparation for my degree'. Conversely, it could also be argued that an overlap in A level Further Mathematics content and first year undergraduate Mathematics could mean that students become bored. However, repeating material that students are already familiar with would give them an advantage.

Overall, 84.9% of participants agreed that they enjoyed Further Mathematics and 90.9% agreed that they were glad that they had taken it. However, whilst 78.8% reported that they found it challenging, and 81.8% that it was more demanding than A level Mathematics, only 56.3% reported that it was their most difficult A level.

How useful are the A levels?

The data suggest that students were largely in agreement that A level Mathematics and Further Mathematics were good preparation for the mathematical component of their degree (see Figure 7).

A large majority of participants (83.7%) indicated that Mathematics was good preparation for their degree, with a smaller majority (65.6%) indicating the same of Further Mathematics. No participants with a Further Mathematics qualification described it as bad preparation, and only one participant reported the same of A level Mathematics.

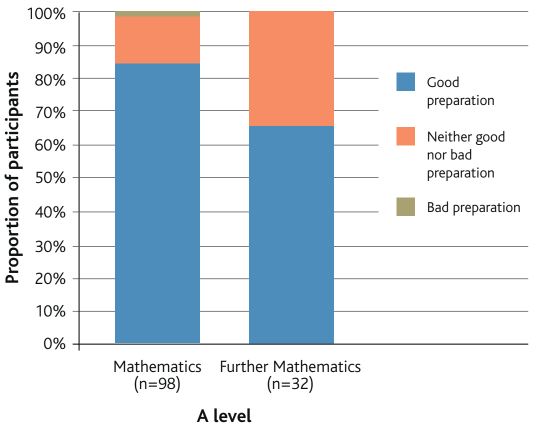


Figure 7: Students' perceptions of the utility of the A levels as preparation for their degree

What improvements could be made?

In addition to the multiple choice questions, participants were asked two open response questions. The first question asked whether there were any additional topics that are not currently incorporated in the A level courses that would have been useful. There were 46 responses. The majority of suggestions for additional topics focused on the inclusion of Financial Mathematics, especially Basic Accountancy for students on Accountancy courses and Basic Economics for participants on Economics-based courses. A smaller proportion of students also suggested that the statistical content at A level should be made harder and more in-depth, with specific topics focusing on a greater variety of distributions. These topics are depicted in Table 5.

The second question asked participants whether there were any improvements that could be made to the A levels to make them better preparation for FBM undergraduate courses. There were 61 responses. Comments about potential improvements to both A levels centred on suggestions that they should cover a greater depth of content. Most participants also suggested that the inclusion of more real-world applications would be beneficial, particularly in contexts relevant to Business or Economics. The responses also indicated that students had varying opinions about the difficulty of the A levels. Broadly similar proportions of participants reported that an increase in difficulty would be welcome, or that the existing level of challenge had been sufficient preparation for their degree course.

Additional, but less frequent, suggestions were that there could be a stronger relationship between the style of examination question at A level and at university, and that a greater understanding of the material and theory would have been beneficial.

Table 5: Topics participants suggested for inclusion at A level

Topic area	*Topics*	*OCR Mathematics Unit(s) covering topic*[7]
Financial Mathematics	Basic Accountancy	
	Econometrics	
Probability and Statistics	Computer-based software	
	Continuous probability distribution	S2, S3
	Advanced hypothesis testing	
	p, *z*- and *t*-values	S2
	Bayes' theorem	S4
	Regression models	
Calculus	Matrices	FP1
	Partial differentiation	

[7] Source: OCR (2013).

Limitations

A self-selecting study of this nature suffers from a number of classic limitations, as well as some limitations specific to this study:

- Participation was doubly self-selecting. That is, students self-selected in their decision to complete the questionnaire, but their opportunity to do so was also based on self-selection on the part of the university departments which were the vital link between the researchers and the students. Data reported in Figures 2, 3 and 5 were therefore compared with national data in order to give an indication as to whether this sample might be skewed in terms of its composition.
- It could be that students who felt particularly strongly (either positively or negatively) about their mathematical preparedness and its impact on their transition to tertiary study may have felt more compelled to take part.
- This study only incorporates the views of students who had taken post-compulsory Mathematics qualifications. We cannot contrast their responses with students who did not take A levels in Mathematics and/or Further Mathematics.
- Finance, Business and Management are a wide field of study. Therefore, it is possible that degrees in Management, for example, may be less mathematically demanding than degrees in Finance. The responses of participants studying across these different areas were compared using statistical analysis where sample sizes were large enough to do so. No significant differences were found between groups, though caution should be taken when interpreting the data outlined in this article.

Implications and recommendations

The data collected in this work suggests that current FBM students regard both A level Mathematics and Further Mathematics as good preparation for the mathematical content of their degree. In particular, Statistics units were considered to be the most useful applied units, with 96.6 per cent of participants describing them as either very or somewhat useful preparation.

These findings indicate that, despite conflicting results in the studies outlined in the introduction, prior Mathematics qualifications may benefit students' performance in undergraduate FBM courses. That students regarded Statistics as the most useful optional units may seem unsurprising when considering the type of Mathematics commonly required in FBM courses. In particular, sampling methods, hypothesis testing, probability and confidence intervals have been found to be commonly taught topics in first-year courses (Haskin & Krehbiel, 2011), and the Statistics units in A level Mathematics offer basic grounding in these areas. Consequently, prospective FBM students would benefit from specialising in Statistics during their A level studies.

The proposals for the reformed A level Mathematics mean that some Statistics content will become compulsory for all students. This will reduce the variability in undergraduates' Mathematics backgrounds, which is beneficial for admissions tutors. However, it also has implications for students who would have benefitted from specialising in one strand. For example, a student going on to study FBM at university, under the current system, would benefit from taking Units S1 and S2 in A level Mathematics. However, new proposals mean that learning Statistics in depth would require a student to take Further Mathematics.

Sampling, hypothesis testing, *t*-tests, and statistical significance will become compulsory content in the reformed A level Mathematics. Furthermore, there is a new requirement that students handle real, large datasets, although there is currently no guidance on exactly how large these datasets are expected to be, nor how this will be assessed. It therefore seems likely that taking A level Mathematics will continue to be good preparation for FBM courses once reforms have taken place. Furthermore, studying Further Mathematics also appears to be beneficial preparation for prospective FBM undergraduates. Although most participants reported that they did not believe the majority of their peers had taken Further Mathematics, they reported that they had covered some material from Further Mathematics in their first year at university. Additionally, no participant reported that Further Mathematics had been poor preparation for their undergraduate studies. Participants' positive opinions about Further Mathematics, coupled with the overlap in material, suggest that Further Mathematics is a useful qualification for FBM undergraduates to have.

It is not immediately clear whether the benefit of taking Further Mathematics lies in the opportunity to study more advanced Statistics units, or in the exposure to advanced Pure Mathematics content. Participants were very positive about the utility of Statistics units, but Further Pure Mathematics units were also well-received. Moreover, students' suggestions for additional topics to be included at A level incorporated calculus and matrix algebra as well as statistical topics. This suggests that both areas are beneficial preparation. The reform of A level Further Mathematics thus has implications for the preparedness of FBM undergraduates in the future, as the awarding bodies decide what optional content should be available for students to choose.

A further implication of this study is that, given the mathematical entry requirements for FBM courses are very low, universities may wish to reconsider their current requirements and schools and careers advisers should take note. Given that Mathematics has been the most popular A level subject overall for the past two years, and participants in this study were enthusiastic about their experience of post-compulsory Mathematics, it is not unreasonable to suggest that universities ask prospective students for at least AS level Mathematics. A level reform provides an opportune time for admissions departments to review their current entry requirements in light of the forthcoming changes. Those giving students advice when choosing A level subjects should also be made aware that, though A levels in Mathematics or Further Mathematics are not generally required of students going on to study FBM, there are clear benefits.

Additionally, the introduction of new Level 3 Core Mathematics qualifications[8] may also be of interest to FBM departments, as these courses will allow students who do not wish to study A level Mathematics to develop their statistical competency. With the introduction of compulsory statistical content in AS and A level Mathematics and the proliferation of post-compulsory Mathematics courses, the opportunities available to prospective FBM students to increase their mathematical preparedness before university are increasing. Universities can therefore take advantage of these developments in order to increase the overall mathematical proficiency of new cohorts.

[8] Core Mathematics qualifications will be aimed at students who achieved at least a grade C in GCSE Mathematics, but who do not wish to study A level Mathematics. Some specifications will contain substantial amounts of statistical content.

References

ALCAB. (2014). Report of the ALCAB Panel on Mathematics and Further Mathematics. Retrieved from https://alevelcontent.files.wordpress.com/2014/07/alcab-report-on-mathematics-and-further-mathematics-july-2014.pdf

Alcock, J., Cockcroft, S., & Finn, F. (2008). Quantifying the advantage of secondary mathematics study for accounting and finance undergraduates. *Accounting & Finance*, *48*(5), 697–718.

Bartlett, S., Peel, M. J., & Pendlebury, M. (1993). From Fresher to Finalist. *Accounting Education*, *2*(2), 111–122.

Brookshire, R. G., & Palocsay, S. W. (2005). Factors Contributing to the Success of Undergraduate Business Students in Management Science Courses. *Decision Sciences Journal of Innovative Education*, *3*(1), 99–108.

Cottee, M. J., Relph, A., & Robins, K. (2014). *Skills in Mathematics and Statistics in Business and Management and tackling transition*. York: The Higher Education Academy.

Darlington, E., & Bowyer, J. (2016). The Mathematics Needs of Higher Education. *Mathematics Today*, *52*(1), 9.

DfE. (2013). *Reformed GCSE Subject Content Consultation: Government Response*. London: Department for Education.

Dunham, B. (2002). Short Course: Teaching Statistics in Finance. *MSOR Connections*, *2*(4), 50.

Gammie, E., Jones, P. L., & Robertson-Millar, C. (2003). Accountancy Undergraduate Performance: A Statistical Model. *Accounting Education*, *12*(1), 63–78.

Gist, W. E., Goedde, G., & Ward, B. H. (1996). The influence of mathematical skills and other factors on minority student performance in principles of accounting. *Issues in Accounting Education*, *11*(1), 49–59.

Griffith, J. D., Adams, L. T., Gu, L. L., Hart, C. L., & Nichols-Whitehead, P. (2012). Students' Attitudes Toward Statistics Across the Disciplines: A Mixed-Methods Approach. *Statistics Education Research Journal*, *11*(2), 45–56.

Gul, F. A., & Fong, S. C. (1993). Predicting Success for Introductory Accounting Students; Some Further Hong Kong Evidence. *Accounting Education*, *2*(1), 33–42.

Guney, Y. (2009). Exogenous and Endogenous Factors Influencing Students' Performance in Undergraduate Accounting Modules. *Accounting Education*, *18*(1), 51–73.

Haskin, H. N., & Krehbiel, T. C. (2011). Business Statistics at the Top 50 US Business Programmes. *Teaching Statistics*, *34*(3), 92–98.

Higher Education Statistics Agency. (2015). Table 16 – HE qualifications obtained by subject of study, level of qualification and class of first degree 2013/14. Retrieved from https://www.hesa.ac.uk/dox/dataTables/studentsAndQualifiers/download/Qualsub1314.xlsx

Higher Education Statistics Agency. (2016). Table 4 – HE student enrolments by level of study, subject area, mode of study and sex 2010/11 to 2014/15. Retrieved from https://www.hesa.ac.uk/dox/pressOffice/sfr224/061046_student_sfr224_1415_table_4.xlsx

Joint Council for Qualifications. (2015). GCE A Level Results. Retrieved from http://www.jcq.org.uk/Download/examination-results/a-levels/2015/a-as-and-aea-results-summer-2015

Keef, S. P. (1988). Preparation for a First Level University Accounting Course: The Experience in New Zealand. *Journal of Accounting Education*, *6*(2), 293–307.

Koh, M. Y., & Koh, H. C. (1999). The Determinants of Performance in an Accountancy Degree Programme. *Accounting Education*, *8*(1), 13–29.

Levine, D. M. (1992). Business Statistics Curricula Lack Quality. *Quality Progress*, 77–79.

MacFarlane, B. (1997). In search of an identity: lecturer perceptions of the business studies first degree. *Journal of Vocational Education & Training*, *49*(1), 5–20.

OCR. (2013). OCR Advanced GCE and Advanced Subsidiary GCE in Mathematics, Pure Mathematics and Further Mathematics. Retrieved from http://www.ocr.org.uk/Images/67746-specification.pdf

Parker, R. S., Pettijohn, C. E., & Keillor, B. D. (1999). The nature and role of statistics in the business school curriculum. *Journal of Education for Business*, *75*(1), 51–54.

Qualifications and Curriculum Authority. (2006). Evaluation of Participation in A Level Mathematics: Interim Report, Autumn 2005.

Quality Assurance Agency for Higher Education. (2007a). *Subject Benchmark Statement: Accounting*. Mansfield: QAA.

Quality Assurance Agency for Higher Education. (2007b). *Subject Benchmark Statement: General Business and Management*. Mansfield: QAA.

Rowbottom, N. (2013). A-Level Subject Choice, Systematic Bias and University Performance in the UK: The Case of Accounting. *Accounting Education*, *22*(3), 248–267.

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

Standing, L. G. (2006). Why Johnny Still Can't Add: Predictors of University Students' Performance on an Elementary Arithmetic Test. *Social Behavior and Personality: an international journal*, *34*(2), 151–159.

Tho, L. M. (1994). Some Evidence on the Determinants of Student Performance in the University of Malaya Introductory Accounting Course. *Accounting Education*, *3*(4), 331–340.

Vidal Rodeiro, C. L. (2012). *Progression from A level Mathematics to Higher Education*. Cambridge, UK: Cambridge Assessment.

Vidal Rodeiro, C. L., & Sutch, T. (2013). *Popularity of A level Subjects Among UK University Students: Statistical Report Series No. 52:* Cambridge, UK: Cambridge Assessment.

Zandi, G., Shahabi, A., & Bagheri, M. (2012). The Relationship Between Mathematics Excellency and Efficiency of Accounting Students. *Journal of Modern Accounting and Auditing*, *8*(10), 1419–1427.

Collaboration in the 21st century: Implications for assessment

Simon Child Research Division and **Stuart Shaw** Cambridge International Examinations

Background

In recent years, there has been an increasing focus on conceptualising and defining so-called *21st century skills*. The literature on 21st century skills includes a number of frameworks for categorising the skills and knowledge required for participation in the workplace and in society (Lai & Viering, 2012). These frameworks have been motivated by observed changes in how students (and others) have to apply and demonstrate their acquired knowledge; using advanced technologies within multicultural societies in an age of increasing economic competition (Suto, 2013). Examples include the Partnership for 21st Century Learning (P21®), Assessment and Teaching of 21st Century Skills (ATC21S) and the National Research Council (NRC).

Whilst definitions of 21st century skills differ in terms of the placement of individual skills within their frameworks (Silva, 2009), there is a degree of consensus established with regards to skill identification. Skills include creativity and innovation, critical thinking, problem solving, metacognition, information and ICT literacy, citizenship, communication, and collaboration (see Suto, 2013, for an overview). Recently, these skills have been linked to future economic prosperity for individuals and nations, as they provide key qualities required to succeed in the global skills race (see Development Economics, 2015; P21, 2008).

Given the current status of 21st century skills, there is an increased motivation to develop modes of assessment that allow students to demonstrate their abilities in these domains. As Shute and Becker (2010) note:

> *We need to re-think assessment, identify new skills and standards relevant for the twenty-first century, and then determine how to best assess students' acquisition of the new competencies… Moreover, the envisioned new competencies should include not only cognitive variables (e.g., critical thinking, reasoning skills) but also non-cognitive variables (e.g., teamwork, tolerance, tenacity) as the basis for new assessments to support learning.* *(p.3)*

The appropriate assessment of 21st century skills is also important as it provides value and motivation to students, and can help structure pedagogical approaches (e.g., Swan, Shen, & Hiltz, 2006). However, any assessment has to resolve tensions related to its validity, reliability, comparability and delivery. Satisfactory construct definition for the purposes of assessment has always been considered an essential principle in testing. If these constructs are not well defined, then it is difficult to support the claims awarding bodies make about the usefulness of their assessments. Awarding bodies are challenged with the task of articulating how their assessments represent the target construct, how potential contaminating factors related to the assessment are controlled, and how the assessment achieves a desired level of reliability. This is challenging for 21st century skills due to the potential for subjectivity in the assessment process (Suto, 2013).

The status of collaboration in the 21st century

The focus of this article is the skill of collaboration. Collaboration has recently been identified as an important educational outcome in its own right, rather than just a means to develop or assess knowledge, which is learned through engagement and practice (Kuhn, 2015; Lai, 2011). Collaboration has been described as a skill that encourages learning mechanisms (such as induction, deduction and associative learning) to be enacted (Dillenbourg, 1999; Hunter, 2006).

The NRC (2011) outlined several justifications for collaboration's status as a key 21st century skill. First, there is a growing emphasis on project and enquiry-based learning. This is motivated by research that shows that collaboration has influential effects on student learning and knowledge retention (Fall, Webb, & Chudowsky, 1997; Rojas-Drummond & Mercer, 2003; Saner, McCaffrey, Stecher, & Bell, 1994; Webb, 1993). It is claimed that collaboration has distinct advantages over individual problem solving because it allows for: an effective division of labour; the incorporation of information from multiple sources of knowledge, perspectives, and experiences; and enhanced creativity and quality of solutions stimulated by ideas of other group members (Organisation for Economic Co-operation and Development [OECD], 2013). Similarly, collaboration has also been found to increase students' social competency (e.g., conflict resolution skills and use of helping behaviours) and academic self-concept (Ginsburg-Block, Rohrbeck, & Fantuzzo, 2006).

Secondly, there is an increasing need for students to be able to apply their knowledge and problem-solving skills in social settings (OECD, 2013). Organisations, faced with the need to innovate, use collaboration to combine the potential and expertise of their employees (Knoll, Plumbaum, Hoffmann, & De Luca, 2010). This is linked to recent advancements in technology, which have opened up new opportunities for how collaboration can be enacted (Salas, Cooke, & Rosen, 2008). The application of social technologies by individuals and across organisations has become a legitimate mode of enquiry (Blaskovich, 2008), and this ability has been regarded as important for the workforce of the future (OECD, 2013).

The stated importance of collaboration means that appropriately defining its construct remains an important aim. The main issue here is that the notion of collaboration, although almost universally accepted as being useful for application in the classroom and beyond, is conceptually vague (Brna, 1998). Different frameworks of 21st century skills place collaboration as either a *learning* skill (P21, 2015), an *interpersonal* skill (NRC, 2011) or a *way of working* (ATC21S, 2015). These frameworks have different conceptualisations of collaboration as a construct, and in terms of its interaction with other skills (Lai & Viering, 2012).

Aims of the article

When assessing collaboration, there is a need for a clear understanding of what is being tested, based on a theoretically-sound and agreed upon definition. In light of this important issue, this article has two main intentions. First we aim to provide an overview of how collaboration is conceptualised, and how it is distinguished from other related group activities (e.g., cooperation). Integral to this aim is the ambition to develop a coherent understanding of the abilities underlying the targeted construct.

The second aim is to discuss how the conceptualisations of collaboration underpin the development of appropriate methods of assessment. Specifically, we explore how the task given to students can potentially optimise the opportunities for collaboration to occur amongst group members. We also consider how different conceptualisations of collaboration are currently assessed, and the issues raised in the development of large-scale assessment.

Defining the construct of collaboration

The basic facets of what constitutes a collaborative activity are reasonably well rehearsed in the literature. Academics who have attempted to delineate collaboration from other related activities have articulated three fundamental aspects to collaborative learning. These three aspects are expressed in the definition provided by the OECD (2013):

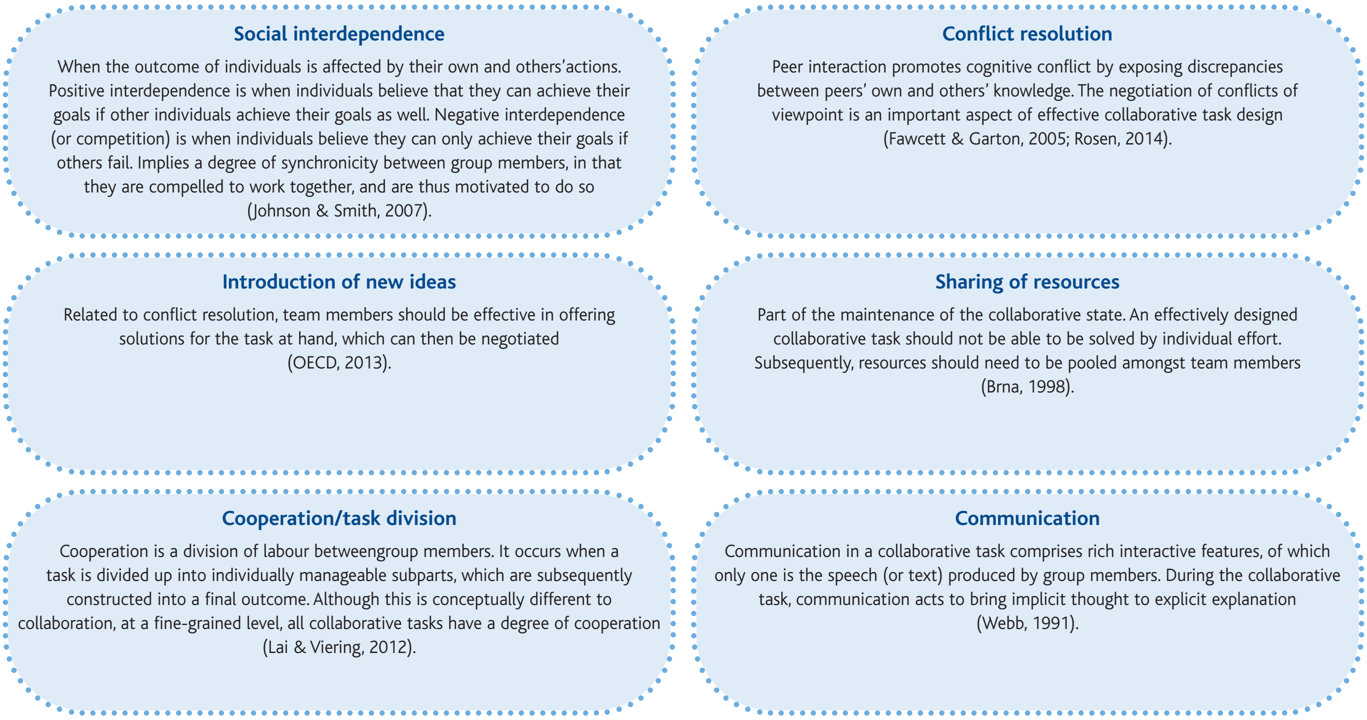
> *Collaborative problem solving competency is the capacity of an individual to effectively engage in a process whereby* ***two or more agents*** *attempt to* ***solve a problem*** *by* ***sharing the understanding and effort*** *required to come to a solution and pooling their knowledge, skills and efforts to reach that solution. (p.6) [emphasis added]*

Each of the three emphasised aspects are important factors in the maintenance of collaborative activity. For a collaborative 'state' to be constructed (Brna, 1998) there has to be a task where the achievement of the goal requires more than one person to pool resources. This view is shared by Roschelle and Teasley (1995), who broadly define collaboration as a "coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem" (p.70), and Dillenbourg (1999) who defines collaboration as "a situation in which two or more people learn or attempt to learn something together" (p.1). The sharing of roles and responsibilities during collaboration closely relates to the concept of the 'joint problem space' (or JPS, Roschelle & Teasley, 1995). The JPS implies that group members enter into a social contract with the joint aim of achieving a desirable outcome. In this sense, group members enter into a collaborative 'state' (Brna, 1998) that has to be effectively maintained until the problem is solved, or the outcome is reached.

There is an important distinction to be made here between collaboration as it has been defined above, and *cooperation*. These terms have often been used synonymously when referring to group-related activities (Lai & Viering, 2012) yet have important conceptual differences. Cooperation is typically a *division of labour* among group members, but can also be part of a process of allowing information to be accessed amongst group members. It occurs when a task is divided up into individually manageable subparts, which are subsequently constructed into a final outcome. To cooperate in this way, group members do not need to maintain a mutual understanding of the task goals, as individuals simply focus on their subtasks. It can also encourage asymmetric individual contributions towards the task goal. Collaboration, on the other hand, contains inherent flexibility of roles and responsibilities with regards to the various subtasks in achieving a goal (Lai, 2011).

Another key distinction is between collaboration *as process* and collaboration *as outcomes*. The collaborative 'state' is related to *process*. This broadly relates to how well the collaborative state is maintained and progressed. We have identified six fundamental facets of the collaborative process in Figure 1.

On the other hand collaboration as *outcome* implies that the final product takes precedence over the means to achieve the goal.



Social interdependence
When the outcome of individuals is affected by their own and others'actions. Positive interdependence is when individuals believe that they can achieve their goals if other individuals achieve their goals as well. Negative interdependence (or competition) is when individuals believe they can only achieve their goals if others fail. Implies a degree of synchronicity between group members, in that they are compelled to work together, and are thus motivated to do so (Johnson & Smith, 2007).

Conflict resolution
Peer interaction promotes cognitive conflict by exposing discrepancies between peers' own and others' knowledge. The negotiation of conflicts of viewpoint is an important aspect of effective collaborative task design (Fawcett & Garton, 2005; Rosen, 2014).

Introduction of new ideas
Related to conflict resolution, team members should be effective in offering solutions for the task at hand, which can then be negotiated (OECD, 2013).

Sharing of resources
Part of the maintenance of the collaborative state. An effectively designed collaborative task should not be able to be solved by individual effort. Subsequently, resources should need to be pooled amongst team members (Brna, 1998).

Cooperation/task division
Cooperation is a division of labour betweengroup members. It occurs when a task is divided up into individually manageable subparts, which are subsequently constructed into a final outcome. Although this is conceptually different to collaboration, at a fine-grained level, all collaborative tasks have a degree of cooperation (Lai & Viering, 2012).

Communication
Communication in a collaborative task comprises rich interactive features, of which only one is the speech (or text) produced by group members. During the collaborative task, communication acts to bring implicit thought to explicit explanation (Webb, 1991).

Figure 1: The six facets of the collaborative process

This approach assumes that the task itself has encouraged collaborative processes to be enacted, and that the task is sufficiently complex that collaboration is required for its successful completion. The separation of process (i.e., how well the collaborative state is maintained and progressed) from *outcome* (i.e., the final product following a period of collaboration) is a key distinction that emerges from the literature, and has implications for how collaboration is optimally assessed. This is discussed in the next section.

Implications for the assessment of collaboration

The complexity of collaboration as a construct leads to two main challenges for assessors. First, they must create the working conditions necessary for collaboration to be engendered and encouraged. Secondly, they must be able to pinpoint aspects of an individual's behaviours within a group task, so that a judgement can be made about that individual's general capacities for collaboration. These issues are intimately related, with aspects related to collaboration formulation constraining (or optimising) the possibilities for assessment. The approach to assessment (e.g., the distribution of individual or group marks) can also influence the potential for collaboration to be engendered.

Here we first explore how the task given to candidates can be optimised so that collaborative processes can be observed. We then consider the modes of observation available to assessors interested in either the collaborative process or outcomes.

Pre-task

It is important to note that collaboration among group members is not an automatic outcome of setting a task with a shared goal (Kreijens, Kirschner, & Jochems, 2003). Indeed, there are significant barriers to collaboration taking place at all. For example, in some cases, group members may value achieving a quick consensus above the potential difficulties that can be encountered when introducing new ideas or negotiating contrasting positions. This 'rapid' consensus may be of detriment to the eventual outcome (e.g., Janis, 1982; Rimor, Rosen, & Naser, 2010).

Collaborative success is therefore dependent on establishing optimal group dynamics. Key aspects include the development of a sense of community among individual group members, setting up relationships among group members so that they all have the opportunity to perform the same range of actions, and an equality of status of individuals. Dillenbourg (1999) refers to symmetry on three planes, all of which are required for collaboration to occur:

1. **Symmetry of action:** The same range of actions is allowed to each group member.
2. **Symmetry of knowledge:** The group members have a similar level of expertise (but different viewpoints on the task).
3. **Symmetry of status:** Individual group members have a similar status with respect to other group members.

Whilst these points of symmetry refer to collaboration in numerous (although not all) contexts, it has some important implications for the effective assessment of collaboration. The first implication for assessors is that, before the group task is set, practitioners need to engender a sense of group identity and rapport amongst the group members. Similarly, high levels of trust and shared understanding, and depth of relationships have been identified as pre-conditions to collaboration (Monteiro & Morrison, 2014; Peters & Manz, 2007). Crucial to this is the role of the task setter, as they can encourage group members to build trust and mutual understanding *before* the assessment task (Mercer, 1996; Laurillard, 2012). To encourage true collaboration (which can then be observed and measured) assessors need to manipulate group members' experiences with one another so that channels of communication and mutual understanding are optimised *before* assessment commences.

Task setting

A fundamental element of a successful assessment of collaboration is that the task itself should provoke all members of the group to share their views and ideas on potential courses of action (Dillenbourg, 1999). As mentioned in the previous section, this can be encouraged by setting up effective pre-task relationships among group members. However, this alone is unlikely to be sufficient for collaborative strategies to be utilised. We have identified five criteria that assessors should meet when devising a collaborative problem-solving task. Some of these criteria relate specifically to the task itself, whilst others relate to aspects of group composition.

1. **Task is sufficiently complex:** The common factor in all assessments of collaboration is that group members are set a problem. Ideally the problem engenders alternative suggestions from within the group about the best course of action, or requires group members to research potential solutions to the problem. Overly simplistic or trivial tasks do not encourage group members to collaborate because there is little need to share cognitive load. High-quality collaborative tasks are thus likely to include elements of constructive argumentation (Brna & Burton, 1997).
2. **Task is ill-structured:** A good collaborative task is one that cannot be solved by one capable member of the group. Task complexity is, at least in part, determined by the structure of the task. Tasks should be sufficiently open, with more than one plausible (or defensible) solution (Webb, Nemer, Chizhik, & Sugrue, 1998). Furthermore, individual roles should not be designated by the task setter (at least initially) as this encourages unnecessary processing constraints. Strictly defined roles can also create the illusion of collaboration. This also introduces the problem of the group being led by the expectations of the task setter, which may restrict novel or innovative solutions.
3. **Task should utilise technologies that facilitate the collaborative process:** There are a number of ways in which technology can be introduced into a collaborative task: as a resource in information gathering; as a focus of the interaction; or as a collaborative partner. Tasks that involve group members collaborating using computers as a means of communication typically use email, instant messaging applications, discussion forums or videoconferencing. The advantage of these modes of communication is that they can enhance the reach of

communication, and increase the potential for different perspectives to be expressed (e.g., Thorpe, 1998). Despite these perceived advantages, it remains to be seen whether computer-supported means of communication within a collaborative task can overcome challenges created by the initial distance of participants from each other, both physically and in terms of the creation of a JPS (Kreijns, Kirschner, & Vermeulen, 2013).

4. **Group member dynamics engender negotiation:** Negotiation is unlikely if all group members agree on a solution to a problem, or if one group member forces their will or assumed knowledge onto another (e.g., in a tutoring scenario). When assessing collaboration, it is therefore important to place students in groups where there may be differences in opinion (Brna & Burton, 1997). However, the evidence on creating effective heterogeneous groups is mixed (see Webb et al., 1998). Some research has found that groups manifesting a range of abilities collaborate more productively compared to more homogeneous groups. This effect is observed more clearly in ill-structured tasks. Where the task is clearly specified, low-ability group members are more likely to display negative behaviours such as 'social loafing' (Salomon & Globerson, 1989). Social loafing appears to also be a function of group size. In general terms, the larger the group the more likely that some group members will not contribute to the task due to asymmetric interactions among group members.
5. **Group is motivated to work together:** In setting the task, the assessor needs to motivate group members to work together. If the criteria outlined above are met, then the group dynamic and the task itself are likely to be highly motivating. This is closely related to the concept of *social interdependence*, which is based on mutual encouragement and accountability (Johnson & Johnson, 1989, 2002 – see Figure 1). How group members are assessed during the collaborative task may contribute to developing social interdependence among group members. Research has found that productivity is improved when members are rewarded as a group, within a context of individual accountability (Bossert, 1988; Slavin, 1983).

Towards an assessment model for collaboration

The appropriate assessment of collaboration as a *process* or as an *outcome* reflects the distinct characteristics of these two conceptualisations.

Assessment of the collaborative process

The first aspect to consider regards the desirable characteristics of an individual who is effectively collaborating with their peers. We have identified six elements that comprise effective maintenance and progress of the collaborative state, as depicted in Figure 1. This framework may be a useful starting point in directing assessors to consider the fundamental skills within the collaborative process.

The next issue relates to how the process of collaboration can be optimally observed, from which judgements on performance can be made. Assessors have the challenging task of relating individuals' behaviours to both the context of the task and to the dynamics of the group. Appropriate adjustments of these judgements are required as group members negotiate and progress towards a solution, with a final 'best-fit' decision being made. In this process, the assessor implicitly creates an evidence base from which to ground their decision-making. The use of technology has been identified as a potential means from which observation of the collaborative process can be enhanced (e.g., Austin, Smyth, Rickard, Quirk-Bolt, & Metcalfe, 2010; MacDonald, 2003). For example, the use of wikis can provide a full record of individuals' contributions to a task, in addition to the responses from other group members (Judd, Kennedy, & Cropper, 2010). Taken together, assessors can analyse and reflect on these interactions off-line, potentially improving the evidence base from which judgements are made. However, different methods of analysis of these data are possible, and so careful consideration of how this evidence is used alongside more typical observational approaches is required.

Interestingly, there have been recent attempts to standardise the process of collaboration through the use of computer partners (see Rosen & Tager, 2013; OECD, 2013). These assessment procedures have the advantage of controlling the task scenario, so that the student can be encouraged to negotiate and offer different courses of action. It is debatable as to whether the level of control possible using this assessment method outweighs issues of ecological validity.

A third issue relates to the distribution of marks among individuals and the group. When marks are given to individuals, there is the potential for collaboration to become competition, and for individuals to feel aggrieved if their contributions are not noted. However, when marks are given at the group-level, this potentially obscures individual contributions. Further issues are raised when we consider that the usual aim of assessment (and qualifications more broadly) is for a judgement to be made on individuals. For any assessment of collaboration, then, it is imperative that group members are given individual marks. The focus of this individual mark, however, should centre on positive contributions to the collaborative process. The balance between group-level and individual-level marks for a collaborative task is an important consideration in the future development of models of assessment of collaboration.

Related to this issue is the origin of the marks: can a case be made for the assessors to be located *within the group*, via either self- or peer-assessment? These models of assessment have been identified as improving group processes, motivation and engagement, and achieving a good level of reliability (Mills & Glover, 2006; Race, 2001). However, concerns remain about their appropriateness as part of an assessment strategy for large-scale qualifications.

Assessment of collaborative outcomes

If the focus of assessment is on the learning achieved during collaboration, then the assessment itself should specifically relate to the quality of the final product. This is typically assessed by a terminal demonstration of learning either by a group presentation or the creation of a portfolio, where learning could be showcased (MacDonald, 2003). The use of portfolios, which are held centrally within a shared network, allows a longitudinal record of learning to be held by the assessor over time (Hauge & Wittek, 2002). This can encourage the assessor to understand each student's development of understanding of a topic area.

In assessing the outcome of individual learning within a collaborative context, two main considerations need to be made. First, the assessor needs to have measured each student's understanding of the topic of interest prior to the task, so that the 'before' and 'after' of learning can be established. Secondly, the assessor needs to set a task where learning relies to an extent on the collaborative process.

Conclusions and future questions

This article first aimed to briefly outline different conceptualisations of collaboration, and made the important distinction between the collaborative *process* (which is demonstrated within the collaborative activity) and the *outcome* (which is demonstrated by the quality of the knowledge or understanding of the group members). The article has also explored the implications for how the different constructs of collaboration can be assessed, focusing mainly on task conditions that need to be met for collaboration to be encouraged.

There remain several questions for future research. Specifically, the future development of effective assessments of collaboration relies on several decisions being made by developers regarding the desired direction of the assessment. These include the following:

What should be the focus of the assessment of collaboration?

The main distinction made here is between collaboration as process and collaboration as outcome. This decision routes the possible options for assessment. If the purpose of the assessment is to target the collaborative process, then assessment must focus on targeting individual contributions to the collaborative effort. However, some focus on project outcomes may be required for the purpose of student motivation, and to gather a more holistic view of a student's collaborative skill. If the aim of assessment is to measure student learning via collaboration (a specific form of collaborative outcome), then group-based assessment, for example, is not appropriate. Assessment of individual learning would likely rely on multi-stage assessment procedures.

Furthermore, the focus of assessment will be closely related to the other objectives of the target qualification. The relative importance of collaboration within the entire structure of the target qualification framework will have implications for its assessment.

If the focus of assessment is the collaborative process, how should the identified subskills be weighted?

We have identified six subskills that contribute to the collaborative process. However, the status of these skills, and how they can best be observed, is a source for future investigation.

What is the desired division of individual/group marks for students?

Giving an individual score to candidates meets the imperative for them to be rewarded for their contributions, and to prevent negative collaborative behaviours. The inclusion of a group score encourages a degree of mutual accountability which is essential in encouraging students to display the desired construct. To encourage full participation, both individual and group effort therefore need to be assessed. However, the weighting of this scoring approach remains an open question. For example, the idea of providing a single mark for an entire group related to the final outcome has been criticised on the basis of fairness.

How can technology be utilised to optimal effect?

Maintaining a consistent and reliable record of interaction is problematic, particularly when assessing large groups. For example, the use of online-based forums and wikis can provide a useful record of interactions between participants which can then be utilised for assessment purposes. Interestingly, the very process of introducing technology fundamentally changes the aspects of the interaction that makes collaboration more likely. Technology will need to overcome significant challenges for it to be a suitable mode from which collaboration can be derived and observed.

Acknowledgements

We are grateful to Irenka Suto, Research Division, for her helpful suggestions on an earlier draft of this article. We would also like to thank Sylvia Green, Research Division, Paul Bullen-Smith, Cambridge International Examinations, and Helen Eccles, formerly of Cambridge International Examinations, for their insightful discussion over the course of the research.

References

ATC21S (2015). *Assessment and Teaching of 21st Century Skills*. Official website. Available online at: www.atc21s.org

Austin, R., Smyth, J., Rickard, A., Quirk-Bolt, N., & Metcalfe, N. (2010): Collaborative digital learning in schools: Teacher perceptions of purpose and effectiveness. *Technology, Pedagogy and Education*, *19*(3), 327–343.

Blaskovich, J. L. (2008). Exploring the effect of distance: An experimental investigation of virtual collaboration, social loafing, and group decisions. *Journal of Information Systems*, *22*(1), 27–46.

Bossert, S. T. (1988). Cooperative activities in the classroom. *Review of Research in Education*, *15*, 225–250.

Brna, P. (1998). Models of collaboration. *Proceedings of the Workshop on Informatics in Education, XVIII Congresso Nacional da Sociedade Brasileira de Computação*, Belo Horizonte, Brazil.

Brna, P. & Burton, M. (1997). The computer modelling of students collaborating in learning about energy. *Journal of Computer Assisted Learning*, *13*, 193–204.

Development Economics. (2015). The value of soft skills to the UK economy. Retrieved from http://www.backingsoftskills.co.uk/The%20Value%20of%20Soft%20Skills%20to%20the%20UK%20Economy.pdf

Dillenbourg, P. (1999). What do you mean by 'collaborative learning?' In P. Dillenbourg (Ed.), *Collaborative-learning: Cognitive and Computational Approaches* (pp.1–19). Oxford: Elsevier.

Fall, R., Webb, N., & Chudowsky, N. (1997). *Group Discussion and Large-Scale Language Arts Assessment: Effects on Students' Comprehension*. CSE Technical Report 445. Los Angeles: CRESST.

Fawcett, L. M., & Garton, A. F. (2005). The effect of peer collaboration on children's problem-solving ability. *The British Journal of Educational Psychology*, *75*(2), 157–169.

Ginsburg-Block, M. D., Rohrbeck, C. A., & Fantuzzo, J. W. (2006). A meta-analytic review of social, self-concept, and behavioral outcomes of peer-assisted learning. *Journal of Educational Psychology*, *98*(4), 732–749.

Hauge, T. E. & Wittek, L. (2002). *Portfolios as mediators for collaborative learning and professional development in a distributed environment of teacher education*. Paper presented at the European Association for Research on Learning and Instruction (EARLI) Assessment Conference: Learning Communities and Assessment Cultures, Newcastle, England.

Hunter, D. (2006). Assessing collaborative learning. *British Journal of Music Education*, *23*(1), 75–89.

Janis, I. L. (1982). *Counseling on personal decisions: Theory and research on short-term helping relationships*. New Haven: Yale University Press.

Johnson, D. W., & Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Edina, MN: Interaction Book Company.

Johnson, D. W., & Johnson, R. T. (2002). Learning together and alone: Overview and meta-analysis, *Asia Pacific Journal of Education*, *22*(1), 95–105.

Johnson, D. W., & Johnson, R. T., & Smith, K. (2007). The state of cooperative learning in postsecondary settings. *Educational Psychology Review*, *19*(1), 15–29.

Judd, T., Kennedy, G. & Cropper, S. (2010). Using wikis for collaborative learning: Assessing collaboration through contribution. *Australasian Journal of Educational Technology*, *26*(3), 341–354.

Knoll, S. W., Plumbaum, T., Hoffmann, J. L., & De Luca, E. W. (2010). Collaboration ontology: Applying collaboration knowledge to a generic group support system. In G-J. De Vreede (Ed). *Proceedings of the Group Decision and Negotiation Conference 2010*, Delft, The Netherlands (p.37).

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in Human Behavior*, *19*, 335–353.

Kreijns, K., Kirschner, P. A., & Vermeulen, M. (2013). Social aspects of CSCL environments: A research framework. *Educational Psychologist*, *48*(4), 229–242.

Kuhn, D. (2015). Thinking together and alone. *Educational Researcher*, *44*(1), 46–53.

Lai, E. R. (2011). Collaboration: A Literature Review. Retrieved from http://images.pearsonassessments.com/images/tmrs/Collaboration-Review.pdf

Lai, E. R., & Viering, M. (2012). *Assessing 21st century skills: Integrating research findings*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, B.C., Canada.

Laurillard, D. (2012). *Teaching as a design science: Building pedagogical patterns for learning and technology*. Abingdon, UK: Routledge.

MacDonald, J. (2003). Assessing online collaborative learning: process and product. *Computers and Education*, *40*, 377–391.

Mercer, N. (1996). The quality of talk in children's collaborative activity in the classroom. *Learning and Instruction*, *6*(4), 359–377.

Mills, J., & Glover, C. (2006) *Using assessment within course structure to drive student engagement with the learning process*. Retrieved from http://www.open.ac.uk/fast/pdfs/John%20Mills.pdf

Monteiro, E. & Morrison, K. (2014). Challenges for collaborative blended learning in undergraduate studies. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *20*(7–8), 564–591.

National Research Council (2011). *Assessing 21st century skills*. Washington, DC: National Academies Press.

OECD (2013). Programme for International Student Assessment (PISA) 2015: Draft Collaborative Problem Solving Framework. Retrieved from http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Partnership for 21st Century Learning (2015). Official website. Available online at: http://www.p21.org/

Peters, L. M., & Manz, C. C. (2007). Identifying antecedents of virtual team collaboration. *Team Performance Management*, *13*, 117–129.

Race, P. (2001). *A briefing on self, peer, and group assessment. Assessment Series Number 9*. York, UK: Learning and Teaching Support Network.

Rimor, R., Rosen, Y., & Naser, K. (2010). Complexity of social interactions in collaborative learning: The case of online database environment. *Interdisciplinary Journal of E-Learning and Learning Objects*, *6*, 355–365.

Rojas-Drummond, S. & Mercer, N. (2003). Scaffolding the development of effective collaboration and learning. *International Journal of Educational Research*, *39*, 99–111.

Roschelle, J. & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem-solving. In C.E. O'Malley (Ed.), *Computer-supported collaborative learning* (pp.69–97). Berlin: Springer-Verlag.

Rosen, Y. (2014). Comparability of Conflict Opportunities in Human-to-Human and Human-to-Agent Online Collaborative Problem Solving, *Tech Know Learn*, *19*, 147–164.

Rosen, Y., & Tager, M. (2013). *Computer-based assessment of collaborative problem solving skills: Human-to-agent versus human-to-human approach*. Research & Innovation Network: Pearson Education.

Salas, E., Cooke, N. J., & Rosen, M. A. (2008). On teams, teamwork, and team performance: Discoveries and developments. *Human Factors*, *50*, 540–548.

Salomon, G. & Globerson, T. (1989). When teams do not function the way they ought to. *International Journal of Educational Research*, *13*(1), 89–100.

Saner, H., McCaffrey, D., Stecher, B., Klein, S., & Bell, R. (1994). The effects of working in pairs in science performance assessments. *Educational Assessment*, *2*(4), 325–338.

Shute, V. J., & Becker, B. J. (2010). *Innovative Assessment for the 21st Century*. *Supporting Educational Needs*. New York: Springer.

Silva, E. (2009). Measuring skills for 21st-century learning. *Phi Delta Kappa*, *90*(9), 630–634.

Slavin, R.E. (1983). When does cooperative learning increase achievement? *Psychological Bulletin*, *94*, 429–445.

Suto, I. (2013). 21st Century skills: Ancient, ubiquitous, enigmatic? *Research Matters: A Cambridge Assessment publication*, *15*, 2–8.

Swan, K., Shen, J., & Hiltz, S. R. (2006). Assessment and collaboration in online learning. *Journal of Asynchronous Learning Networks*, *10*(1), 45–62.

Thorpe, M. (1998). Assessment and "third generation" distance education. *Distance Education*, *19*(2), 265–286.

Webb, N. M. (1991). Task-related verbal interaction and mathematical learning in small groups. *Research in Mathematics Education*, *22*(5), 366–389.

Webb, N. M. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. *Educational Assessment*, *1*(2), 131–152.

Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative group assessment: Group composition and performance. *American Educational Research Journal*, *35*(4), 607–651.

The effect of subject choice on the apparent relative difficulty of different subjects

Tom Bramley Research Division

Introduction

The work presented here was prompted by a survey carried out by the Office of Qualifications and Examinations Regulation (Ofqual), of opinions on whether the grading of high-stakes academic examinations in England taken at age 16, for General Certificate of Secondary Education (GCSE), and at 18, for General Certificate of Education Advanced level (A level), should attempt in some way to make the different subjects equally 'difficult'. Ofqual published a suite of working papers to inform the debate[1], the second of which (Ofqual, 2015a) was a review of the United Kingdom literature on the topic. In our response to this survey (Cambridge Assessment, 2016), we expressed the view that while:

> *… in a small number of specific cases there may be some reasons for providing decision-makers with an indication of differences in subject difficulty, these are generally substantially outweighed by a much larger number of arguments against taking any of the options outlined by Ofqual to control for inter-subject comparability.* *(p.2)*

One of those arguments – the particular topic of this article – concerns whether it is valid to calculate statistical measures of relative subject difficulty based on the examinee-by-subject matrix containing the grades of each examinee on each subject in a particular examination session (For example, all GCSEs taken in the June 2016 session). There are several different methods of varying complexity that can be used to do this (see Coe, 2007). All of them face the same problems of first defining what is meant by 'difficulty', and second of dealing with the fact that the matrix of data to be analysed contains a large amount of missing data – the grades of examinees on subjects that they did not take. The non-random nature of this missing data (created by the fact that students only choose a subset of the possible subjects) makes the calculation of any statistical adjustment somewhat problematic. It is also likely to make subjects that measure something different to the majority of other subjects appear easier. These two claims are illustrated in this article with a simple example using simulated data.

Simulated data

Consider a scenario where only four subjects are available: Mathematics, Physics, Chemistry and Art. Assume that in the entire cohort of potential examinees that Mathematics, Physics and Chemistry are highly correlated with each other, but less so with Art – for example with a correlation matrix as in Table 1.

Table 1: Correlation matrix of scores for simulations (all potential examinees)

	Maths	*Physics*	*Chemistry*	*Art*
Maths	1.00	0.90	0.90	0.50
Physics		1.00	0.90	0.50
Chemistry			1.00	0.50
Art				1.00

The scores of 10,000 examinees were simulated to yield the above correlation matrix (scores in each subject normally distributed with a mean of 0 and a standard deviation (SD) of 1). The scores were converted to grades on an A* to G scale giving a value of 8 to A* and 1 to G, such that the overall distribution was roughly the same in each subject, and reasonably realistic (in fact it matched the national distribution of GCSE Mathematics grades in 2012[2]). Treating the grades as numeric variables, Tables 2 and 3 give the descriptive statistics and correlations among the grades in the different subjects.

Table 2: Summary of simulated grade distribution (all potential examinees)

Variable	*N*	*Mean*	*SD*	*Minimum*	*Maximum*
MathGrade	10,000	4.66	1.80	0	8
PhysGrade	10,000	4.67	1.81	0	8
ChemGrade	10,000	4.67	1.80	0	8
ArtGrade	10,000	4.67	1.81	0	8

Table 3: Correlation matrix of simulated grades (all potential examinees)

	Maths	*Physics*	*Chemistry*	*Art*
Maths	1.00	0.87	0.87	0.48
Physics		1.00	0.87	0.48
Chemistry			1.00	0.48
Art				1.00

We now define 'subject difficulty' statistically such that all these subjects are *by definition equally difficult* because the grade distributions in each of them are the same for the entire cohort of potential examinees.

Effect of subject choice

We now imagine a situation where each student chooses only two subjects to be examined in, and, for the sake of simplicity, each student

[1] Available online at https://www.gov.uk/government/collections/inter-subject-comparability-research-documents

[2] Cumulative percentage: A* 5.5%, A 15.5%, B 30.2%, C 58.7%, D 77.6%, E 86.7%, F 93.9%, G 98.2%.

chooses their best two subjects (according to the original simulated scores). Tables 4 and 5 show the new descriptive statistics and correlations for the 'observed data'.

Table 4: Summary of simulated grade distribution (after examinees have chosen their two best subjects)

Variable	*N*	*Mean*	*SD*	*Minimum*	*Maximum*
MathGrade	5,016	5.21	1.66	0	8
PhysGrade	4,956	5.19	1.65	0	8
ChemGrade	5,018	5.18	1.67	0	8
ArtGrade	5,010	5.40	1.55	0	8

Table 5: Correlation matrix of simulated grades (after examinees have chosen their two best subjects)

	Maths	*Physics*	*Chemistry*	*Art*
Maths	1.00	0.90	0.90	0.74
Physics		1.00	0.90	0.76
Chemistry			1.00	0.74
Art				1.00

We see from Table 4 that all subjects now appear around half a grade 'easier' (have a higher mean grade) than previously, but that Art is 0.2 of a grade easier than the other three subjects. It is interesting to note from Table 5 that Art is now much more highly correlated with the other subjects (the correlation has risen from 0.48 to 0.75). The effect of subject choice on the grades is easier to see if the six possible subject combinations are considered separately, as in Table 6.

Table 6: Average grades for each combination of subjects

		N	*Mean*	*SD*	*Minimum*	*Maximum*
Maths and Physics	MathGrade	1,659	5.54	1.55	0	8
	PhysGrade	1,659	5.54	1.52	0	8
	ChemGrade	0				
	ArtGrade	0				
Maths and Chemistry	MathGrade	1,672	5.54	1.53	0	8
	PhysGrade	0				
	ChemGrade	1,672	5.54	1.53	0	8
	ArtGrade	0				
Maths and Art	MathGrade	1,656	4.56	1.69	0	8
	PhysGrade	0				
	ChemGrade	0				
	ArtGrade	1,656	5.43	1.56	0	8
Physics and Chemistry	MathGrade	0				
	PhysGrade	1,668	5.52	1.55	0	8
	ChemGrade	1,668	5.52	1.53	0	8
	ArtGrade	0				
Physics and Art	MathGrade	0				
	PhysGrade	1,668	4.52	1.68	0	8
	ChemGrade	0				
	ArtGrade	1,668	5.40	1.55	0	8
Chemistry and Art	MathGrade	0				
	PhysGrade	0				
	ChemGrade	1,677	4.50	1.71	0	8
	ArtGrade	1,677	5.37	1.55	0	8

Table 6 shows that the examinees choosing Art have achieved on average 0.8 to 0.9 of a grade better in Art than in the other subject they chose. The 'subject pairs' method of comparing subjects (e.g., Forrest & Smith, 1972; Coe, 2007) would therefore deem Art to be easier than the other three subjects. A more complex method, used in Scotland to calculate the statistical adjustment needed to align the difficulty of different subjects, is Kelly's method (Kelly, 1976; Coe, 2007). The adjustments represent the amount (in grades) that needs to be added or subtracted from each subject such that examinees on average achieve the same grade in that subject than they do on average in their other subjects.

Table 7: Subject difficulty according to Kelly's method (after examinees have chosen their two best subjects)

	N	*Difficulty*
Maths	5,016	0.211
Physics	4,956	0.216
Chemistry	5,018	0.219
Art	5,010	-0.644

We see that Kelly's method has resulted in Mathematics, Physics and Chemistry being 'harder' and Art being 'easier'. Because this is such a simple scenario we can verify the Kelly result by applying it to Table 6. For example, adding 0.219 to the Chemistry mean and subtracting 0.644 from the Art mean of those taking Chemistry and Art gives approximately equal means of 4.72 and 4.73.

If the difficulty adjustments from Kelly's method were applied, when numeric grades in the two subjects were added together (e.g., to form an index of 'general academic ability' like the University and Colleges Admissions Service (UCAS) points score often used by UK universities as part of the student admission process) a student not taking Art would get a boost of ≈ 0.43, whereas a student taking Art would get a reduction of ≈ -0.43. In other words there would appear to be nearly a grade's worth (0.86) of difference between two students with the same raw points score who differed in whether or not they had taken Art. But of course we know from the simulation that (by definition) all the subjects were equally difficult.

Discussion

In this example, the lower correlation between Art and the other subjects means that there is more 'regression to the mean' – hence for a given score (grade) in Art, the conditional mean score on Mathematics, Physics or Chemistry will be closer to the mean than it would for comparisons of pairs within those three subjects. Because in this simulation examinees are choosing their best subjects, scores on Mathematics, Physics and Chemistry for those examinees for whom Art is one of their top two subjects will be relatively lower (closer to the overall mean) than they are for examinees for whom Art is not one of their top two subjects. Conversely, examinees who are poor at one of Mathematics, Physics and Chemistry are more likely to be poor at the other two than they are to be poor at Art, making Art a more likely best or second best subject. This is illustrated in Figure 1.

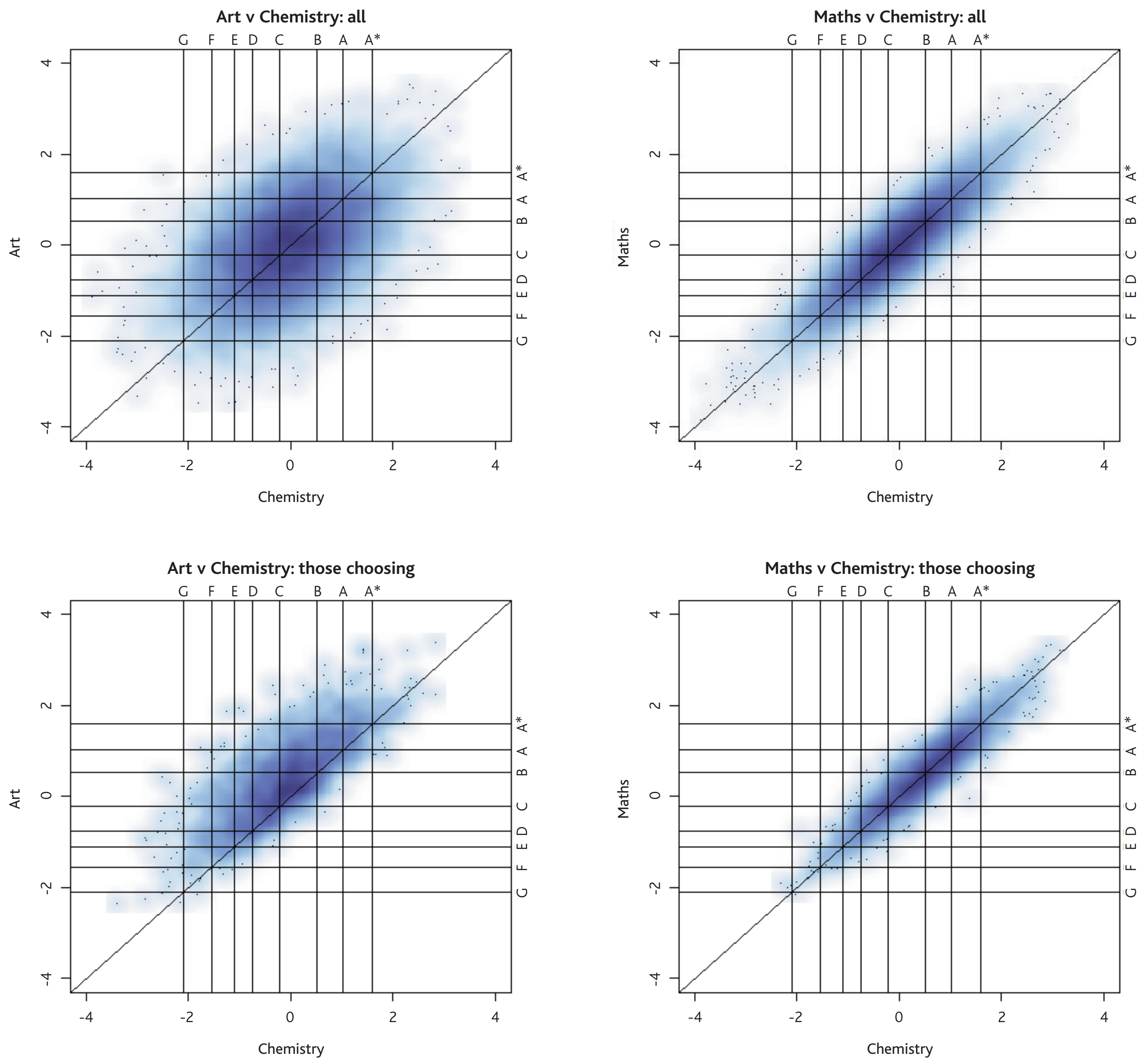


Figure 1: Effect of choice for those choosing Chemistry and Art compared with those choosing Chemistry and Mathematics (axis values are the original normally distributed simulated scores with mean 0 and SD 1)

This example highlights the problem in interpreting the data from all methods of measuring inter-subject comparability – the definition of difficulty is based upon a theoretical notion of every examinee having taken every subject. However, in practice examinees (or perhaps in some cases their schools) choose what subjects to take. We do not know, and can never know, what the results would look like if the entire GCSE or A level aged cohort took all the qualifications. So, while it is true that the same overall ranking of subjects by difficulty appears stable across time and even across jurisdictions (as noted in Ofqual [2015a] p.4), all methods for calculating a single adjustment for difficulty are making the same unjustifiable assumption that the 'missing' data (grades on subjects not taken) resembles the data at hand in the relevant way.

This assumption is brought out especially clearly in approaches that use Item Response Theory (IRT), as used, for example, by Coe (2008) and Ofqual (2015b). Here the different subjects have the role of different items (questions) on a single test, and the examinees have a single 'ability' that is supposed to reflect their probability of achieving a given grade in any particular subject. However, the relationship between differences in difficulty between items in a test on the one hand, and differences in difficulty between different academic subjects on the other, is only analogical (Bramley, 2011). Items within a test are usually selected to measure a construct that is explicitly defined via a specification (syllabus). Different academic subjects within a qualification family, such as GCSE, are not designed to measure any particular overall construct connected with the qualification family, so the construct has to be inferred retrospectively as something like 'general academic ability'. However, it is debatable whether there is any underlying ability that can usefully be said to underpin the wide range of subjects on offer at GCSE and A level.

Furthermore, subjects that are often taken together by large numbers of examinees (e.g., Mathematics and Sciences) are likely to dominate the retrospective definition of the construct. This presents two issues.

First, many minority subjects will correlate less well with mainstream subjects because there are far fewer common skills, content and understanding between them. Secondly, the self-selection effect is different for minority subjects. Students take minority subjects because they have a particular talent, interest or future requirement for them whereas it could be argued that, although they may also take English, Mathematics, Sciences and History for those reasons, they also take them because they are generally and widely considered good subjects for general progression in Higher Education and employment. This weakens the assumption that there will be a strong relationship between performance in different subjects and means that the subset of students taking minority subjects are often very successful in them.

In summary, when examinees choose subjects that measure something different (from mainstream subjects) on the basis of those examinees' strengths in, and preferences for, those subjects, then it is very likely they will appear easier. Psychometricians have cautioned against the dangers of making statistical adjustments to allow for differences in question difficulty in scenarios where choice of questions is allowed within a single examination (e.g., Wang, Wainer, & Thissen, 1995). It is clearly far more problematic to adjust for differences in difficulty at the subject level.

The simulation reported here resembles A level more closely than GCSE since at A level examinees usually choose three subjects (albeit from a much wider range of possibilities). At GCSE, examinees usually choose eight to ten subjects with Mathematics and English taken by virtually all examinees, and a small subset of other subjects taken by large numbers, meaning that these subjects form an effective 'anchor' setting the scale by which the relative difficulty of less popular subjects is determined. But the above conclusion should still hold: less popular subjects that correlate worse with the anchor will appear easier than they really are, if people choose them based on their ability in those subjects.

Of course, the simulation described here greatly oversimplifies the reality. Not only do examinees have a wider choice of subjects, they do not know beforehand which ones they will score best in, and even if they did they might need to take one of their weaker subjects in order to follow their desired future academic or employment path. The simulation could of course be extended to make it resemble more closely the actual situation at GCSE or A level. One sophisticated approach to this would be that of Korobko, Glas, Bosker, and Luyten (2008) who build statistical models allowing for both multidimensionality (of examinee ability) and non-random subject choice. But the purpose of this very simplified simulation was merely to illustrate the point that multidimensionality and non-random choice of subjects can lead statistical methods for measuring differences in subject difficulty towards the wrong answer. Perhaps the question is where the burden of proof should lie – with those who argue for the use of statistical adjustments to align subjects in terms of difficulty (to show that the example in this article is exaggerated or irrelevant); or with those who argue against (to show that the effect demonstrated here is also likely to apply with more realistic data).

References

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

Cambridge Assessment (2016). *A Cambridge Assessment response to Ofqual's subject comparability reports*. Cambridge, UK: Cambridge Assessment.

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

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.

Kelly, A. (1976). A study of the comparability of external examinations in different subjects. *Research in Education*, *16*, 37–63.

Korobko, O. B., Glas, C. A. W., Bosker, R. J., & Luyten, J. W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, *45*(2), 139–157.

Ofqual (2015a). *Inter-subject comparability: a review of the technical literature*. *ISC Working Paper 2*. Coventry: Office of Qualifications and Examinations Regulation.

Ofqual (2015b). *Inter-subject comparability of exam standards in GCSE and A Level*. *ISC Working Paper 3*. Coventry: Office of Qualifications and Examinations Regulation.

Wang, X.-b., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education*, *8*(3), 211–225.

On the impact of aligning the difficulty of GCSE subjects on aggregated measures of pupil and school performance

Tom Benton Research Division

Introduction

It is empirically demonstrated that adjusting aggregated measures of either student or school performance to account for the relative difficulty of General Certificate of Secondary Education (GCSE) subjects makes essentially no difference. For either students or schools, the correlation between unadjusted and adjusted measures of performance exceeds 0.998. This indicates that suggested variations in the difficulty of different GCSE subjects do not cause any serious problems either for school accountability, or for summarising the achievement of students at GCSE.

Data source

The analysis in this article is based upon data from the National Pupil Database (NPD) provided by the Department for Education (DfE). In particular the analysis is based upon the GCSE results[1] of all students in Year 11 in England in the academic year 2014/15. Only full GCSE qualifications taken by at least 50 pupils were included within analysis and only each student's best grade in any given subject was retained. Thus the final data set included over 4.5 million GCSE grades from around 600,000 students across 83 GCSE subjects.

Analysis of the impact of adjustments on mean GCSE scores

One simple way to summarise a student's GCSE achievement is to convert their grades to numbers (A*=8, A=7, B=6, C=5, D=4, E=3, F=2, G=1, U=0) and then to calculate their mean grade across all of the GCSEs they took. These summary measures can be averaged across pupils within a school to provide a simple measure of school performance. This section considers the impact of using a particular statistical method to adjust these summary measures.

For the purposes of this analysis, GCSE scores were adjusted using the Kelly method (see Bramley, 2014, for a brief description). This method has been historically used by the Scottish Qualifications Authority (SQA) to rank the difficulty of different Scottish Highers. In essence this method defines a subject as easy if the candidates taking it tend to achieve higher grades in this subject than in their other subjects. On the basis of this assumption, the method is designed to calculate adjustments to grades so that, across the group of pupils taking a particular subject, their mean grade in that subject will equal the average of the mean grades they achieve in all of their other subjects.

The Kelly adjustments for the 83 subjects included in the analysis are shown in Table 1. This shows, for example, that under the assumptions of the method, the "easiest" subject is GCSE Polish. The Kelly rating for this subject is calculated using the fact that the average grade achieved in Polish is 6.9 compared to an average grade of 4.5 across all other GCSEs taken by the same candidates[2]. For this reason the Kelly method suggests an adjustment of subtracting 2.4 from the grades achieved in Polish which (after adjustments have also been made to all other subjects) makes the means match up.

Of course, the fact that so many minor Modern Language GCSEs are amongst those rated "easy" by Table 1 immediately reveals a weakness with the statistical method. It is suspected that many of the candidates taking these subjects are native speakers. For this reason, their tendency to do better in these GCSEs than in others is not necessarily an indication of the GCSEs being easy at all but rather a result of their particularly strong aptitude for the subject. However, notwithstanding this obvious weakness in the Kelly method, it is still of interest to see the impact of applying the Kelly adjustments to all GCSEs on the overall summary measures of achievement.

The mean GCSE score both before and after applying the Kelly adjustments noted in Table 1 were calculated for each pupil in the data set. Across all students, the correlation between these measures was 0.998[3]. To get a measure of overall school performance, the mean of both of these measures was taken across all pupils within each school. The correlation between the school means for the two measures was 0.999 across all 5,236 schools in the analysis, as well as across the 2,928 schools with at least 100 pupils.

To illustrate these findings further, a random sample of 10 schools with between 100 and 200 pupils was selected. The differences between the adjusted and unadjusted measures are illustrated for these schools in Figure 1. The left hand side of the chart compares the measures at pupil level (restricted to students taking at least five GCSEs) whilst the right hand side compares the measures at school level. A line representing equality between the two measures is included in each chart. Within the data used for these charts the correlations between the measures are 0.998 and 0.999 at pupil and school level respectively[4].

At pupil level, there are no very large differences between the measures. In fact there are only five pupils (out of 1,289) where the difference exceeds 0.4 grades and only one where the difference exceeds 0.5. In these cases, the differences are explained by the fact that all five of

[1] Since this analysis is based upon the initial unamended version of the NPD, the GCSE results included will not account for changes to students' grades made as part of the Enquiries About Results (EARs) process. Also note that GCSEs taken by this group prior to June 2015 (i.e., early entries) were included within the analysis.

[2] Calculations restricted to candidates taking at least two GCSEs.

[3] The same value for the correlation was found when analysis was restricted to pupils who had taken at least five GCSEs.

[4] Thus matching the correlations reported for the national data.

Table 1: Kelly difficulty ratings for 83 GCSE subjects sorted from lowest ("easiest") to highest ratings

Rank	*Subject*	*Number of candidates*	*Kelly Rating*
1	Polish	4,080	-2.38
2	Turkish	1,558	-2.31
3	Portuguese	2,045	-1.87
4	Dutch	396	-1.60
5	Persian	422	-1.52
6	Russian	2,098	-1.28
7	Modern Hebrew	441	-1.08
8	Modern Greek	479	-0.94
9	Art & Design (Photography)	22,080	-0.90
10	Chinese	3,355	-0.82
11	Gujarati	597	-0.75
12	Italian	3,985	-0.74
13	Urdu	4,209	-0.71
14	Art & Design (3D Studies)	2,156	-0.63
15	Arabic	3,167	-0.63
16	Applied Art & Design	874	-0.58
17	Home Economics: Textiles	296	-0.55
18	Art & Design (Textiles)	7,692	-0.55
19	Art & Design	87,940	-0.47
20	Bengali	897	-0.43
21	Art & Design (Fine Art)	51,786	-0.41
22	Japanese	865	-0.40
23	Art & Design (Graphics)	7,440	-0.37
24	Punjabi	794	-0.34
25	English Language & Literature	69,086	-0.33
26	Film Studies	6,971	-0.31
27	D&T Textiles Technology	24,177	-0.31
28	Home Economics: Child Devt	18,096	-0.30
29	D&T Food Technology	38,357	-0.28
30	Expressive Arts & Performance	3,343	-0.27
31	Media/Film/TV Studies	52,715	-0.24
32	Performing Arts	6,256	-0.23
33	Drama & Theatre Studies	71,340	-0.15
34	English Literature	407,758	-0.14
35	PE/Sports Studies	110,846	-0.14
36	Mathematics	549,695	-0.12
37	Social Science: Citizenship	20,792	-0.12
38	Geology	638	-0.10
39	Religious Studies	268,738	-0.09
40	Health & Social Care	7,178	-0.08
41	English Language	307,818	-0.07
42	D&T Resistant Materials	51,017	-0.06
43	ICT	99,160	-0.05
44	Dance	11,982	-0.05
45	Home Economics: Food	8,623	-0.03
46	Physics	123,822	0.01
47	Science (Core)	371,451	0.01
48	Methods in Mathematics	12,438	0.01
49	Chemistry	124,507	0.02
50	Biology	127,778	0.03
51	Electronics	538	0.05
52	Applications of Mathematics	12,179	0.09
53	Additional Science	304,991	0.09
54	Office Technology	13,969	0.12
55	D&T Product Design	37,870	0.12
56	Sociology	21,336	0.13
57	Statistics	51,901	0.14
58	Music	43,519	0.16
59	D&T Electronic Products	7,882	0.16
60	Geography	211,167	0.21
61	D&T Engineering	289	0.23
62	Classical Greek	1,191	0.25
63	Other Classical Languages	506	0.28
64	D&T Graphic Products	31,779	0.28
65	History	228,674	0.28
66	Business Studies: Single	74,023	0.28
67	Latin	8,297	0.31
68	Environmental Science	2,721	0.33
69	Spanish	85,138	0.33
70	D&T Systems & Control	2,976	0.34
71	Ancient History	980	0.40
72	French	150,486	0.50
73	Classical Civilisation	3,937	0.52
74	Psychology	15,961	0.53
75	German	52,677	0.54
76	Economics	9,444	0.56
77	Humanities: Single	8,389	0.57
78	Computer Studies/Computing	32,223	0.59
79	English Studies	720	0.65
80	General Studies	9,341	0.74
81	Applied Engineering	6,358	0.85
82	Astronomy	2,320	1.06
83	Law	2,214	1.19

these pupils took Polish GCSE. As noted earlier, the statistical method used to calculate subject difficulty may be particularly inappropriate for Minor Language GCSEs and so these adjustments may not be valid in any case. However, more importantly, the analysis shows that, even when such subjects are included, the impact of statistically adjusting the difficulty of subjects is almost zero with the ranking of students remaining largely unaffected.

There are two reasons for this: First, it is because the differences between subjects in terms of difficulty are dwarfed by differences in the abilities of pupils across the population. For example, once we account for the number of candidates taking each subject, the standard deviation of the adjustments (Table 1) that will be applied to individual GCSE grades is 0.25. This compares to a standard deviation of 1.6 in the unadjusted mean GCSE scores of pupils. In addition, most pupils take a range of subjects meaning that these adjustments will tend to average out. Secondly, because most pupils will take English, Mathematics and at least one Science GCSE, this ensures some comparability between mean GCSE grades.

On the right hand side of Figure 1 we can see that the rank order of schools is almost entirely preserved regardless of whether adjustments are applied, with the only changes in rank order being amongst schools with extremely similar ratings on both measures[5]. This demonstrates how adjusting for subject difficulty makes essentially no difference to this method of quantifying school performance.

Analysis of the impact of five grade A*-C performance measures

Up to this point this article has only considered measures of performance based on averaging GCSE grades across subjects. However, many school performance measures are based on the percentage of students achieving above some given threshold. Historically, there has been considerable focus upon the percentage of pupils within a school achieving at least five good GCSEs – that is,

[5] In fact, to the naked eye only nine data points are visible on the right hand side of the chart due to the fact that the points for 'School 2' and 'School 10' coincide more or less precisely.

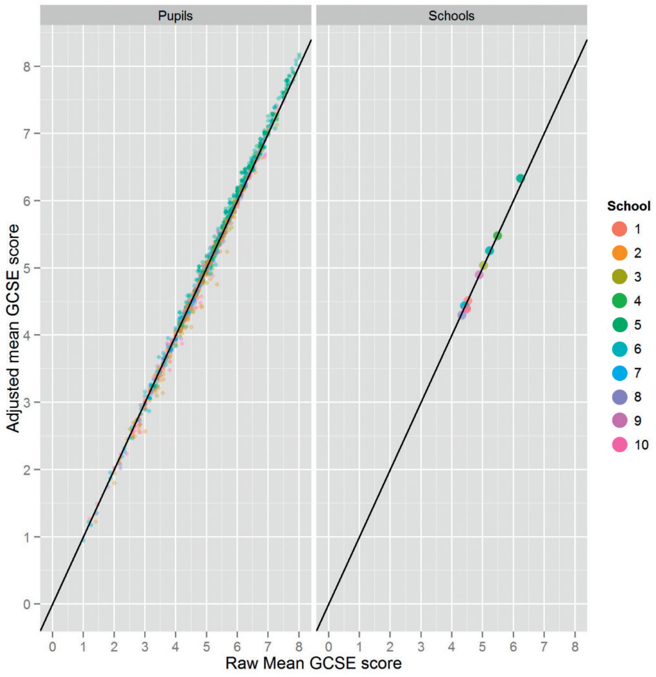


Figure 1: A comparison of adjusted and unadjusted school and pupil GCSE performance measures for a random selection of 10 schools

at grades A*-C[6]. This section estimates what the impact of imposing a statistically defined definition of inter-subject comparability upon GCSEs might be upon this measure.

As the Kelly method only provides adjustments to mean grades, it does not provide an appropriate tool for this analysis. Instead we use an alternative approach: First, we split all pupils into 10 groups (deciles) dependent upon their overall mean GCSE grade. Next, across all subjects combined, we calculate the percentage of GCSEs that are achieved at grade A*-C within each decile. For example, just less than 3% of GCSEs taken by pupils in the lowest decile are awarded grades A*-C compared to 72% for pupils in the fifth decile and 99.96% for the highest decile. Using this information we can predict the percentage of candidates that would be awarded grade A*-C in each subject if the relationship between deciles of achievement and grades awarded was consistent within every subject. This percentage can be compared to the number of candidates that were actually awarded grades A*-C.

Although the NPD does not include a record of the marks achieved by each candidate, it contains sufficient information for us to estimate for each individual pupil the probability that their grade would be awarded at least a grade C if all subjects were adjusted statistically. An example of this is given in Table 2 for GCSE German.

Table 2: Intended percentage of candidates achieving A*-C in GCSE German and cumulative percentage of candidates currently at each grade

Percentage to achieve grade A-C after alignment*	*Percentage of candidates achieving each grade or above*							
	A*	A	B	C	D	E	F	G
85.4	8.0	22.5	45.4	74.2	92.1	97.5	99.3	99.8

Predictions based upon performance deciles suggest that, under this definition of subject comparability, 85.4% of candidates should have been awarded grade A*-C. This compares to 74.2% who were actually awarded these grades. The cumulative percentage of candidates awarded each grade is shown in Table 2. Since only 74.2% of candidates achieved grade C or above, any adjustments to grading would leave these candidates within the grade A*-C band. In contrast, 92.1% of candidates achieved grade D or above so that it is clear no candidates with their current grades below D would be reclassified. However, in order to ensure that 85.4% of candidates achieved grade C or above overall it, would be necessary to reclassify some of those candidates who were awarded grade D to grade C. In fact, the top 62.6% of these candidates[7] should be reclassified. On this basis we can say that:

- all candidates currently awarded grades A*-C in German would have a 100% chance of being awarded grades A*-C after statistical aligning of grading standards across subjects;
- all candidates currently awarded grades E, F, G and U in German would have a 0% chance of being awarded grades A*-C after adjustment; and
- candidates currently awarded grade D would have a probability of 62.6% of being awarded grade A*-C after adjustment.

The above calculations were completed for each GCSE subject. Using the probabilities calculated in this way it was possible to calculate the overall probability that each individual student would achieve at least five A*-C grades[8]. By averaging these probabilities across all pupils within a school, it was then possible to estimate the percentage of pupils that would achieve at least five A*-C grades if statistical alignment of GCSE subjects was implemented. This can be compared to the current percentage that actually achieved five A*-C grades.

Figure 2 shows this comparison for all schools with at least 100 pupils. As can be seen, adjusting grading standards to account for any (supposed)

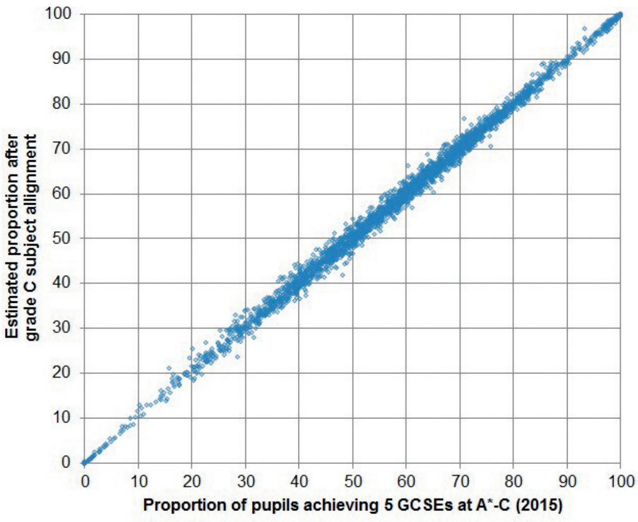


Figure 2: A comparison of adjusted and unadjusted estimates of the percentage of Year 11 pupils in each school who achieve at least 5 A*-C grades at GCSE

[6] More recently the main performance measure has been the percentage of pupils achieving at least five A*-C grades *including* English and Mathematics. However, since this more recent measure places a restriction upon which subjects are used, it is of less interest for a piece of research concerned with the impact of differences in subject difficulty.

[7] Calculated as 100*(85.4–74.2)/(92.1–74.2).

[8] The process for doing this was fairly complicated and is not described in full in this article. Briefly, it required an assumption of independence between a pupil having their grade adjusted in one subject and having their grade adjusted in another. In essence, this implies an assumption that pupils' marks in different subjects are independent given their grades. Although this assumption is unlikely to hold precisely, given that grades capture nearly all of the useful information in marks, it provides a reasonable starting point. Calculations then treat the number of A*-C grades achieved by each candidate as the sum of independent Bernoulli trials which will (by definition) follow a Poisson binomial distribution.

differential subject difficulty make almost no difference to the ranking of schools. Overall there is a correlation of 0.998 between the original percentage of candidates achieving five A*-C grades and the estimated percentage after adjustments. Furthermore, there are only 8 schools (out of 2,928) where the difference exceeds 5 percentage points and none where it exceeds 10 percentage points. This again indicates that adjustments to grading to account for variations in subject difficulty are unlikely to have any substantial effect upon school performance measures.

Reference

Bramley, T. (2014) Multivariate representations of subject difficulty. *Research Matters: A Cambridge Assessment publication*, *18*, 42–47. Available online at: http://www.cambridgeassessment.org.uk/Images/174492-research-matters-18-summer-2014.pdf.

Statistical moderation of school-based assessment in GCSEs

Joanna Williamson Research Division

Introduction

School-based assessment (SBA) such as coursework is included in high-stakes qualifications around the world. In the United Kingdom (UK) for example, selected General Certificate of Secondary Education (GCSE) and General Certificate of Education (GCE) Advanced level (A level) examinations include SBA components[1] alongside examination components[2]. Moderation is required in order to address the question of comparability of SBA marks across different centres. Under current procedures for GCSEs and A levels (see Gill, 2015), moderators re-mark a sample of each centre's SBA work. The awarding body uses the relationship between the moderator mark and centre mark (in the re-marked sample) to decide what adjustment, if any, should be applied to that centre's SBA marks.

Statistical moderation is an alternative form of moderation that calibrates and/or monitors the marks of an assessment on the basis of a statistical relationship with another assessment. Its validity depends on the two assessments having a strong relationship in terms of both assessment content and candidate performance, but they need not measure precisely the same construct. In the context of SBA, the most common statistical moderation practice is to calibrate candidate marks on SBA component(s) using marks from the exam component(s) of the same overall assessment. The motivation for statistical moderation is to preserve information about candidates' SBA performance (such as their ranking within the centre) whilst acknowledging that marking may vary between centres. Statistical moderation removes the absolute meaning of SBA marks, and calibrates them to a new scale that is common to all candidates, that is, the exam component.

During recent reforms of GCSEs and A levels, the Office of Qualifications and Examinations Regulation (Ofqual) proposed the use of statistical moderation in GCSE assessment (Ofqual, 2015a). Previous research by Taylor (2005), using results data from the AQA awarding body, found that statistical moderation generally adjusted marks downward, since SBA marks for GSCE and A level were usually higher than exam marks. The study also found that many candidates would have been awarded different grades under statistical moderation, and that there was a disappointing "absence of any pattern, across different specifications" in terms of statistical moderation outcomes (Taylor, 2005, p.51). The present article outlines methods of statistical moderation that are used in jurisdictions around the world, and explores the effect of applying these methods to results data from three Oxford, Cambridge and RSA Examinations (OCR) GCSEs. This involved statistically moderating all SBA components, aggregating SBA marks with exam marks, and then calculating candidates' statistically moderated final grades from these aggregate scores. Analysis focuses on comparing the statistically moderated results to operational results (moderated under existing, non-statistical procedures) in terms of marks, grades, and the rank-order of candidates and centres.

Methods of statistical moderation

Statistical moderation is a form of assessment linking, where "the goal is to put scores from two or more tests on the same scale – *in some sense*." (Kolen & Brennan, 2004, p.423). Given a suitable pair of assessments (e.g., SBA unit and exam unit), there exist multiple ways to statistically moderate. Table 1 shows the methods investigated in this article: the first four methods are variations of linear scaling, the next two are forms of curvilinear scaling and the final method is rank mapping. Of these, the most commonly used method is linear scaling that matches the mean and standard deviation (SD) of SBA marks within each centre to those of the exam marks (Method 2). The three simplest linear methods (1, 2 and 4) and rank mapping (Method 7) were previously investigated by Taylor (2005). Despite different statistical procedures, many of the methods share common outcomes, as summarised in Table 2.

[1] Recent qualification reforms have reduced the use of SBA in GCSE assessment (Ofqual, 2015b). Of the 23 'new' GCSEs (9–1) ready for first teaching in September 2015 or 2016, 7 contain SBA components.

[2] In GCSE and A level, examination components are always externally set and assessed. They are usually written exams.

Table 1: Methods of statistical moderation

Description	*Moderation formula*	*Examples of use*	*Advantages*	*Criticisms*
1 Adjusts SBA mean to match exam mean	$y_i = \bar{z} + (x_i - \bar{x})$	South Africa	Transparency Few parameters to estimate	Out-of-range marks Potentially unfair when mark distributions skewed
2 Adjusts SBA mean and SD to match exam mean and SD	$y_i = \bar{z} + \frac{\sigma_z}{\sigma_x}(x_i - \bar{x})$	West African Senior School Certificate Western Australia Certificate of Education	Transparency Few parameters to estimate	Out-of-range marks Potentially unfair when mark distributions skewed 'Company you keep' factor unacceptably high
3 Adjusts SBA mean and SD, taking into account both inter- and intra- group differences	$y_i = x_{mean} + \frac{s_y}{\sigma_x}(x_i - \bar{x})$ $+\beta(\bar{z} - z_{mean})$	Hong Kong Diploma of Secondary Education	Allows for global difference in SBA and exam performance	Low transparency
4 Adjusts SBA marks based on regression of exam marks onto SBA marks	$y_i = \bar{z} + \frac{\sigma_z}{\sigma_x}(x_i - \bar{x}) \cdot r$			Moderated marks are 'compressed' about the mean Potentially unfair when mark distributions skewed
5 Quadratic polynomial mapping, fixing max, mean and min SBA marks onto max, mean and min exam marks	$y_i = ax_i^2 + bx_i + c$	New South Wales High School Certificate	Copes with differences in SBA/exam mark distributions High-attaining candidates protected No out-of-range marks	Low transparency Does not preserve ratio between pairs of marks
6 Simplified equipercentile mapping, with linear interpolation	$mod(x_{max}) = z_{max}$, $mod(x_{Q3}) = z_{Q3}$, $mod(x_{Q2}) = z_{Q2}$, $mod(x_{Q1}) = z_{Q1}$, $mod(x_{min}) = z_{min}$	Victorian Certificate of Education (Australia)	Copes with differences in SBA/exam mark distributions Mark intervals somewhat preserved	Vulnerable to effects of individual marks Low transparency Unsuitable for small groups
7 Maps SBA marks to equivalently-ranked exam marks	$mod(x_{rank\ n}) = z_{rank\ n}$			'Company you keep' factor unacceptably high Mark intervals not preserved

- $mod(x)$ is the statistically moderated mark corresponding to raw mark x;
- r is the within-centre correlation coefficient of SBA and exam marks;
- $s_y = \sqrt{w_x\sigma_x^2 + w_z\sigma_z^2}$, where w_x and w_z are weightings such that $w_x + w_z = 1$;
- β is the (pooled) slope after regressing raw SBA marks onto exam marks in a two-level random intercept model;
- x_i, y_i, and z_i are the i^{th} candidate's raw SBA mark, moderated SBA mark and exam mark respectively;
- x_{mean}, $\bar{x}$ and σ_x are the global mean, centre mean and centre SD of raw SBA marks;
- y_{mean}, $\bar{y}$ and σ_y are the global mean, centre mean and centre SD of moderated SBA marks;
- z_{mean}, $\bar{z}$ and σ_z are the global mean, centre mean and centre SD of exam marks; and
- The formulae to calculate coefficients a, b, c (Method 5) are given by MacCann (1996).

Table 2: Statistical moderation outcomes

Aspect of statistical moderation	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Are moderated SBA marks distributed about centres' mean or median exam marks?	Y	Y	N	Y	Y	Y	Y
Do moderated SBA marks have the same mean as the centre's exam marks?	Y	Y	N	Y	Y	N	Y
Is a global difference between SBA and exam marks allowed for?	N	N	Y	N	N	N	N
Can a centre be comparatively 'better at coursework than exams' than other centres?	N	N	N	N	N	N	N
Do moderated marks ever fall out of range?	Y	Y	Y	Y	N	N	N
Is the within-centre rank order of candidates, by SBA mark, preserved?	Y	Y	Y	Y	Y	Y	Y
Are the intervals between candidate SBA marks preserved?	Y	Y	Y	Y	N	N	N
Is the within-centre rank order of candidates, by aggregated mark, preserved?	N	N	N	N	N	N	N
Is the rank order of centres, by mean aggregated mark, preserved?	N	N	N	N	N	N	N

For Methods 1, 2 and 4, there exists a variant form that allows for a global difference between the level of SBA marks and exam marks, as Method 3 does already. The 'allowed difference' variant adjusts marks so that each centre's mean moderated SBA mark differs from its mean exam mark by an 'allowed difference', defined as the difference between the global SBA mark mean and global exam mark mean. To achieve this, occurrences of the centre mean exam mark ($\bar{z}$) in the mark adjustment formulae are replaced by the centre mean exam mark plus allowed difference ($\bar{z} + (x_{mean} - z_{mean})$).

For all methods of statistical moderation, the perceived fairness (and acceptability to stakeholders) is affected by validity, transparency, and assessment context, as suggested by the advantages and criticisms noted in Table 1. The list that follows on page 32 expands upon some particularly important factors:

Anomalous or 'flop' scores

Anomalous scores can distort the mark adjustment deemed necessary for a moderation group. The difficulty is that it is impossible to be sure that a score is 'anomalous' since authentic differences in SBA and exam performance may occur for many reasons. This issue is particularly critical for methods, such as Method 6, that are highly sensitive to individual marks.

Small moderation groups

The smaller the moderation group, the greater the risk of a misleading score distribution which can lead to unfair adjustments to candidate marks. Small moderation groups may necessitate adaptations to statistical procedures and/or manual intervention. As well as increasing cost and complexity, this can harm perceived fairness since different processes are applied to different centres and candidates.

Transparency

It is usually considered important that the statistical procedures leading to moderated marks are transparent to stakeholders. A difficulty is that steps to address other concerns, such as validity, often result in more sophisticated statistical procedures (e.g., Method 3) that are less transparent.

'Company you keep' factor

Under statistical moderation, candidate marks are "inevitably affected" by the performance of others in their moderation group (Wilmut & Tuson, 2005, p.52). The degree to which this occurs is difficult to quantify, but a high degree is perceived as very unfair. Methods 1 and 2 are criticised on the basis that results are too strongly influenced by the moderation group. As an example, Table 3 and Figure 1 show a group of 12 candidates statistically moderated by Method 2. Two cases are shown: (1) where all candidates complete the qualification, and (2) where candidates 1–3 do not complete the qualification. The SBA and exam marks of the other candidates (4–12) remain the same, but their moderated marks differ substantially depending on whether the three lowest-attaining candidates complete the qualification or not.

Disadvantaging particular candidates

There are concerns whenever statistical moderation appears to affect some candidates differently. Substantial changes to the relative intervals between pairs of candidate marks are perceived as unfair, for example, since it is difficult to justify candidates with very similar raw SBA marks receiving very different moderated marks. Truncation of marks (after statistical moderation results in marks out of range) also results in problematic results: loss of discrimination between candidates, marks of zero for valid SBA efforts, and different effective mark reductions for different candidates.

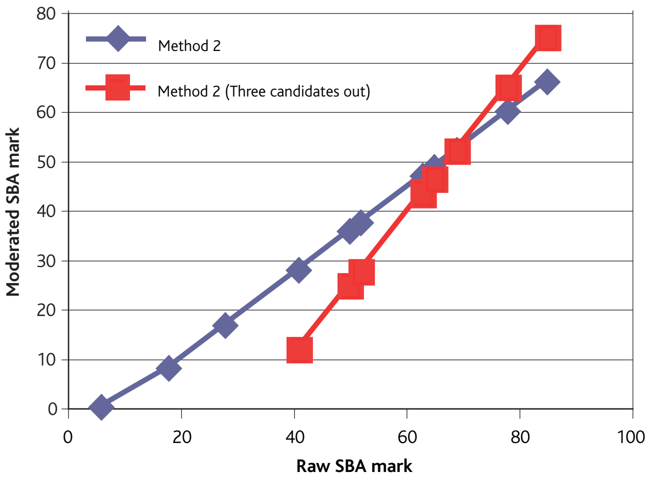


Figure 1: Illustration of 'Company you keep' factor

Another area of concern is large downward mark adjustments, which are perceived as especially unfair for high-attaining candidates in competitive contexts. Stanley, MacCann, Gardner, Reynolds, and Wild (2009, p.54) note that moderation using a linear method "often fails to work satisfactorily" due to negatively skewed SBA mark distributions that result from teachers "setting assessment tasks at which the students can excel" or from overly generous marking applied unevenly across the mark

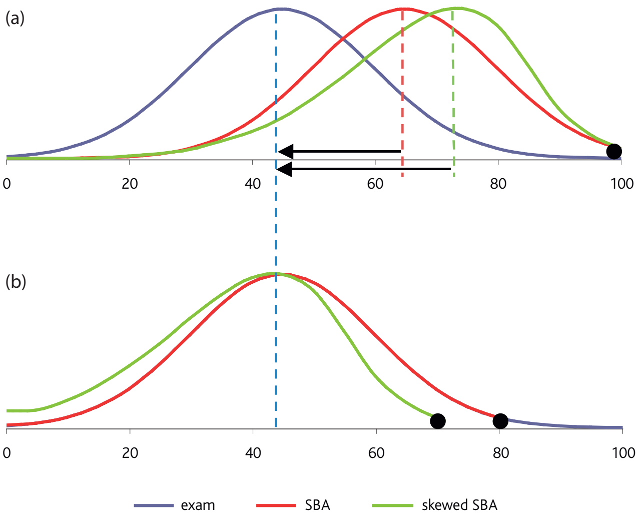


Figure 2: Example mark distributions (a) before moderation and (b) after moderation (dotted lines indicate the mean)

Table 3: Example candidate data, moderated by Method 2

		Candidate No.:												*Mean*	*SD*
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*		
(1)	Raw SBA mark	6	18	28	41	50	52	63	65	65	69	78	85	*52*	*23.2*
	Exam mark	9	28	15	10	32	38	23	70	51	45	65	58	*37*	*20.2*
	Moderated SBA mark	0	8	16	28	36	37	47	49	49	52	60	66		
(2)	Raw SBA mark	-	-	-	41	50	52	63	65	65	69	78	85	*63*	*13.0*
	Exam mark	-	-	-	10	32	38	23	70	51	45	65	58	*44*	*18.7*
	Moderated SBA mark	-	-	-	12	25	28	43	46	46	52	65	75		

range. The SBA mark distribution in these cases will have an 'inflated' centre mean, and the mark adjustment resulting from Methods 1 and 2 will also therefore be inflated (Figure 2a). High-attaining candidates from such a centre will consequently receive lower moderated marks than high-attaining candidates from a centre with a non-skewed distribution (Figure 2b). If the centre's downward mark adjustment was compensating for SBA mark inflation that the higher-attaining candidates within the centre did not benefit from, this appears unfair. A related aspect of perceived fairness is the difficulty in justifying a large mark reduction applied to an 'almost perfect' mark, compared with applying the same reduction to a low- or mid-level mark.

If the standard deviation of a centre's exam marks is lower than the standard deviation of the SBA marks, then Method 2 will also compress SBA marks towards the mean. Where the overall mark adjustment is downward, higher-attaining candidates will therefore receive a larger mark reduction than lower-attaining candidates. This effect is not unique to Method 2, but is mentioned here since it can exacerbate the problem of large mark reductions for high-attaining candidates.

To illustrate the effect of the different methods, Figure 4 and Figure 5, show the effects of statistically moderating an SBA unit for one centre. Twenty-nine candidates took GCSE X at this centre in June 2015. Their raw SBA marks (mean 67.4) and exam marks (mean 39.7) are plotted in Figure 3.

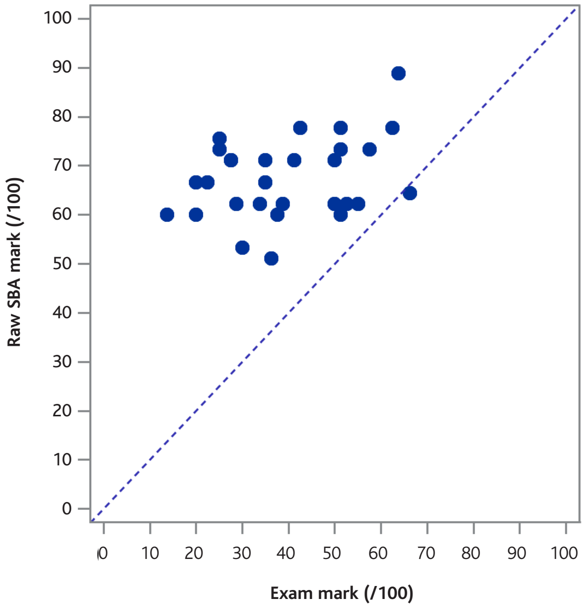


Figure 3: Raw SBA marks against exam marks (r=0.33)

Figure 4 plots the moderated SBA marks resulting from each method of statistical moderation against the candidates' raw SBA marks, and Figure 5 compares the mark distributions resulting from each method. Method 3, the only method not to distribute moderated SBA marks about the mean or median exam mark, is clearly differentiated from the other methods. The highly reduced spread of marks resulting from Method 4 is also very noticeable.

Method

The statistical moderation methods described in Table 1, plus the allowed difference variants, were applied to June 2015 results data from three OCR GCSEs. Centres with fewer than six candidates[3] were excluded, as were centres where SBA and exam marks had zero or negative correlation[4]. For each specification, marks were first converted onto a scale of 0–100, and then all SBA components were statistically moderated by the corresponding exam unit (or a linear combination of the exam units if the specification had multiple). Statistically moderated SBA marks were truncated to the allowed mark range (if outside this), rounded to the nearest whole number, and combined with the exam marks using the weightings implied by Uniform Mark Scale (UMS) allocations. From these aggregated marks, a statistically moderated final grade was calculated for each candidate. New grade boundaries were calculated for each method, such that each statistically moderated grade distribution matched that of June 2015. The present study differs in this respect from that of Taylor (2005), which calculated statistically moderated grades using operational grade boundaries.

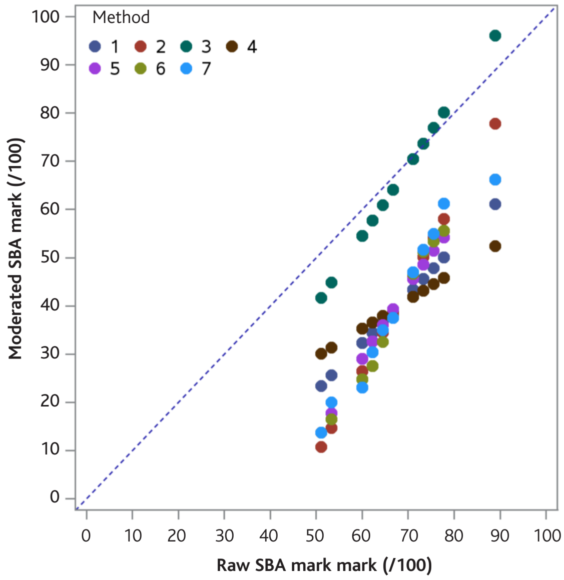


Figure 4: Moderated SBA marks against raw SBA marks

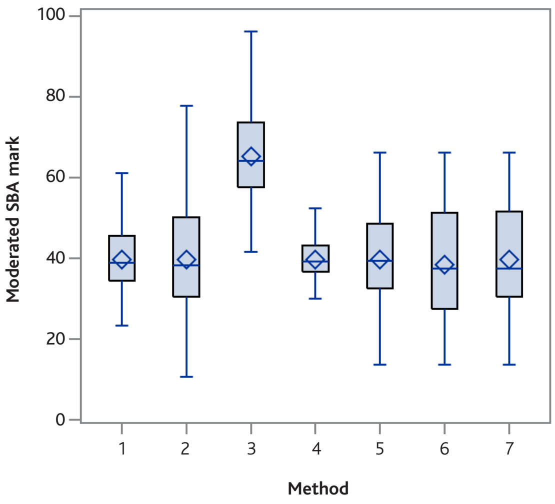


Figure 5: Statistically moderated SBA mark distributions

[3] The smallest definition of acceptable group size found in the literature.

[4] This excluded 1.6% of GCSE X candidates, 3.9% of GCSE Y candidates, and 4.1% of GCSE Z candidates.

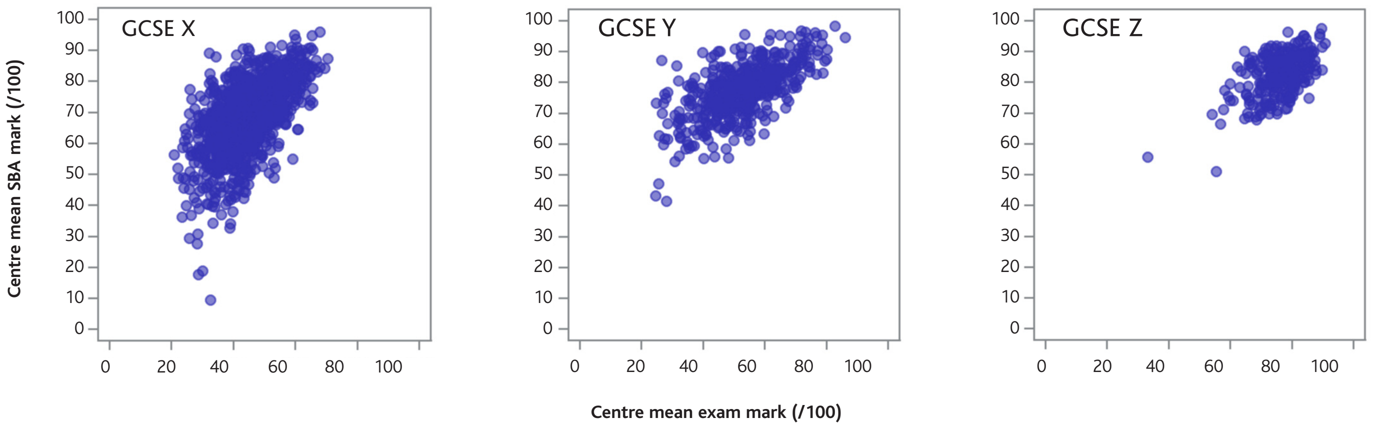


Figure 6: Scatter plots of mean SBA marks against mean exam marks, by specification

Data

The chosen GCSE specifications each had at least one exam unit and at least one SBA unit, had overlap between the assessment objectives of SBA unit(s) and exam unit(s), and were awarded to at least 8,000 candidates in June 2015. Table 4 summarises the characteristics of component units for each specification. Only one SBA unit is shown per specification, since for the specifications with additional SBA units, the characteristics of additional units differed very little from those shown.

Table 4: Summary of component characteristics (0–100 mark scales)

	GCSE X	*GCSE Y*	*GCSE Z*
Difference between raw SBA mean and exam mean	-24.2 marks	-24.7 marks	-6.9 marks
Variability of SBA and exam mark difference[5]	9.6 marks	9.1 marks	6.7 marks
Level of spread in SBA marks compared with exam marks	SD ~3 marks higher	SD ~3 marks lower	SD ~2 marks lower
Shape of mark distributions	Strong negative skew (SBA) vs. negligible skew (exam)	Strong negative skew (SBA) vs. negligible skew (exam)	Both highly negatively skewed
Mean within-centre correlation of SBA and exam marks	r = 0.62	r = 0.58	r = 0.46

Figure 6 compares each centre's mean SBA mark with its mean exam mark. For GCSE X and GCSE Y, the difference between centres' mean SBA mark and mean exam mark was highly variable, particularly for centres with low mean exam marks.

Findings and discussion

Candidate marks and grades

The SBA units of GCSEs X, Y and Z were each statistically moderated ten times, using each method in turn. Figure 7 summarises the resulting mark adjustments for the three SBA units shown in Table 4. For Method 3 and the allowed difference variants, the mean mark adjustment is close to zero. For all other methods, the mean mark adjustment reflects the difference between the mean SBA mark and the mean exam mark, hence a reduction of about 24 marks for GCSE X and GCSE Y, and a reduction of about 7 marks for GCSE Z. The variability of mark adjustments reflects the variability of SBA and exam mark levels, as shown in Figure 6, and therefore is substantially lower for GCSE Z than for the other two specifications.

Across all three specifications, the method resulting in the lowest standard deviation of mark adjustments is Method 3, the Hong Kong linear scaling method. The lower levels of mark and grade changes under

[5] Standard deviation of the difference between centres' mean SBA mark and mean exam mark.

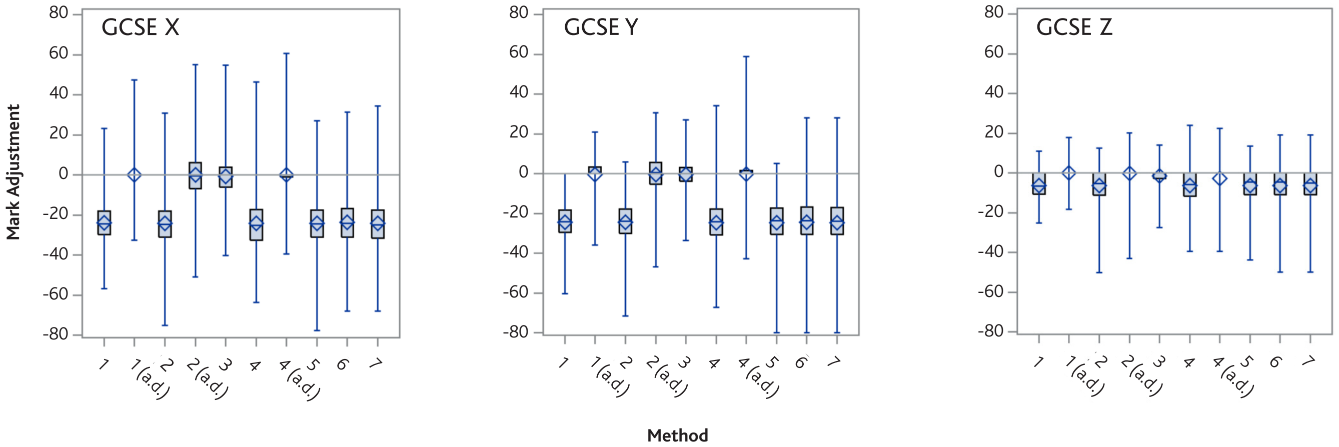


Figure 7: Adjustments to raw marks, by method

Method 3 can be attributed to factors accounted for by the Method 3 moderation formula that other methods in this study do not address. Method 3 does not assume that a centre's mean SBA mark will equal its mean exam mark, and or even the mean plus an allowed difference. Rather, the formula accounts for the reality of regression to the mean and does not 'expect' centres to over-/under-perform equally on different assessments. In addition, the formula adjusts the spread of SBA marks by taking into account the weighted average of spread in the SBA and exam units, so that the spread of moderated marks more closely resembles that of the original SBA marks than under most other methods. These aspects to the moderation formula minimise the overall changes to candidate marks.

The mark adjustments shown here are far larger than the typical mark adjustments made under current moderation practice (see Gill, 2015). As a result of this discrepancy, candidates' statistically moderated marks differed substantially from their operationally moderated marks[6]. Table 5 summarises the differences for the SBA units described in Table 4 and shows that they are almost as large, and variable, as the raw mark adjustments. For some methods (only among Method 3 and allowed difference variants), the mean difference between statistically and operationally moderated marks is positive, indicating that statistical moderation resulted in marks on average *higher* than operationally moderated marks.

Table 5: Differences between statistically moderated and operationally moderated marks

Method	*GCSE X*		*GCSE Y*		*GCSE Z*	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
1	-22.6	10.26	-22.64	9.06	-5.51	6.6
1 (a.d.)	1.36	9.62	1.48	8.42	0.81	5.9
2	-22.77	11.24	-22.67	10.1	-5.51	7.38
2 (a.d.)	1.26	11.15	1.31	9.51	0.79	6.44
3	0.67	9.14	1.5	7.14	-0.65	4.11
4	-22.82	12.69	-22.68	10.24	-5.51	7.08
4 (a.d.)	1.45	11.22	1.54	9.23	-1.75	7.21
5	-22.87	11.73	-22.68	10.74	-5.49	7.71
6	-22.25	11.98	-22.43	10.78	-5.49	7.7
7	-22.82	12.07	-22.68	10.77	-5.51	7.57

Because grade boundaries were recalculated for each statistically moderated mark distribution, a large mean difference between statistically and operationally moderated marks did not itself cause differences between statistically moderated and operational candidate grades. If mark differences had been uniform across centres and candidates, then overall candidate rank orders and consequently grades would have matched those of June 2015 (with lowered grade boundaries). In practice, however, differences between statistically moderated and operationally moderated marks varied substantially across centres and candidates, as already noted. The overall rank orders of candidates after statistical moderation were therefore substantially different to the June 2015 rank orders, leading to differences between statistically moderated and operational grades.

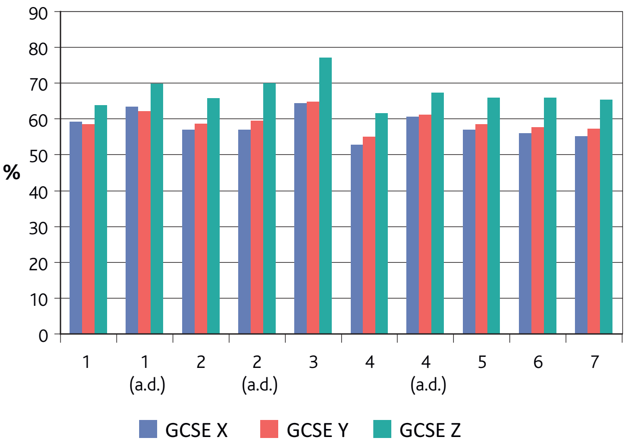


Figure 8: Percentage of candidates awarded the same grade as the June 2015 grade, by method, by GCSE

Figure 8 shows the proportions of candidates whose statistically moderated grade matched their June 2015 grade, for each method and specification. For all three specifications, Method 3 resulted in the highest proportion of candidates retaining their grade, and Method 4 resulted in the lowest proportion of candidates retaining their grade, reflecting the results of Table 5. The majority of statistically moderated grades were within one grade of candidates' June 2015 grade. The distribution of grade differences for Method 3 (Figure 9) shows the typical spread.

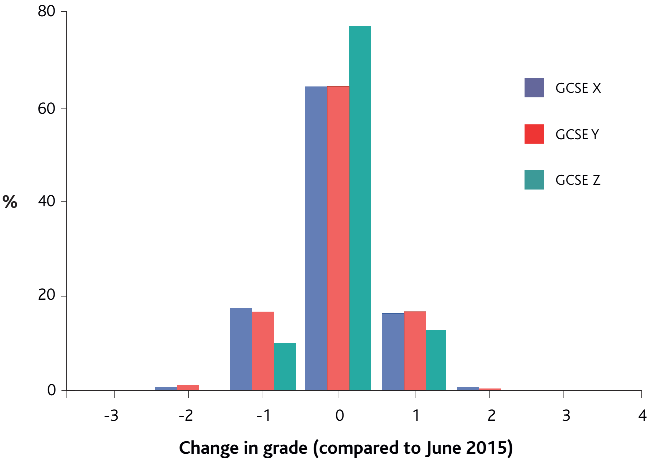


Figure 9: Percentage of candidates per grade change, by specification, for Method 3

A higher proportion of GCSE Z candidates retained their June 2015 grade than in the other two specifications, under all methods. This reflects the lower variability in differences between statistically moderated and operationally moderated marks for GCSE Z (Table 5), which itself reflects the lower variability in SBA and exam mark levels for GCSE Z (Figure 6). It is important that the variability in SBA and exam mark levels for GCSE Z was not only low in absolute terms, but low in relation to the mark width of individual grades. For GCSE Y and GCSE Z, variability in SBA and exam mark levels was higher in relation to the mark width of individual grades, and grade changes were thus more likely to occur.

Rank order

None of the statistical moderation methods altered the within-centre rank ordering of candidates by SBA mark, but all methods changed the rank order by aggregated mark. Statistical moderation also resulted in a

[6] The June 2015 SBA marks after current moderation practices, but before conversion to UMS marks.

different rank order of centres (by mean aggregated mark) compared with the June 2015 rank order. For each of the methods in this study, the final rank order of centres is fundamentally determined by exam performance. Considering moderation formulae alone, the final rank order of centres is *entirely* determined by exam performance. In practice, however, factors beyond the basic formulae, such as rounding to integer marks and truncating scores to the allowed mark range, led to the rank order of centres differing between methods.

Conclusions

This study set out to find methods of statistical moderation used to moderate SBA and to investigate the outcomes of applying these methods to OCR GCSEs. The study identified and explored seven methods: four variations of linear scaling (Methods 1 to 4), two forms of curvilinear scaling (Methods 5 and 6) and finally rank mapping (Method 7).

Statistically moderated marks for the three GCSEs considered were generally lower than both raw marks and operationally moderated marks, in line with Taylor's (2005) findings. In terms of changes to candidate grades, this study agrees that "there were … large numbers of candidates who would change grade" (Taylor, 2005, p.51), even though the present study recalculated grade boundaries in order to preserve overall grade distributions. The high frequency of grade changes reflects high variability in the level of SBA marks compared with exam marks, as illustrated by the scatter plots of Figure 6. For GCSE X and GCSE Y, this variability was particularly large in comparison with the mark widths of grades, and so mark adjustments led to frequent grade changes.

Method 3, the Hong Kong linear scaling method, consistently resulted in lower levels of change to candidate results than other methods, and this is well accounted for by mathematical features of the mark adjustment formula. This formula minimised the overall changes to candidate marks whilst, like the other statistical moderation methods in this study, ensuring that the overall ranking of centres was determined by exam performance rather than SBA performance. It is important to note that the resulting levels of mark and grade changes were still high for both GCSE X and GCSE Y, with results very different from operational results. In terms of appropriateness for GCSE assessment, the level of transparency of Method 3 is also a potential concern, since the moderation procedure uses a more complex formula than the other methods. In Hong Kong, the complete statistical procedures and formulae are published for the public[7], but it is not clear whether a statistical procedure of this complexity would be fully understood by all stakeholders in GCSE assessment.

Under all methods, mark and grade changes for GCSE Z were smaller than those for GCSE X and GCSE Y, and these differences can be linked to clear differences in the original mark distributions: specifically, the SBA and exam mark distributions of GCSE Z had similar shape, and much lower variation in mark levels, than those of GCSE X and GCSE Y. In contrast to Taylor (2005, p.51), who concluded that there was "an absence of any pattern, across different specifications, with respect to the sizes of the adjustments arising from statistical moderation", the present study found that the magnitudes of mark adjustments and levels of grade change appeared to relate fairly directly to the characteristics of the mark distributions of the individual specifications considered.

Overall, the findings support Taylor's conclusion that "the outcomes appear to be very different (at least at candidate level) from those obtained under the current system of moderation by inspection" (2005, p.51). The present study cannot say which of the statistically moderated or operational marks is more 'correct', but clearly demonstrates that the marks resulting from statistical moderation procedures are very different to the marks awarded under current procedures. Careful work would be required in order to explain and justify statistical moderation procedures, if mark adjustments of the level seen in this study were to be accepted. In particular, it would be important for stakeholders to understand that statistically moderated marks carry relative rather than absolute meaning, and in this respect are fundamentally different to moderated marks under current procedures.

References

Gill, T. (2015). The moderation of coursework and controlled assessment: A summary. *Research Matters: A Cambridge Assessment publication*, *19*, 26–31. Available online at: http://www.cambridgeassessment.org.uk/Images/202665-research-matters-19-winter-2015.pdf

Hong Kong Examinations and Assessment Authority. (2010). *Moderation of School-based Assessment Scores in the HKDSE*. Hong Kong: HKEEA. Retrieved from http://www.hkeaa.edu.hk/DocLibrary/Media/Leaflets/HKDSE-SBA-ModerationBooklet_r.pdf

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*. New York: Springer.

MacCann, R. G. (1996). *The Moderation of Higher School Certificate Assessments using a Quadratic Polynomial Transformation: a Technical Paper*. Sydney: New South Wales Board of Studies.

Ofqual. (2015a). *GCSE Computer Science: Consultation on Conditions and Guidance*. Coventry: Ofqual. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/404598/2015-02-07-gcse-computer-science-consultation-on-conditions-and-guidance.pdf

Ofqual. (2015b). *Guidance: Summary of changes to GCSEs from 2015. Coventry: Ofqual*. Retrieved from https://www.gov.uk/government/publications/gcse-changes-a-summary

Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of Teacher Assessment: Evidence of What Works Best and Issues for Development*. London: Qualifications and Curriculum Authority.

Taylor, M. (2005). *Teacher Moderation Systems*. London: Qualifications and Curriculum Authority.

Wilmut, J., & Tuson, J. (2005). *Statistical Moderation of Teacher Assessments*. London: Qualifications and Curriculum Authority.

[7] See Hong Kong Examinations and Assessment Authority (2010)

Good - better - best? Identifying highest performing jurisdictions

Gill Elliott Research Division

Introduction

Jurisdictions which appear at the upper positions of comparative rankings exercises such as PISA (Programme for International Student Assessment) and TIMSS (Trends in International Mathematics and Science Study) are known as high-performing jurisdictions (HPJs). The number of HPJs grows with the number of comparisons in existence, but it is probable that more than 20 jurisdictions might reasonably be given the title, following performance in one or other of the comparisons.

It is becoming increasingly difficult to identify a smaller number of the highest performing jurisdictions, owing to the abundance of comparisons from which to choose. For the purposes of Cambridge Assessment's research into different education systems worldwide, the definition below was proposed to identify the high***est*** performing jurisdictions. In the abbreviation 'HPJ' an asterisk is used to signify 'highest' rather than 'high'-performing jurisdictions; hence H*PJ.

Definition

An H*PJ is identified by its appearance in one of the top 20 positions of at least six of the following seven recent comparisons: TIMSS 2011 8th Grade Science (Martin, Mullis, Foy, & Stanco, 2012); TIMSS 2011 8th Grade Maths (Mullis, Martin, Foy, & Arora, 2012); PIRLS (Progress in International Reading Literacy Study) 2011 Reading (Martin, Mullis, Foy, & Drucker, 2012); PISA 2012 Reading (Organisation for Economic Co-operation and Development [OECD], 2013); PISA 2012 Maths (OECD, 2013); PISA 2012 Science (OECD, 2013); The Global Index of Cognitive Skills and Educational Attainment 2014 (Pearson, 2014).

There are two limitations to this definition; firstly, not all jurisdictions participate in every comparison, so absence from a top 20 position may be due to this fact alone, which might seem unfair. Secondly, if many of the comparisons are influenced by the same overriding factors (e.g., congruence of testing style to jurisdiction educational culture) then it would be expected that the same jurisdictions reappear. Nevertheless, these limitations aside, this seems a reasonable pragmatic approach to obtaining a manageable list of the highest performers.

H*PJs

Application of the definition above resulted in the following list of H*PJs: **Hong Kong, Singapore, Finland, Chinese Taipei, Australia, Japan, South Korea**.

Full details are shown in Figure 1 on page 38.

- No attempt has been made to change jurisdiction names in Figure 1; they are retained in the form in which they appear in each comparison. Thus, 'Korea, Rep.' and 'South Korea' each appear, but are treated as the same jurisdiction.
- In some instances a country is listed in one comparison (e.g., United Kingdom [UK] in PISA (Science) 2012 whilst jurisdictions within that country are listed in another (e.g., England and Northern Ireland in PIRLS (Reading) 2011. In these instances, the count is made separately for each; that is, England receives a count of three, the UK receives a count of two, and Northern Ireland receives a count of one. They are not combined into a single count of five for the UK.
- In some cases the specific order of jurisdictions within a particular comparison will differ from other published sources. This occurs where multiple jurisdictions have equal ranking, so the specific order in which they appear in the figure is determined by other methods. As we are making no attempt to use the specific rankings in this exercise, and are merely counting the number of occurrences of that jurisdiction in the figure, this is immaterial.

References

Martin, M.O., Mullis, I.V.S., Foy, P., & Stanco, G.M. (2012). *TIMSS 2011 International Results in Science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2011/reports/international-results-science.html

Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/timss2011/international-results-mathematics.html

Mullis, I.V.S., Martin, M.O., Foy, P., & Drucker, K.T. *PIRLS 2011 International Results in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/pirls2011/international-results-pirls.html

OECD. (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. PISA, OECD Publishing, Paris. doi: 10.1787/9789264190511-en

Pearson (2014). The Global Index of Cognitive Skills and Educational Attainment 2014. *The Learning Curve*. Retrieved from http://thelearningcurve.pearson.com/index/index-ranking

Figure 1: Ranked positions of jurisdictions in seven recent comparisons

	TIMSS 2011 (8th Grade) Science	TIMSS 2011 (8th Grade) Maths	PIRLS 2011 Reading	PISA (Reading) 2012	PISA (Maths) 2012	PISA (Science) 2012	Pearson Index of Cognitive Skills and Educational Attainment 2014[1]
1	Singapore	Korea, Rep.	Hong Kong	Shanghai – China	Shanghai – China	Shanghai – China	South Korea
2	Chinese Taipei	Singapore	Russian Fed.	Hong Kong – China	Singapore	Hong Kong	Japan
3	Korea, Rep.	Chinese Taipei	Finland	Singapore	Hong Kong – China	Singapore	Singapore
4	Japan	Hong Kong – China	Singapore	Japan	Chinese Taipei	Japan	Hong Kong
5	Finland	Japan	Northern Ireland	Korea	Korea	Finland	Finland
6	Slovenia	Russian Fed.	United States	Finland	Macao – China	Estonia	United Kingdom
7	Russian Fed.	Israel	Denmark	Ireland	Japan	Korea	Canada
8	Hong Kong	Finland	Croatia	Chinese Taipei	Liechtenstein	Vietnam	Netherlands
9	England	United States	Chinese Taipei	Canada	Switzerland	Poland	Ireland
10	United States	England	Ireland, Rep.	Poland	Netherlands	Canada	Poland
11	Hungary	Hungary	England	Estonia	Estonia	Liechtenstein	Denmark
12	Australia	Australia	Canada	Liechtenstein	Finland	Germany	Germany
13	Israel	Slovenia	Netherlands	New Zealand	Canada	Chinese Taipei	Russia
14	Lithuania	Lithuania	Czech Rep.	Australia	Poland	Ireland	United States
15	New Zealand	Italy	Sweden	Netherlands	Belgium	Netherlands	Australia
16	Sweden	New Zealand	Italy	Belgium	Germany	Australia	New Zealand
17	Italy	Kazakhstan	Germany	Switzerland	Vietnam	Macao – China	Israel
18	Ukraine	Sweden	Israel	Macao – China	Austria	New Zealand	Belgium
19	Norway	Ukraine	Portugal	Vietnam	Australia	Switzerland	Czech Rep.
20	Kazakhstan	Norway	Hungary	Germany	= Ireland	= United Kingdom	Switzerland
					= Slovenia	= Slovenia	

⬜ Appears in all seven top 20s: Hong Kong, Singapore, Finland

🟩 Appears in six of the top 20s: Chinese Taipei, Australia, Japan, South Korea

🟦 Appears in five of the top 20s

🟫 Appears in four of the top 20s

🟧 Appears in three of the top 20s

🟨 Appears in two of the top 20s

🟪 Appears in one of the top 20s

= Joint 20th position

[1] The Pearson Index is not entirely independent from all of the other comparisons charted here as it is a 'basket' comparison which draws partly from the PISA, TIMSS and PIRLS scores and partly from literacy and graduation rates.

Research News

Karen Barden Research Division

Conferences and seminars

Tenth Annual UK Rasch User Group meeting

Tom Bramley, Research Division, attended the tenth UK Rasch User Group meeting at Durham University in March. The Group provides a forum for those applying the Rasch model in different fields to get together to share ideas and present research. Tom presented a paper entitled *Rasch – a look under the carpet*.

Educational Collaborative for International Schools (ECIS) Leadership Conference

The ECIS Leadership Conference took place in Rome, Italy in April under the theme Designing on Purpose. Stuart Shaw, Cambridge International Examinations, presented a paper based on work co-authored with his colleagues Helen Imam and Sarah Hughes on *How can your school understand and communicate your bilingual programme?*

International Education Conference (IEC)

Organised by The Clute Institute, the conference was held in Venice, Italy in June. It provided a forum to share proven and innovative methods in teaching at all levels of education and covered a range of topics from accreditation to teaching methods. Jackie Greatorex, Research Division, presented a paper on *Analysing the cognitive demands of reading, writing and listening tests*.

7th Nordic Conference on Cultural and Activity Research

The 7th Nordic Conference took place in Helsingør, Denmark at the University of Copenhagen in June. It was organised in association with the International Society for Cultural-historical Activity Research (ISCAR) and provided a cross-disciplinary forum for researchers and practitioners to share interests in cultural and activity theoretical approaches. Martin Johnson, Research Division, presented a paper entitled *Researching effective feedback in a professional learning context*.

European Conference on Educational Research (ECER)

Held at University College Dublin, Ireland, in August, the ECER Conference provided an opportunity to debate the theme *Leading Education: The Distinct Contributions of Educational Research and Researchers*. Sylvia Vitello presented a paper on *Employers' views on assessment design in vocational qualifications: a preliminary study*. The paper was co-authored with Jackie Greatorex and Jo Ireland, Research Division, and Prerna Carroll, formerly of the Research Division.

European Association for Research on Learning and Instruction (EARLI) – SIG 1: Assessment & Evaluation

In August, Jackie Greatorex and Filio Constantinou, Research Division, attended the EARLI SIG 1 Conference in Munich, Germany. The main theme was *Building bridges between assessment and evaluation*. Jackie presented a paper on *Extending educational taxonomies from general to applied education: can they be used to write and review assessment criteria?* The paper was co-authored with Irenka Suto, Research Division. Filio presented a paper co-authored with Research Division colleagues Victoria Crisp and Martin Johnson entitled *Writing questions for examination papers: a creative process?*

Publications

The following articles have been published since Issue 21 of *Research Matters*:

Darlington, E. and Bowyer, J. (2016). The Mathematics Needs of Higher Education. *Mathematics Today*, *52*(1), 9.

Dunn, K. and Darlington, E. (2016). GCSE Geography teachers' experiences of differentiation in the classroom. *International Research in Geographical and Environmental Education*. Advance online publication available at: http://www.tandfonline.com/doi/full/10.1080/10382046.2016.1207990

Dunn, K. and Darlington, E. (2016). Making resources available to visually impaired students. *Teaching Geography*, *41*(1), 34-36.

Johnson, M. (2016). Feedback effectiveness in professional learning contexts. *Review of Education*, *4*(2), 195–229. Available online at: doi /10.1002/rev3.3061

Johnson, M. (2016). Reading between the lines: exploring methods for analysing professional examiner feedback discourse. *International Journal of Research & Method in Education*. Advance online publication available at: http://www.tandfonline.com/doi/full/10.1080/1743727X.2016.1166484#abstract

Newton, P.E and Shaw, S.D. (2016). Agreements and disagreements over validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 316-318. Available online at: doi:10.1080/0969594X.2016.1158151

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

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

Build your expertise in assessment

Cambridge Assessment Network provides training events, courses and bespoke programmes for assessment professionals and organisations in the UK and internationally.

Our free seminars are a catalyst for topical professional debate, bringing authoritative voices and the wider education community together.

Join us **www.canetwork.org.uk**

the NETWORK
BUILDING EXPERTISE IN ASSESSMENT

Statistics Reports

The Research Division

Examinations generate large volumes of statistical data (approximately 800,000 candidates sit general qualifications each year in the United Kingdom). Our on-going *Statistics Reports Series* provides statistical summaries of various aspects of the English examination system. The objective of the series is to provide statistical information about the system, such as trends in pupil uptake and attainment, qualifications choice, subject combinations and subject provision at school.

In March 2016, we reached a milestone with the publication of the 100th Statistics Report. The reports are part of the Group's commitment to transparency and access to exams data and provide information about the English exam system that can be used by all.

Upon the publication of the 100th report, Tim Oates, CBE, Group Director of Assessment Research and Development, said: "The reports are consistently accessible and clear and, most importantly, available to all. They often reveal important trends and patterns in education and we see them as an important part of our educational mission."

The reports are available in both PDF and Excel format on our website: http://www.cambridgeassessment.org.uk/our-research/all-published-resources/statistical-reports/

The most recent additions to this series are:

- *Statistics Report Series No.101: Uptake and results in the Extended Project Qualification 2008-2015*
- *Statistics Report Series No.102: Provision of level 2 qualifications in English schools 2015*
- *Statistics Report Series No.103: Uptake of level 2 qualifications in English schools 2015*
- *Statistics Report Series No.104: Provision of level 3 qualifications in English schools 2015*
- *Statistics Report Series No.105: Uptake of level 3 qualifications in English schools 2015*
- *Statistics Report Series No.106: Provision of GCSE subjects 2015*
- *Statistics Report Series No.107: Uptake of GCSE subjects 2015.*
- *Statistics Report Series No.108: Provision of GCE A level subjects 2015*
- *Statistics Report Series No.109: Uptake of GCE A level subjects 2015*
- *Statistics Report Series No.110: The re-sitting patterns of a cohort of A level students.*

Introducing Data Bytes

James Keirstead, Tom Sutch and **Nicole Klir** Research Division

Data Bytes is a series of data graphics from Cambridge Assessment's Research Division that is designed to bring the latest trends and research in educational assessment to a wider audience.

High-quality graphics are increasingly used by researchers to communicate complex subject matter both to other researchers and to the general public (Healy & Moody, 2014). This may include the presentation of "raw" data sets, or the results of statistical analyses. However the clear visual communication of quantitative information can be obscured by so-called "chartjunk" (Tufte, 2001). This may be as simple as the use of poorly-chosen fill patterns, or overly dense grid lines that make the visual interpretation of a graphic difficult. But Tufte also warns of graphics "when the overall design purveys Graphical Style rather than quantitative informative" (p. 116). Many "infographics" arguably fall into this latter category. David McCandless (2010) in particular has been criticised for using graphics that "make a simple statement in a way that looks light-hearted and fun. As such, they invite viewers to accept the message superficially, not to explore or contemplate deeply." (Few, 2011). With this caution in mind, we have designed Data Bytes to be informative, accurate and easy to understand.

Each Data Byte consists of a single graphic designed to present a notable data set or research finding relevant to educational assessment. The graphic is accompanied by a brief text explaining what the image shows and why it is significant. Topics for Data Bytes are often chosen to coincide with contemporary news or recent Cambridge Assessment research outputs. Since the series began in October 2015, we have published approximately one graphic per month on topics such as global trends in educational attainment, changing uptake in secondary education subjects, teacher mobility within Europe, and the gender gap in attainment.

One recent example demonstrates the link between achieving an A* grade at A level and a student's likelihood of achieving a First-class university degree. The research was originally published in a peer reviewed journal (Vidal Rodeiro & Zanini, 2015) with the results summarised as a table of odds ratios, a format useful to an academic audience but difficult for the general public to interpret. The corresponding Data Byte presented the same information more intuitively as predicted probabilities, as shown in Figure 1. The graphic illustrates that the number of A* grades a student attained at A level was a strong predictor of their likelihood of achieving a First-class degree at university, and that this relationship was particularly strong for A levels in STEM (Science, Technology, Engineering and Mathematics) subjects. An interactive version of the graphic is available on our website, allowing readers to explore how these probabilities vary by university subject, A level subject, gender, and other factors.

The Data Bytes series can be found at http://www.cambridgeassessment.org.uk/our-research/data-bytes/

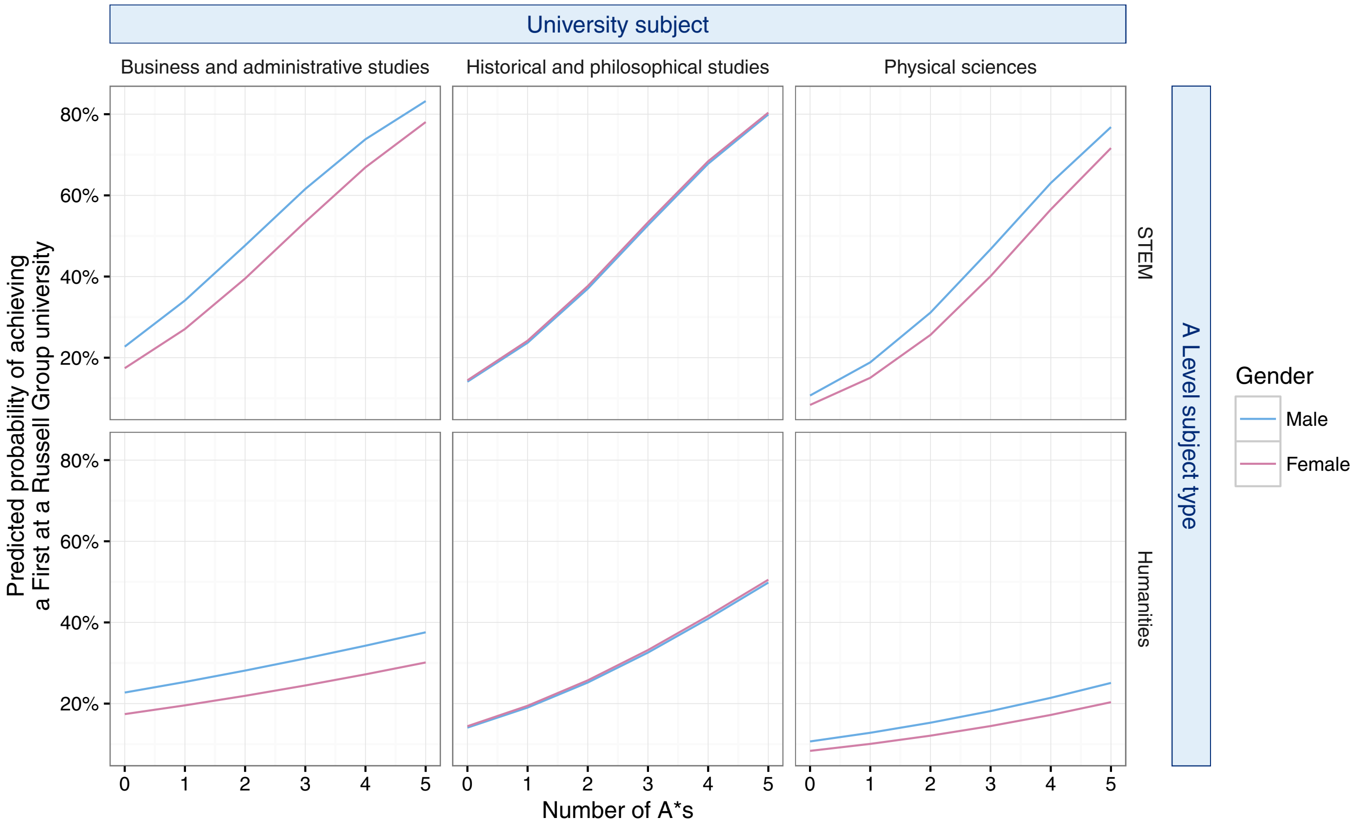


Figure 1: The effect of the A* grade on a student's probability of achieving a First-class degree from a Russell Group university in different subjects

References

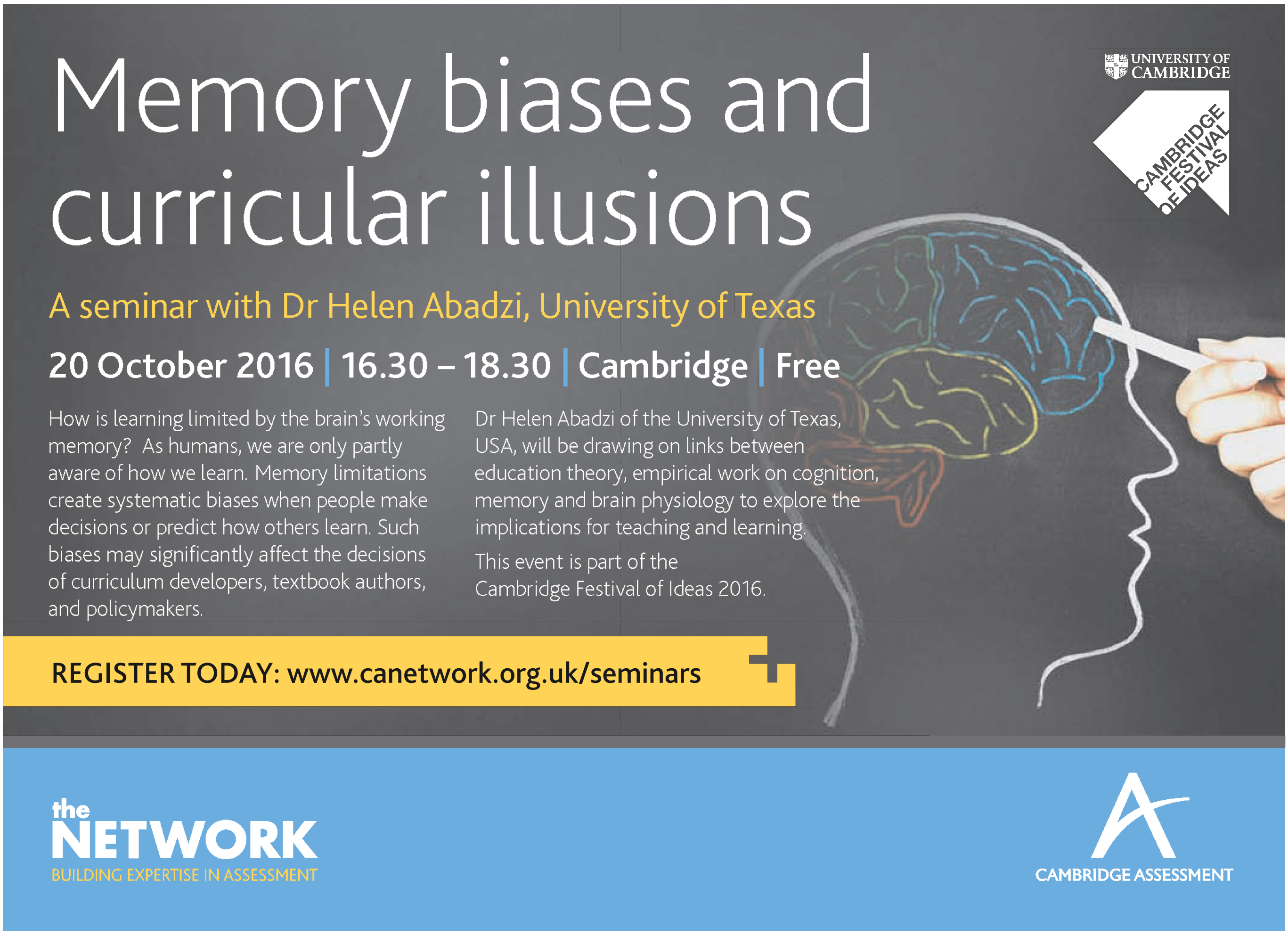
Few, S. (2011). *Visual Business Intelligence – Teradata, David McCandless, and yet another detour for analytics*. Retrieved from http://www.perceptualedge.com/blog/?p=935

Healy, K., & Moody, J. (2014). Data Visualization in Sociology. *Annual Review of Sociology, 40*(1), 105–128. doi.org/10.1146/annurev-soc-071312-145551

McCandless, D. (2010). *Information is Beautiful*. London: Collins.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd edition). Cheshire, Conn: Graphics Press.

Vidal Rodeiro, C., & Zanini, N. (2015). The role of the A* grade at A level as a predictor of university performance in the United Kingdom. *Oxford Review of Education, 41*(5), 647–670. doi: 10.1080/03054985.2015.1090967



Memory biases and curricular illusions

A seminar with Dr Helen Abadzi, University of Texas

20 October 2016 | 16.30 – 18.30 | Cambridge | Free

How is learning limited by the brain's working memory? As humans, we are only partly aware of how we learn. Memory limitations create systematic biases when people make decisions or predict how others learn. Such biases may significantly affect the decisions of curriculum developers, textbook authors, and policymakers.

Dr Helen Abadzi of the University of Texas, USA, will be drawing on links between education theory, empirical work on cognition, memory and brain physiology to explore the implications for teaching and learning.

This event is part of the Cambridge Festival of Ideas 2016.

REGISTER TODAY: www.canetwork.org.uk/seminars

UNIVERSITY OF CAMBRIDGE

CAMBRIDGE FESTIVAL OF IDEAS

the NETWORK
BUILDING EXPERTISE IN ASSESSMENT

CAMBRIDGE ASSESSMENT

CONTENTS : Issue 22 Summer 2016

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU
UK

Tel: +44(0)1223 552666
Email: researchprogrammes@cambridgeassessment.org.uk
www.cambridgeassessment.org.uk

7674774011

© UCLES 2016

ISSN: 1755–6031