CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Research Matters

Issue 36 / Autumn 2023

Proud to be part of the University of Cambridge

Cambridge University Press & Assessment unlocks the potential of millions of people worldwide. Our qualifications, assessments, academic publications and original research spread knowledge, spark enquiry and aid understanding.

Citation

Articles in this publication should be cited using the following example for article 1: Cooke, G. (2023). The prevalence and relevance of Natural History assessments in the school curriculum, 1858–2000: a study of the Assessment Archives. *Research Matters: A Cambridge University Press & Assessment publication, 36*, 6–24. https://doi.org/10.17863/CAM.101742

Credits

Reviewers: Tim Gill, Jo Ireland, Martina Kuvalja, Dominika Majewska and Nicky Rushton

Editorial and production management: Lisa Bowett

Additional proofreading: Alison French

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, please contact our team: Research Division, ResearchDivision@cambridge.org

If you need this document in a different format contact us, telling us your name, email address and requirements and we will respond within 15 working days.

Research Matters Issue 36, https://doi.org/10.17863/CAM.101741

Contents

Foreword

Tim Oates

"...Education will never be the same again, post-pandemic..." is one view. "Back to normal..." is another. Neither position captures the reality of what we experienced and what we now see. For sure, following the wide and accelerated adoption of digital learning during the pandemic we need to understand and implement best practice in hybrid learning. We must understand this massive natural experiment in educational innovation, ensuring that we can regain the progress in closing gaps which we saw prior to COVID. Research and monitoring needs to regain the front foot, not simply seeing in hindsight the impact of what we have done, but providing evidence to better guarantee the quality of what we are doing and plan to do. "Back to normal" denies the fundamental and wide-ranging shifts which COVID response effected. "Everything has changed" undermines the continuity of good science and system improvement which we have seen for the decades preceding COVID. The article on checklists exemplifies this. Checklists were not in use when the flight crew tested the complex Boeing B-17 in 1935, leading to the death of the crew and loss of the aircraft. Modern practice in aviation on checklists began with the analysis of that specific incident and has protected the lives of millions of people. Checklists now are fundamental throughout the aviation industry, in maintenance as well as active flying. But despite their proven value, adoption in other fields has been strangely slow in the past 80 years, with research papers in medicine repeatedly expressing surprise that checklists have not been routinely introduced into areas such as interventional cardiology. COVID may have changed some things, but not the need for expanding and rolling out things which were of demonstrable value prior to the pandemic. The humble checklist, deployed properly, can yet improve the development and practice of assessment, digital or not. Despite COVID, principles of good assessment and high-quality assessment continue to obtain. "Everything has changed" is a voice which discourages something of vital importance – the need for scientific accumulation of knowledge of what works and what does not. The need for this has not diminished at all, nor has it eroded our need to understand exactly how each innovation impacts on both equity and attainment.

Editorial

Tom Bramley

Welcome to the autumn issue of *Research Matters*. With Nadhim Zahawi's announcement in April of a new GCSE in Natural History, to be delivered by OCR in 2025, this is a good time to consider the place of natural history in the school curriculum. In our first article Gillian Cooke traces its role via the record of assessments in natural history and related subjects (e.g., Botany, Zoology, Environmental Science) from the Cambridge University Press & Assessment archives.

Our second article, by Carmen Vidal Rodeiro and Joanna Williamson, explores the effect of the reforms to GCSE Mathematics in England on progression to, and achievement in, post-16 mathematics.

A theme of several *Research Matters* articles in recent years has been the nature of error in assessment materials (and how to avoid it). In our third article Sylvia Vitello, Victoria Crisp and Jo Ireland now report on the practical application of that work in terms of redesigned checklists used in OCR for different professional roles in the question paper production process.

Our fourth article, by Tim Gill, looks at the relationship between the Cambridge Checkpoint tests taken at the end of lower secondary (around age 14) in some international schools, and subsequent performance on the Cambridge IGCSE (taken at around age 16). These kinds of analyses can be used to indicate the predictive value of the earlier qualification and be one factor for schools to consider when deciding whether to adopt the Checkpoint programme.

It has become increasingly clear that the lockdowns imposed in many countries around the world as a response to the COVID-19 pandemic had a particularly harmful effect on young people. School closures and the subsequent partial re-openings required school teachers to cope with the demands of giving lessons to classes where some pupils were physically present while others were attending online. Our fifth article, by Filio Constantinou, presents an analysis of the challenges of this "synchronous hybrid teaching" based on in-depth interviews with 12 teachers from six different European countries.

The prevalence and relevance of Natural History assessments in the school curriculum, 1858–2000: a study of the Assessment Archives

Gillian Cooke (Group Archivist)

Introduction

Concern about our natural environment is at an unprecedented level. It permeates through all levels of media as the effects of climate change and fluctuations of biodiversity manifest throughout the world. There is a thirst for knowledge to understand our environment better and the impact of humans on the natural world. But while the introduction of a new GCSE in Natural History by OCR chimes with our times, natural history itself is not a new qualification, as shown from an archive of over 160 years of qualification documentation from the University of Cambridge Local Examinations Syndicate (UCLES), a predecessor of OCR, which contains a rich resource of natural history type qualifications available to children at all levels and ages.

Natural History is variously defined, but commonly described as "a domain of inquiry involving organisms, including animals, fungi, and plants, in their natural environment, leaning more towards observational than experimental methods of study". Fuelled by curiosity and, in some cases, imperialistic vanity, wealthy explorers of the early modern period sought to dispel images of fantastical creatures and flora from folklore with accurate scientific observation. So began a trend to collect and display natural history findings, and a rise in the popularity of natural history museums, which was at a peak in Britain between the 1880s and 1900s (Rader & Cain, 2015). However, as tastes changed and funding became scarcer, conflict grew between curators, scientists, and conservationists. The educational value of this type of objectification of the natural world, with its safe and organised view of nature, seemed at odds with escalating environmental issues, and the role of natural history education gradually shifted to meet demands for a broader understanding of the laws of nature.

The Cambridge Assessment archive provides a commentary on our social history. The examinations set for school leavers reflect contemporary educational expectations and give us an insight into the aspirations of society and the achievements of students. This repository of primary sources includes question papers, examination regulations, statistics, examiners' reports, and syllabuses (or "schedules") all arranged chronologically. This article attempts to cover the period from the first examinations in 1858 to 2000, two years after the creation of OCR. I have focused on end-of-school examinations offered to 16-year-olds, but have included references to other levels and qualifications to provide context and continuity. The article aims to show the types of qualifications available to candidates and the range of documentation held in the archives. It also looks at the relative success of the subjects over time. UCLES was just one of several examination boards throughout this period. Initially a pioneer of school examinations, when few children attended school beyond primary years, UCLES was one of over 10 boards serving school leavers in England and Wales by the mid 20th century. These were reduced to just four in the 1990s when OCR was created within the UCLES Group specifically for examinations in England. It should be noted that all the information in this article only covers UCLES and OCR.

But what about the definition of natural history? Few subjects can trace a direct line from the 19th to the 21st centuries and natural history is no exception. It takes a circuitous route through several differently named subjects and overlaps with Biology, Geography, Hygiene and Drawing. There are also potential natural history type questions in examinations on Environmental, General, and even Domestic Science. It is difficult to find one common clear definition of natural history but, as the scientific study of animals and plants, there is one consistent emphasis that the study is *observational* rather than *experimental*, and this is the guiding definition used for this article.

This study will form three chronological groups: from 1858 to 1906, 1907 to 1942, and 1943 to 2000. The first and second periods are separated by a major revision in the natural history curriculum in 1906, but throughout both periods the certificate for the qualification was dependent on candidates passing examinations from a range of subjects grouped in different "sections" of broadly similar content, such as languages and sciences. No candidate, therefore, took a single subject qualification in natural history, and subjects such as Botany and Zoology do not consistently appear in one defined section, as is shown in this article. The main qualifications in the first period were the Juniors for under 16s and the Seniors for under 18s, which, after 1917, became the School Certificate and Higher School Certificate. A review of natural history assessment was carried out in 1943, the beginning of my third group of study, and the findings of this review fed into the new single subject national examinations of Ordinary (O) and Advanced (A) level qualifications first taken in 1951. During this period, end-of-school examinations became more widespread, and a national curriculum was first applied to secondary schooling in the 1990s. In 1987 the O level was replaced by the General Certificate of Secondary Education (GCSE) for UK examinations. For each period, I will look at trends and statistics, and natural history coverage in other subjects and at other levels. In the final part I will consider themes and comments from examiners, which may indicate threads common to the whole period from 1858 to 2000.

1858-1906

"The candidate is required to make a careful drawing from memory of a portion of the stem of the white lily, bearing a flower, flower-buds and leaves"

This apparent question for natural history candidates is, in fact, an hour-long question set for Senior candidates in Drawing in the first examinations of 1858.

These first examinations were pioneering, a manifestation of a Victorian ethos of self-improvement, and the first candidates were young male students, eager to acquire a recognisable standard of education before taking up professional work.

While Junior candidates could select Botany or Zoology, the equivalent subjects available to Seniors were called "Botany and Vegetable Physiology" and "Comparative Anatomy and Animal Physiology". Long names for which questions were set on the description and classification of animals and plants.

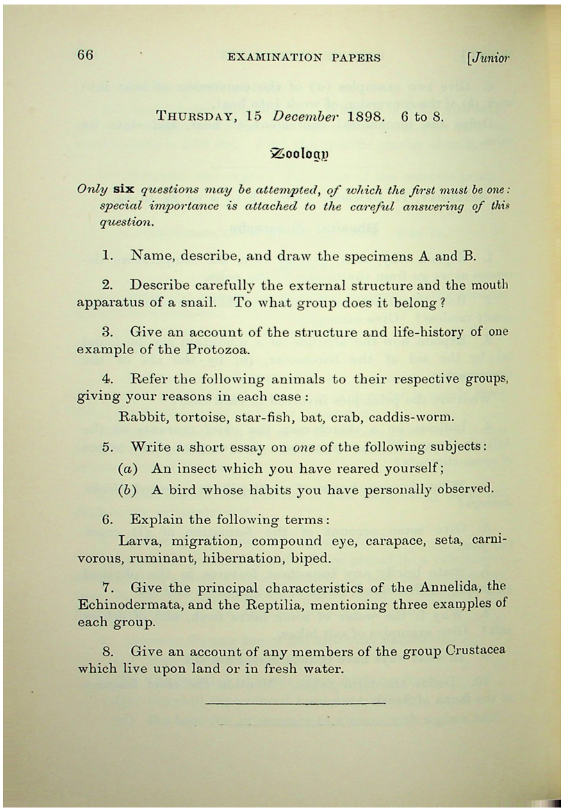
"It is humbling", wrote Linda Few, Biology Subject Officer, "to note that these exams were sat in the same year that Charles Darwin and Alfred Wallace independently proposed the theory of Evolution by Natural Selection" (1858 Question Paper Book, Cambridge University Press, 2008).

Indeed, the subject names and crossover of subjects gives a strong sense that these first examinations, which were independently run by the University, were experimental. There were several drawing examinations and Drawing and Design Drawing candidates alike faced tasks to draw a convolvulus plant, a branch of the woody nightshade, the flowering spike of a foxglove, a human figure in action or a thistle plant – all from memory.

By 1867, Comparative Anatomy had become Zoology, and Vegetable Physiology was renamed as Botany. Geology and Physical Geography was also added, expanding the section to three subjects. Three of the eight Botany questions in this year refer to the "natural order" while the questions in the Geography papers are chiefly concerned with locations and trade.

In 1879, Zoology, Botany and Geology formed a designated "Natural Sciences" section which, by 1895 included Chemistry, Heat, Statistics, Dynamics, Hydrostatics, Electricity and Magnetism, Physical Geography, Physiology and Hygiene. In the early years, Physiology and Hygiene examination questions asked candidates about heart function, breathing, and skin but also about purity of water. By then candidates were required to pass different subjects from up to eight different sections to make up their qualification.

In 1903, the Natural Sciences section was replaced by a section called Biology and Physical Geography. Botany and Zoology were moved to this section and new, additional, sections for Chemistry and Physics took the total number of sections to 16, with between one and six subjects per section. Therefore, as the options for candidates expanded, so did the possibilities for candidates to take sciences without choosing Botany or Zoology.



66 EXAMINATION PAPERS [*Junior*

THURSDAY, 15 *December* 1898. 6 to 8.

Zoology

Only **six** *questions may be attempted, of which the first must be one: special importance is attached to the careful answering of this question.*

1. Name, describe, and draw the specimens A and B.

2. Describe carefully the external structure and the mouth apparatus of a snail. To what group does it belong?

3. Give an account of the structure and life-history of one example of the Protozoa.

4. Refer the following animals to their respective groups, giving your reasons in each case:
Rabbit, tortoise, star-fish, bat, crab, caddis-worm.

5. Write a short essay on *one* of the following subjects:
(*a*) An insect which you have reared yourself;
(*b*) A bird whose habits you have personally observed.

6. Explain the following terms:
Larva, migration, compound eye, carapace, seta, carnivorous, ruminant, hibernation, biped.

7. Give the principal characteristics of the Annelida, the Echinodermata, and the Reptilia, mentioning three examples of each group.

8. Give an account of any members of the group Crustacea which live upon land or in fresh water.

Figure 1: UCLES Junior Zoology Question Papers, December 1898

Trends

Despite the inclusion of so much natural history in the first examination of 1858, the statistics of 1867 (see Table 1) show a low take-up of Zoology and Botany. In that year just 3.3 per cent of the total Junior entry presented themselves for Zoology and 1.4 per cent for Botany. In the Seniors, take-up was even lower and represented just two candidates, both of whom were in West Buckland in Devon.

By 1877, school examinations had become more embedded into school life. The 1870 Education Act introduced compulsory elementary education for all, and end-of-school examinations – although still only applicable to a select few – began to reflect society's educational expectations. Furthermore, the exams were, by then, also available to girls on the same terms as boys and this increased the take-up of Botany.

The 1877 timetable shows that Botany and Zoology were given an evening slot in the week-long mid December exam series. The issue of evening examinations and candidate "overstrain" was taken up in the *Journal of Education* in 1893, which cited the Botany practical as a potential third exam for candidates that day. Excess aside, an examination involving specimen identification from six until eight on a December evening with no electric light would have challenged both candidates and their presiding examiners. The examiners themselves noted that many Botany Juniors "appeared to have had scarcely any practical teaching". As the timetable expanded and the exams session spilled over into a second week, Botany retained an evening slot.

Table 1: Home candidates for Botany and Zoology, 1867–1907 (December exams).

	Juniors						Seniors					
	Total candidates		Zoology/ Nat History		Botany		Total Candidates		Zoology/ Nat History		Botany	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
1867	1196		40		17		273		8		2	
1877	3093	1308	No subject figures given				590	885	No subject figures given			
1887	4986	2519	52	72	69	420	644	1457	6	62	5	257
1897	5378		25	17	33	719	852	1339	2	15	21	287
1907	4400	2703	55	21	57	759	1966	2301	10	10	19	573

Exams were only taken in December up to 1907 when Natural History of Animals replaced Zoology.

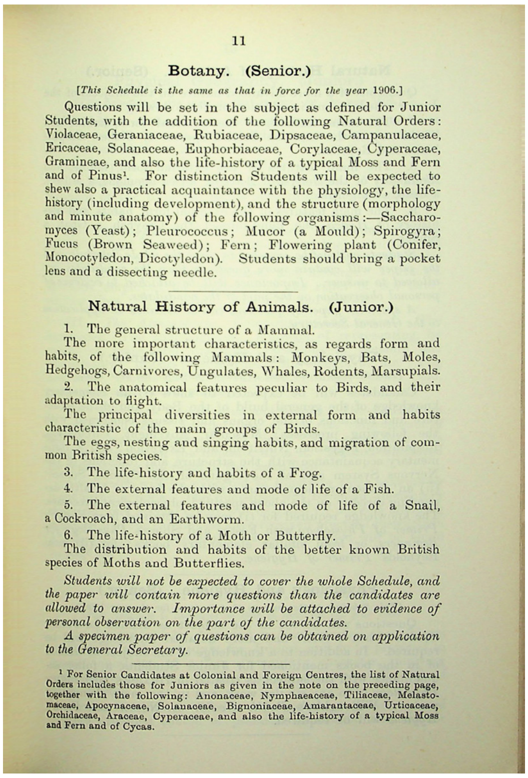
By 1887 the disparity of entries between boys and girls had become marked. Overall performance was first noted by examiners in 1878 when "the work of the girls was decidedly superior to that of the boys" but Botany, as a subject choice, was consistently more popular among girls, with 16.6 per cent choosing the subject in 1887 compared to just 1.3 per cent of boys. It is not clear why this was the case; there was a drawing element to Botany, and drawing was targeted at girls, but there is no discernible focus on medicinal properties of plants, which may have favoured a higher female candidature at that time.

Questions throughout this period are notoriously repetitive with the same themes, and even identical questions, appearing again and again in some subjects. However, this appears to be less common in Botany and particularly Zoology. Although the format of questions is repeated often, the subject matter varies considerably, which would have made these examinations challenging.

Other qualifications

In 1869, the Higher Examinations were introduced for over 18s, predominantly for women training to become teachers, and in 1895, Preliminary Examinations were introduced at the lower end for 14-year-olds. Both qualifications attracted more female than male candidates and included Botany papers. The Higher exams also added Zoology, and there were practical examinations in each. Candidates completing the Higher Examinations of Practical Zoology in June 1897 were asked to dissect a snail and sketch the dissection, while Preliminary candidates of Elementary Botany in December 1896 were asked to botanically describe the edible parts of the potato, strawberry, carrot, almond and plum.

In 1901 a whole new Seniors section was created for Agricultural Science, comprising two compulsory question papers on Agricultural Botany, Chemistry and the Principles of Agriculture. This section germinated into a whole new subject in 1906.



11

Botany. (Senior.)

[*This Schedule is the same as that in force for the year* 1906.]

Questions will be set in the subject as defined for Junior Students, with the addition of the following Natural Orders: Violaceae, Geraniaceae, Rubiaceae, Dipsaceae, Campanulaceae, Ericaceae, Solanaceae, Euphorbiaceae, Corylaceae, Cyperaceae, Gramineae, and also the life-history of a typical Moss and Fern and of Pinus[1]. For distinction Students will be expected to shew also a practical acquaintance with the physiology, the life-history (including development), and the structure (morphology and minute anatomy) of the following organisms:—Saccharomyces (Yeast); Pleurococcus; Mucor (a Mould); Spirogyra; Fucus (Brown Seaweed); Fern; Flowering plant (Conifer, Monocotyledon, Dicotyledon). Students should bring a pocket lens and a dissecting needle.

Natural History of Animals. (Junior.)

1. The general structure of a Mammal.

The more important characteristics, as regards form and habits, of the following Mammals: Monkeys, Bats, Moles, Hedgehogs, Carnivores, Ungulates, Whales, Rodents, Marsupials.

2. The anatomical features peculiar to Birds, and their adaptation to flight.

The principal diversities in external form and habits characteristic of the main groups of Birds.

The eggs, nesting and singing habits, and migration of common British species.

3. The life-history and habits of a Frog.

4. The external features and mode of life of a Fish.

5. The external features and mode of life of a Snail, a Cockroach, and an Earthworm.

6. The life-history of a Moth or Butterfly.

The distribution and habits of the better known British species of Moths and Butterflies.

Students will not be expected to cover the whole Schedule, and the paper will contain more questions than the candidates are allowed to answer. Importance will be attached to evidence of personal observation on the part of the candidates.

A specimen paper of questions can be obtained on application to the General Secretary.

[1] For Senior Candidates at Colonial and Foreign Centres, the list of Natural Orders includes those for Juniors as given in the note on the preceding page, together with the following: Anonaceae, Nymphaeaceae, Tiliaceae, Melastomaceae, Apocynaceae, Solanaceae, Bignoniaceae, Amarantaceae, Urticaceae, Orchidaceae, Araceae, Cyperaceae, and also the life-history of a typical Moss and Fern and of Cycas.

Figure 2: UCLES Botany and Natural History of Animals Schedules, 1907

1907-1942

During the early 20th century, the science options available to candidates continued to expand. Along with Agricultural Science in 1906 (see Table 4), there was Hygiene, Physical Geography, Chemistry, Physics, Heat and Experimental Science and, in this year, Zoology was replaced by the Natural History of Animals. The new syllabuses coincided with the introduction of a summer examination session, which gradually became the standard examination session for "home" candidates (in England).

The new regulations for the Natural History of Animals attempted to address the size of the subject. It was stressed that students would "not be expected to cover the whole schedule" and were encouraged to use common English terms, rather than Latin, whose unfamiliarity would add to the candidate's workload and was "apt to give a distaste for the subject". Botany teachers were asked to "keep in mind the importance of naked-eye work and of experiments performed by the students themselves"; they were encouraged to arrange "excursions into the country … to enable students to observe the plants of different classes of habitats", as "special weight" would be given to candidates' descriptions of specimens.

In 1917, the new Board of Education introduced a national system of school leaving examinations. The School Certificate replaced the Senior Local Exams, and the Higher School Certificate was introduced for under 18s. Despite this change, the schedules for Botany and the Natural History of Animals remained the same as those introduced in 1906 and these subjects were part of a four-subject section with Physiology and Hygiene, and Physical Geography, for which candidates needed to pass just one subject to pass the section. By 1921 the School Certificate had been arranged into four groups, the third of which comprised science subjects. The accessibility of Botany and the Natural History of Animals was compromised by a requirement that candidates could only take these subjects at centres with properly equipped laboratories for which the centre may impose an extra fee on the candidate.

Botany examinations "for Colonial centres" appeared in 1898 and, from 1910, there were Botany and Natural History schedules for "centres in Tropical Countries" with references to arborescent plants and seasonal changes. The introduction of national exams allowed the traditional Cambridge Senior and Junior examinations to become more tailored to the needs of overseas candidates, and schedules were extended to include "oversea alternative" questions for Botany candidates, which in 1937 included questions on bamboo and jack-fruit trees.

In 1931, with entries in continual decline, Natural History was examined for the final time and this marked the end of a "Natural History" titled examination. By then the paper was still a Junior examination, which was predominantly taken by overseas candidates. Thereafter, Zoology existed only as a section within General Science Paper II, which also included sections on Botany, Soil Science, Domestic Science and Physiology and Hygiene, until that, too, disappeared in 1935.

Trends

While more girls than boys continued to choose Botany at both Junior and Senior levels, the entry levels for Zoology or Natural History of Animals (which replaced Zoology in 1907) remained comparatively low but consistent between the sexes (see Tables 2 and 3).

Table 2: UK and overseas candidates for Botany and Natural History of Animals between 1910 and 1918.

	Juniors			Seniors		
Year	Total	Botany	Natural History of Animals	Total	Botany	Natural History of Animals
1910	9030	1219	51	8182	1587	23
1912	9199	1107	35	8157	1610	22
1914	10 187	1483	19	9506	1791	53
1916	9009	1417	57	10 235	2495	45
1918	9177	1555	53	8869	2588	43

Table 3: Junior candidates for Botany and Natural History of Animals, 1910.

	July		December	
	Boys	Girls	Boys	Girls
Botany	35	651	70	463
Natural History of Animals	11	8	25	7

Table 4: Candidates for Agricultural Science (and Rural Science), 1907-1997.

Agricultural Science (Home candidates)				% passes at grade C and above	
Qualification	Year	Boys	Girls	Boys	Girls
Seniors	1907	46	0		unknown
Seniors	1917	9	0		unknown
Seniors	1927	37	2		unknown
None	1937				
None	1947				
O Level	1957	49	6	55.1	83.3
O Level	1967	148	10	47.3	80
O Level	1977	136	34	45.6	44.1
O Level	1987	109	28	61.5	46.4
GCSE Rural Science	1997	863	653	19.7	28.1

After a surge in candidate entries in 1914, possibly in reaction to the uncertainty of war, candidate numbers for all subjects dropped, rising again after the introduction of national qualifications which precipitated a shift of UK candidates towards June, rather than December exams. In the last examinations for the Natural History of Animals in December 1931 just one boy and three girls entered at School Certificate level from the UK.

The inclusion of sections on Botany and Zoology in a single General Science Paper in the 1930s makes it harder to identify candidates who chose these subjects as there is no evidence to show that General Science candidates would have answered any natural history questions at all, and although Botany was retained as a separate science subject, Table 5 shows the decline in its popularity.

Table 5: School Certificate candidates (UK) for Botany, Natural History and General Science, 1920-1945

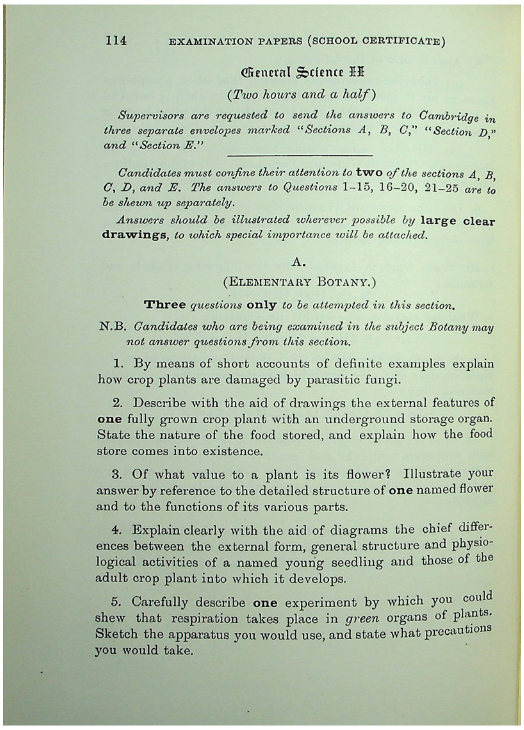
	Total	Botany		Natural History of Animals		General Science		Total % of candidates
		Boys	Girls	Boys	Girls	Boys	Girls	
1920	7785	40	2521	36	22			33.6%
1925	9476	88	2240	39	25			25.2%
1930	9795	169	2353	33	6	135	84	28.3%
1935	9208	113	1536			260	338	24.4%
1940	9569	69	757			895	794	26.2%
1945	12 270	32	405			1600	1222	26.5%

Other qualifications

After the withdrawal of the Higher Examinations in 1923, two new qualifications were introduced for aspiring teachers, but at very different levels. The Rural Pupil Teachers qualification targeted pupils who had completed primary education and could support the education of their younger peers in rural areas. It ran from 1929 to 1936 and included a Natural Science question paper of compulsory sections on Zoology and Botany based on "recognition and familiarity with main crops, trees, flowers ... and the habitats of the common wild life of the countryside" (Rural Pupil Teacher Natural Science Regulations, 1929). A more formal higher grade teaching qualification was set up in collaboration with Homerton College in 1927 and the Homerton College Final Examinations ran until 1951. These included a paper on Gardening with questions on soil cultivation, the cultivation of plants, and insect and fungoid pests.

The 1930s also saw a revival in natural history themed drawing exams with a two-hour question paper entitled Nature Drawing which was placed in the Art Section of the School Certificate. This became Painting or Drawing from Plant Life in 1946. Geography Papers continued to include questions on climate, temperature, and rainfall. Agricultural Science, which was introduced during this period and aimed at boys, was in two parts: part 1 was concerned with the application of chemistry and physics to crop growing, and part 2 was concerned with the biology of farm crops and weeds.

Physiology and Hygiene (later, just Hygiene) was pitched at girls and became a more practical test of candidates' knowledge of first aid and self-care, with questions on air quality and dental hygiene; skills that would have been highly valued before the establishment of the National Health Service.


114 EXAMINATION PAPERS (SCHOOL CERTIFICATE)

General Science II

(*Two hours and a half*)

Supervisors are requested to send the answers to Cambridge in three separate envelopes marked "Sections A, B, C," "Section D," and "Section E."

Candidates must confine their attention to **two** *of the sections A, B, C, D, and E. The answers to Questions* 1–15, 16–20, 21–25 *are to be shewn up separately.*

Answers should be illustrated wherever possible by **large clear drawings,** *to which special importance will be attached.*

A.

(ELEMENTARY BOTANY.)

Three *questions* **only** *to be attempted in this section.*

N.B. *Candidates who are being examined in the subject Botany may not answer questions from this section.*

1. By means of short accounts of definite examples explain how crop plants are damaged by parasitic fungi.

2. Describe with the aid of drawings the external features of **one** fully grown crop plant with an underground storage organ. State the nature of the food stored, and explain how the food store comes into existence.

3. Of what value to a plant is its flower? Illustrate your answer by reference to the detailed structure of **one** named flower and to the functions of its various parts.

4. Explain clearly with the aid of diagrams the chief differences between the external form, general structure and physiological activities of a named young seedling and those of the adult crop plant into which it develops.

5. Carefully describe **one** experiment by which you could shew that respiration takes place in *green* organs of plants. Sketch the apparatus you would use, and state what precautions you would take.

Figure 3: UCLES School Certificate General Science II Question Paper, June 1933

1943-2000

In 1943 a Joint Committee drafted a report containing recommendations for sixth form study covering Biology, Botany and Zoology to address concerns that the syllabuses in these subjects were outdated and "too much concerned with questions which seemed important a generation ago". It included wide representation from three examination boards: Cambridge (UCLES), Oxford (UODLE) and Oxford and Cambridge (OCSEB), plus the university faculties of Biology and Medicine, College Scholarship Groups, and male and female science teaching associations.

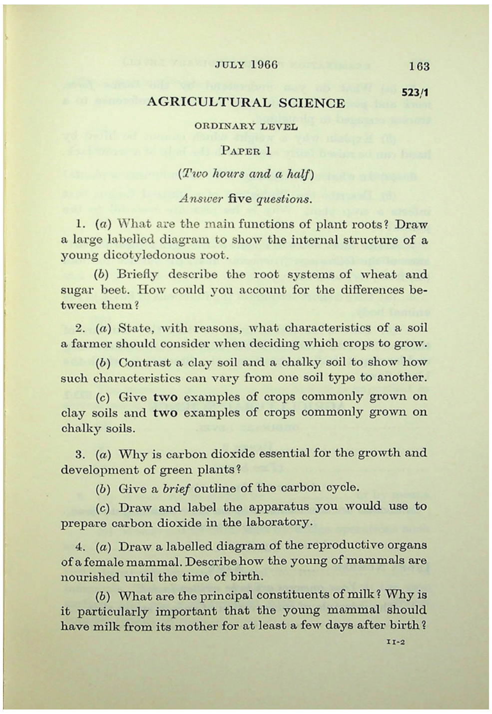
The committee took up a concern raised by examiners in 1925 that the size and scope of natural history could prevent detailed study and "impede the development of free observational skills". It therefore advised that "the student should be encouraged to make a really thorough study of a restricted part" of the syllabus to "train his scientific judgement in the best possible way". The recommendations led to new syllabuses for the Higher School Certificate which remained largely unchanged over the following decade.

During the Second World War subject options reduced and topics within the subjects became more generic. In 1947 a revision of the Science syllabus attempted to address their "increasing popularity" but this did not affect the Botany School Certificate papers. These remained unchanged and separate from the seven General Science question papers and included Biology and Geology but not Zoology.

With the introduction of the first single subject qualifications in 1951, candidates could specialise in specific subjects without having to select a subject group to make up their certificate qualification. In the first year Botany was available to Ordinary (O) level candidates for 16-year-olds but from 1952 the lowest qualification for both Botany and Zoology was the Alternative Ordinary level (AO), which was aimed at sixth formers. These qualifications were available to UK and overseas candidates, but, as earlier, they were restricted to candidates at schools with suitable equipment for practical work.

Agricultural Science survived to become an O level subject, and in 1970 Environmental Science replaced O level General Science within the Combined Sciences section. According to the new syllabus, its principal idea was "the relationship of Man to his environment". It was revised in 1986 when it was offered to UK candidates only, and the course was then described as a natural science course based on "experimentation, observation and logical deduction".

In 1987 the General Certificate of Secondary Education (GCSE) was introduced for UK candidates as a replacement for O level, aimed at a wider cohort. Agricultural Science, however, remained as an O level qualification (presumably to meet an overseas candidature) and, in 1989 Rural Science was introduced at UK GCSE level. Described as "an applied science" with expectation that "full use will be made of the plants, animals and materials found locally", this qualification ran to the millennium.


JULY 1966 163

523/1

AGRICULTURAL SCIENCE

ORDINARY LEVEL

PAPER 1

(*Two hours and a half*)

Answer **five** *questions.*

1. (*a*) What are the main functions of plant roots? Draw a large labelled diagram to show the internal structure of a young dicotyledonous root.

(*b*) Briefly describe the root systems of wheat and sugar beet. How could you account for the differences between them?

2. (*a*) State, with reasons, what characteristics of a soil a farmer should consider when deciding which crops to grow.

(*b*) Contrast a clay soil and a chalky soil to show how such characteristics can vary from one soil type to another.

(*c*) Give **two** examples of crops commonly grown on clay soils and **two** examples of crops commonly grown on chalky soils.

3. (*a*) Why is carbon dioxide essential for the growth and development of green plants?

(*b*) Give a *brief* outline of the carbon cycle.

(*c*) Draw and label the apparatus you would use to prepare carbon dioxide in the laboratory.

4. (*a*) Draw a labelled diagram of the reproductive organs of a female mammal. Describe how the young of mammals are nourished until the time of birth.

(*b*) What are the principal constituents of milk? Why is it particularly important that the young mammal should have milk from its mother for at least a few days after birth?

II-2

Figure 4: UCLES O Level Agricultural Science Question Paper, June 1966

Trends

More data is available for this later period, and although it is potentially confusing to juxtapose different subjects at different levels, Tables 6 and 7 do indicate trends that Zoology, Botany and later Environmental Science trail behind other subjects in both entries and results.

Table 6: Sixth form AO level candidates for Botany, Zoology and Geology, 1955-1970.

		Botany			Zoology		
	Total candidates	No. of candidates	No. of passes	% passes	No. of candidates	No. of passes	% passes
1955	4976	38	26	68.4	9	8	88.9
1960	8590*	62	43	69.4	0		
					Geology		
1965	13 883	60	29	48.3	522	341	65.3
1970	16 120	56	37	66.1	911	552	60.6

*From 1960 candidates counted twice if entered for A and O level.

Table 7: GCE candidates for natural history type examinations (UK summer, 1975–1985).

	O levels				AO levels			
	Environmental Science		Agricultural Science		Botany		Geology	
	No. of candidates	% passes	No. of candidates	% passes	No. of candidates	No. of passes	No. of candidates	% passes
1975	665	41.7	165	46.7	6	2	118	62.4
1980	226	41.2	155	40.6	-	-	722	57.9
1985	18	22.2	130	58.5	-	-	416	61.5

Although this period sees the loss of Botany and Zoology altogether from UCLES school leaving examinations, the subjects continued to be available to sixth formers for a little longer. The suggestion in the 1943 Joint Committee report was that the subjects were too complex and broad for assessment at 16, and the requirement for specialist equipment and resources placed the subjects out of reach of many students. By combining aspects of Botany and Zoology into Rural Science (Tables 8 and 9), however, the potential remit of this new qualification became even bigger, as a remark by the 1991 examiners indicates: "It was noticeable that many candidates displayed a good knowledge of either plants or animals but not both."

Table 8: Candidates for Rural Science, 1990-2000.

	No. of entries	% of OCR GCSE syllabus entries	% of passes at grade C and above
June 1990	1680	0.10	23.6
June 1995	1602	0.11	28.5
June 2000	1028	0.08	23.95

Other qualifications

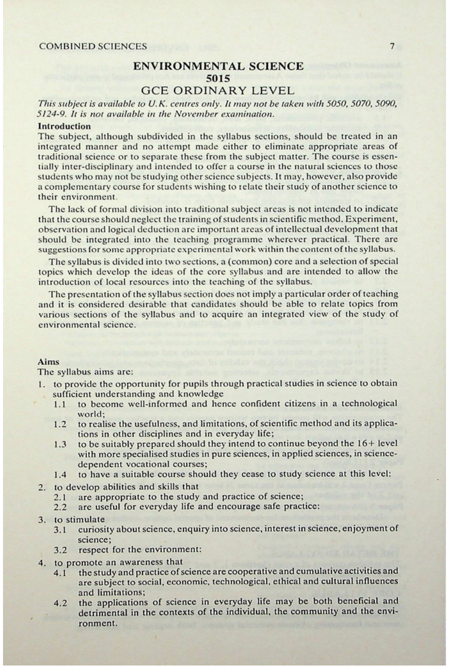
As well as being available to sixth formers, Botany and Zoology were also retained as A level subjects, and the entries for A level give an indication of the respective popularity of Biology, Botany and Zoology.

Table 9: A level entries for Biology, Botany and Zoology 1970-1990.

	Total no. of entries		No. of entries for Biology		No. of entries for Botany		No. of entries for Zoology	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
1970*	6613	5519	1123	1409	121	111	285	237
1980	13 903	13 844	1855	2362	9	15	59	72
1990**	38 712		3717		18		54	

*The 1970 annual report notes that A Level Home Centre entries for Biology rose by 9 per cent on the previous year.

**The published examination statistics for UCLES for 1990 do not give a breakdown of candidates by sex.



COMBINED SCIENCES 7

ENVIRONMENTAL SCIENCE
5015
GCE ORDINARY LEVEL

This subject is available to U.K. centres only. It may not be taken with 5050, 5070, 5090, 5124-9. It is not available in the November examination.

Introduction

The subject, although subdivided in the syllabus sections, should be treated in an integrated manner and no attempt made either to eliminate appropriate areas of traditional science or to separate these from the subject matter. The course is essentially inter-disciplinary and intended to offer a course in the natural sciences to those students who may not be studying other science subjects. It may, however, also provide a complementary course for students wishing to relate their study of another science to their environment.

The lack of formal division into traditional subject areas is not intended to indicate that the course should neglect the training of students in scientific method. Experiment, observation and logical deduction are important areas of intellectual development that should be integrated into the teaching programme wherever practical. There are suggestions for some appropriate experimental work within the content of the syllabus.

The syllabus is divided into two sections, a (common) core and a selection of special topics which develop the ideas of the core syllabus and are intended to allow the introduction of local resources into the teaching of the syllabus.

The presentation of the syllabus section does not imply a particular order of teaching and it is considered desirable that candidates should be able to relate topics from various sections of the syllabus and to acquire an integrated view of the study of environmental science.

Aims

The syllabus aims are:

1. to provide the opportunity for pupils through practical studies in science to obtain sufficient understanding and knowledge
   - 1.1 to become well-informed and hence confident citizens in a technological world;
   - 1.2 to realise the usefulness, and limitations, of scientific method and its applications in other disciplines and in everyday life;
   - 1.3 to be suitably prepared should they intend to continue beyond the 16+ level with more specialised studies in pure sciences, in applied sciences, in science-dependent vocational courses;
   - 1.4 to have a suitable course should they cease to study science at this level:
2. to develop abilities and skills that
   - 2.1 are appropriate to the study and practice of science;
   - 2.2 are useful for everyday life and encourage safe practice:
3. to stimulate
   - 3.1 curiosity about science, enquiry into science, interest in science, enjoyment of science;
   - 3.2 respect for the environment:
4. to promote an awareness that
   - 4.1 the study and practice of science are cooperative and cumulative activities and are subject to social, economic, technological, ethical and cultural influences and limitations;
   - 4.2 the applications of science in everyday life may be both beneficial and detrimental in the contexts of the individual, the community and the environment.

Figure 5: UCLES O Level Environmental Science 5015 Syllabus, 1986

Common themes and concerns

"The answers in July shewed intelligence and good teaching; those in December shewed neither."

The period under consideration is nearly 150 years, during which time there were two major revisions of the natural history curriculum and numerous manifestations of natural history type qualifications. It is, however, possible to trace threads of continuity throughout the entire period.

The examiners' reports are a rich resource of comments, both informative and amusing. They are mostly critical and unashamedly opinionated, as demonstrated by the examiners of 1910 above, but they are a source of continuity in a complex puzzle of perennial change.

Quality of teaching

Criticism on the quality of teaching is common to all subjects and the following, though found in natural history sections, could be applied elsewhere. These include references to the use of "obsolete textbooks", inappropriate use of "technical terms", the need "to give pupils some questions of a problem type during the year", and "considerable evidence of poor reading of the questions".

The following, from the examiners of 1887, is more relevant to Botany teaching, and references a recurring theme related to practical work and observation:

> "It seems clear that if this subject is to be successfully taught in schools the teachers must keep their own knowledge fresh, and endeavour to teach the subject in such a way that it shall be an actual training for the eye and brain and not mere effort of memory."

Practical tests

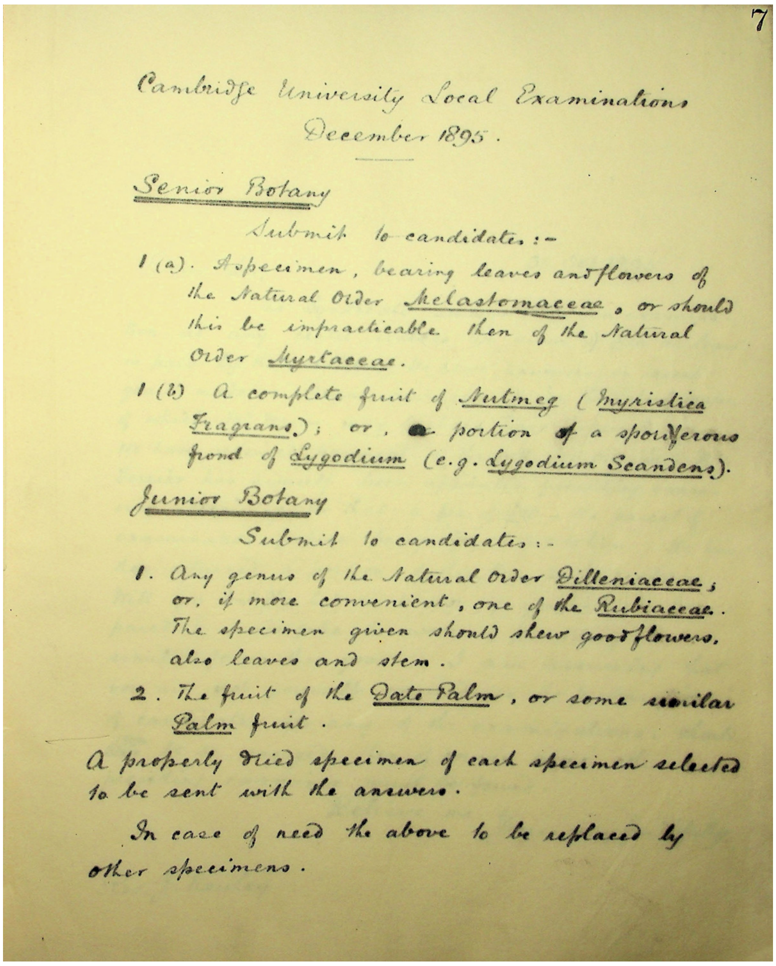
Botany and Zoology examinations included practical tests which attracted many comments on student training and experience. Comments such as that in 1897 that "there still appear to be a few schools where students have no practical instruction" led to clear direction to include practical teaching in the revised regulations of 1906, but examiners in 1908 noted dubious levels of compliance at Junior Botany exams:

> "The many inaccuracies in the description of experiments produced an impression that the apparatus had been arranged by the teacher, and that the students had had no opportunity of performing the experiments themselves."

By 1914 examiners had become frustrated and offered a stark warning to teachers that "unless the teaching of natural history can be assisted by fieldwork and other means of encouraging observation the result is not worth the time devoted to the subject".

The A level Botany report of 1958 repeats the theme, noting "too much preoccupation with the textbook … with too little appreciation of the living plant as observed by naked-eye and lens", and examiners in 1982 commented that "it would appear that field ecology and experimentation receives little attention in schools".

The GCSE Rural Science examiners of 2000 decided on a more proactive approach and stated a clear incentive to incorporate practical work into teaching: "Those with practical understanding of the topic, who planned their answer, or ensured their answers were based on their knowledge of Rural Science, gained most marks."


7

Cambridge University Local Examinations
December 1895.

Senior Botany

Submit to candidates:-

1 (a). A specimen, bearing leaves and flowers of the Natural Order Melastomaceae, or should this be impracticable, then of the Natural Order Myrtaceae.

1 (b) A complete fruit of Nutmeg (Myristica Fragrans); or, a portion of a sporiferous frond of Lygodium (e.g. Lygodium Scandens).

Junior Botany

Submit to candidates:-

1. Any genus of the Natural Order Dilleniaceae; or, if more convenient, one of the Rubiaceae. The specimen given should shew good flowers, also leaves and stem.

2. The fruit of the Date Palm, or some similar Palm fruit.

A properly dried specimen of each specimen selected to be sent with the answers.

In case of need the above to be replaced by other specimens.

Figure 6: Botany samples recorded in UCLES Letter Book, December 1895 (Ref: UCLES/A/LB/1/7)

Observation

"He [the student] should cultivate the habit of observing for himself."

Observation was an important part of the curriculum and was highly encouraged by examiners, as shown in this simple expression in the 1858 report, above. From 1906 to 1925, the Natural History and Botany schedules carry an annual reminder, written in italics, that "importance will be attached to the evidence of observation on the part of the candidate". This message is not only published annually, but published twice, as part of both the Junior and Senior regulations each year.

Throughout this period there are examiner encouragements: "excellent answers were given ... especially where the candidate described direct observations made in the country" (1910), and disappointments: "The answers this year again suggested only too often the lack of actual observation" (1912).

In 1919 the Botany examiners noted "a noticeable lack of ability to make original observations and to attend to details" and, in 1925, the Natural History examiners commented that "too small a part of the knowledge shewn had been gained by first-hand observation".

Some comments reference specific observational failings, as in 1907: “Hardly a candidate could determine either the age of a twig of Beech or the structure of its winter buds” and some, as this in 1912, record commendable effort: “Improvements especially in answers on birds … Knowledge of common British insects was weak. Several good essays on the horse were sent up.”

Candidate observational skills were still being criticised by examiners at the end of the century, and examiners of Rural Science GCSE in 1996 concluded that “questions which relied upon candidates’ practical knowledge and observation were badly answered”.

Complaints on observational skills are not confined to the candidates, and the following comment from the O level Agricultural Science examiners of 1984 is a direct indictment of the teachers:

> “The examiners were amazed, then baffled, and finally rather disconsolate that some teachers were not able to recognise and classify specimens that pupils were expected to identify.”

Drawing and diagrams

In the early years there was strong emphasis on drawing skills, which were essential before photography was commonplace. This is evident by the variety and prevalence of drawing examinations and drawing questions in Botany, Zoology, Natural History and Geography papers. Poor sketches and lack of precision are referenced in several examiners’ reports up to the 1930s, and the School Certificate Botany papers annually instruct candidates to answer with illustrations by large and clear diagrams or drawings.

An A level Zoology report on candidates’ work from 1954 to 1957 devotes a whole section to drawing, stressing the importance of drawing in demonstrating understanding of locomotion: “Candidates should be assured that the rapidly executed simple line drawing is as much a part of writing down zoological observation as is continuous prose.”

Throughout most of the period, drawing requirements changed little. An unconscious echo of the Drawing examination of 1858 finds O level Botany candidates in 1952 being asked to describe, with drawings, the climbing methods of a convolvulus plant.

There is, though, a notable shift in the significance of illustration towards the end of the period as the onus on candidates shifts from making to interpreting drawings in the question paper. The examiners of GCSE Rural Science in 1995 felt it necessary to note that: “It was obvious that many candidates responded without due reference to the diagrams.”

Environmental issues

“On what causes does the climate of a country principally depend?”

Questions on environmental issues can be found throughout the entire period but most, like the question above (from an 1858 Junior paper), are from Geography papers. Geography questions, however, can also overlap with Botany, and a Geography question on peat mosses in 1887 has similarities with a Botany question on the impact of soil and climate on vegetable growing in 1913. Reference to the interdependence of plants and animals features in a Botany examination as early as 1887, and there is repeated reference in both schedules and examination papers to "the natural order".

The 1970 O level Environmental Science syllabus includes, as a subsidiary idea, "the use made by Man of natural resources both of energy and of materials". On the final, 18th page of the syllabus, in a sub, sub, sub section, reference is made to chemical depletion of soil reserves, land use education, wise use of insecticides, the conservation of animal and plant life, and "world health".

The same syllabus in 1975 introduces an optional topic on plastics: raw materials, production and uses, and, in 1986, a further revision includes "respect for the environment" as one of the aims of a course which covers exploitation of resources, conservation and pollution. An optional section offers candidates an ecosystem study involving a pond, wood, or coastal habitat. This syllabus may have pre-empted changing views in society as the examiners' report is not positive and notes a "narrow treatment of conservation and land management" in the candidate answers.

Three years later, however, the examiners record "a noticeable improvement" and a possible cause: "In areas which have been popularised in the media, such as the greenhouse effect, and oil pollution, the quality of answers was particularly good." As environmental issues grew in the collective consciousness, candidate answers continued to impress the examiners who, in 1996, commented that Rural Science GCSE "questions on the environment and conservation were well answered".

Conclusion

This study is limited only by time and interest; the resources are plentiful and there is scope for more detailed study of all the resources, particularly the question paper holdings. Natural history assessments from the Cambridge examination board took several forms in this 150-year period and there are some common themes and trends that could be explored further. Some enduring themes are nicely summarised by the Botany examiner of 1866, just eight years after the examinations were set up:

> "The careful examination (of the flowers and seed-vessels) of common plants seems to be overlooked, though it is indispensable to the acquisition of any real knowledge of the subject. The search for plants in the fields is a healthy exercise and induces a love of nature: it ought to make the study a popular one."

It does seem clear that natural history assessments were initially a mainstream part of the curriculum but that, over time, they became displaced by science subjects and defined more as subsidiary than main subjects. The data in the tables shows that these subjects were not hugely popular with students, and the pass rates were not high. The resources used here reveal entrenched, long-term difficulties in encouraging practical skills and observational exercises, perhaps due to the size and nature of the subjects, which did not lend themselves well to traditional classroom teaching methods of the 19th and 20th centuries. The frequent renaming and repositioning of natural history type subjects indicate that the calibre of the candidates and the fitness for purpose of the qualifications were a constant cause for concern. However, the resilience of both to repeatedly revisit the subject, to absorb new subject areas and to manifest themselves in new ways indicate a strength of purpose that may yet be fulfilled in the new qualification of the 21st century.

Dedicated to Michael Paduano.

Bibliography

Secondary source references

1858 Question Paper Book, Cambridge University Press, 2008.

Charwat, E. (2006). The nature of Replications: Recontextualising Natural History Models and Casts from 19th to early 20th centuries in Britain and beyond.

Rader, K., & Cain, V. (2015). From natural history to science: display and the transformation of American museums of science and nature. *Museum and Society, 6*(2), 152-171.

Spary, E. C. (2018). "Natural History". Europe, 1450 to 1789: Encyclopedia of the Early Modern World.

Primary source references (Cambridge Assessment Archives)

Homerton College Further Education, Examination Regulations 1930, UCLES/JC/HCFE

Journal of Education news cuttings, UCLES/ PP/JNK 2

OCR Examination Statistics, 2000

OCR GCSE Question Papers, Regulations & Reports, Asset Bank Past Papers Resource, 1991–2000

Rural Pupil Teacher Examination Regulations, UCLES/A/RPT

UCLES 1st to 68th Annual Reports, 1858–1925, UCLES/EX/AR

UCLES Examination Regulations and Question Papers, 1858, 1867, 1877 and 1879 (Copies from Cambridge University Library holdings Ref: Cam.c.II.51)

UCLES Examination Regulations, Syllabuses, Question Papers and Reports, Bound Volumes, 1884–1986

UCLES Examination Statistics booklets, 1970–1998

UCLES Higher Examination Question Papers, Bound Volumes of Higher Examinations, 1895–1905

UCLES Letter Book, October 1895–August 1896, UCLES/A/LB/1/7

UCLES Question Papers, Reading Room Booklets, 1937, 1950s–1980s

The impact of GCSE Maths reform on progression to mathematics post-16

Carmen Vidal Rodeiro and Joanna Williamson (Research Division)

Introduction

Students in England aged 14–16 study GCSE (General Certificate of Secondary Education) qualifications in a range of subjects, and generally sit GCSE assessments at age 16. Since GCSEs in English and in mathematics are a prerequisite for most post-16 courses, and for many training programmes and jobs, GCSEs in these subjects are taken by almost all young people. In February 2013, the Secretary of State for Education in England announced his intention to reform GCSE qualifications "to ensure they are rigorous and robust, and give students access to high quality qualifications which match expectations in the highest performing jurisdictions" (Gove, 2013).

For mathematics, in particular, the new GCSE would "focus on ensuring that every student masters the fundamental mathematics that is required for further education and future careers", and, in particular, it would "be more demanding" and "provide greater challenge for the most able students" (Gove, 2013).

The new GCSE in mathematics has, therefore, a revised content framework and aims to better prepare students for progression to future education and employment. It was first assessed in summer 2017. Key changes to the qualification were a greater emphasis on problem-solving and more demanding content, together with a new grading scale from 9 to 1 (with 9 being the highest grade). More details about the subject content and the main assessment features of the new GCSE can be found in DfE (2013) or Ofqual (2017).

Background research

Prior to the GCSE reform, there were longstanding concerns about how well the GCSE in mathematics prepared students for progression to AS and A level[1] study in the subject.

1 GCE AS and A levels are level 3 subject-based academic qualifications available to students aged 16 and above in England.

Hernandez-Martinez et al. (2011) reported, drawing on interviews with students, that the GCSE in mathematics was inadequate preparation for many students with pass grades (especially grade C, but increasingly also grade B) for AS level study, with algebra being mentioned as the key problem. Similarly, Noyes and Adkins (2016) showed that the numbers (and proportions) of GCSE Maths grade C students completing any advanced mathematics were relatively small. In fact, around 99 per cent of students achieving a grade C in 2010 did not complete any advanced mathematics over the following three years.

Rushton and Wilson (2014) carried out a survey of teachers to identify the areas of mathematics that were problematic for students who had just completed the GCSE and wanted to study the subject further at A level. They showed that teachers believed that students were prepared adequately for AS and A level courses in most areas of mathematics, but they also identified other areas (e.g., algebra) where GCSEs were considered not to prepare students well.

In a more recent study exploring the perceptions and experiences of the transition between GCSE and AS level in mathematics of a small group of students, Rigby (2017) reported that the majority of students believed that the GCSE syllabus prepared them for the AS level syllabus but not to the extent they would have hoped. It was believed that a gap existed between the mathematics necessary to pass a GCSE and the mathematics that students need to be able to start AS or A level. As a result, most schools were requiring high grades for entry onto A level Maths courses to make sure students were prepared for the transition: for example, students often had to have achieved a grade B at GCSE or even a grade A in order to be accepted for an AS or A level in mathematics (Noyes & Sealey, 2012). In Rigby's research (Rigby, 2017), one of the suggestions to improve the transition between GCSE and AS level was to change the GCSE syllabus to a more rigorous one, by including more AS level material (this has now been implemented within the reformed GCSEs).

The balance between revising the GCSE qualification to be better preparation for the AS and A levels and ensuring that it was appropriate for students who were not intending to continue to further study was problematic as, for example, including more rigorous content could have undesired effects on the transitions of some students, particularly middle- and low-attaining students.

Despite the A level in mathematics having a period of sustained growth in entries in the years prior to the GCSE reform (see, for example, Gill and Williamson (2016) and Gill (2018)), concerns about participation in post-16 mathematics have emerged in recent years. Stakeholders and researchers worried that a particular combination of structural changes (the decoupling of AS and A level qualifications, curriculum changes to the A level in mathematics, and changes to Key Stage 5 funding) would lead to a reduction in the uptake of mathematics at level 3 (e.g., ALCAB, 2014; Lee et al., 2018; Redmond et al., 2020). Changes to GCSE Maths aimed to encourage students to better manage the transition to the A level. However, the number of entries in A level Maths fell by around 3.5 per cent in

2019[2] (DfE, 2019; 2020), with suggestions from school leaders that students might have been losing confidence in their abilities in the subject or being less inclined to take it as it was perceived as quite hard.

To date, there has been little published research on how the reform of GCSE Maths has affected mathematics learning and, in particular, on how it affected progression to further education (e.g., entries to A levels) or performance in level 3 qualifications in mathematics (e.g., AS and A level Maths; AS and A level Further Maths).

One of the few studies that considered this issue in some detail was carried out by Howard and Khan (2019). They conducted interviews with A level Maths teachers with experience in teaching students who had studied the legacy GCSE in mathematics and students who had studied the reformed GCSE. The interviews explored their perceptions of how the legacy and reformed GCSEs prepared students for A level study. In general, teachers were positive about the extent to which the reformed GCSE prepared students for A level and the majority commented that the reformed GCSE prepared students at least as well, if not better, than the legacy GCSE. Humphries et al. (2017) also carried out a small qualitative study involving a sample of teachers (in 12 schools) who were engaged in delivering the new GCSE. Participating teachers expressed the view that “students sitting the reformed mathematics GCSE would be leaving Key Stage 4 with more mathematical knowledge than previous cohorts”, and that this would apply across all attainment levels. This was an important point as it is well documented that participation in A level Maths has been skewed towards those with high GCSE grades in the subject.

Grima and Golding (2019) and Pearson Education (2019), who carried out a programme of research looking at the introduction of the new GCSE Maths, reported similar findings to those outlined above. However, although the general consensus was that the new GCSE prepared students well for A level, there were concerns about how the weaker students (those with a grade 5 or 6) would feel about their abilities in mathematics. This was also mentioned by the participants in a study by Lee et al. (2018) who reported on a large-scale survey of post-16 mathematics teachers carried out by MEI (Mathematics in Education and Industry). The participants in this study additionally suggested that they had seen a reduction in mathematical confidence for students at a grade 7 level, observing that “with only 52 per cent of the marks[3] required for a grade 7 it may be the case that students who would feel confident and capable of studying maths with a grade A in the past may no longer feel as confident and therefore as motivated to study the subject”.

---

2 Students taking the A level Maths in 2019 would have studied the reformed GCSE Maths.

3 This percentage (52 per cent of the marks required for grade 7) was lower post-reform (in 2019) than in pre-reform years. However, it should be borne in mind that grade boundaries in the first year(s) of reformed qualifications, as it is the case here, are usually lower than in pre-reform years and they gradually increase and stabilise over time to account for candidates’ drop in performance (Cuff et al., 2019).

Aim of the research

The current research aimed to add to the qualitative existing research described above by approaching the question on how the reform of GCSE Maths affected progression to further education (e.g., entries to A levels) or performance in level 3 maths and in maths-related subjects (e.g., achieving at least grade A at A level) via quantitative analysis of entries and performance data. In particular, the main research question was:

> *How does overall performance in GCSE Maths relate to progression to and subsequent attainment in level 3 mathematics, pre- and post-GCSE reform?*

The outcomes of this research will increase understanding of how recent reforms to the qualification have affected students, teachers and schools, and contribute evidence towards further understanding of progression from GCSE to level 3 mathematics.

Data and methods

Data

This work addressed the research question using national results data available in the National Pupil Database (NPD).

The NPD is a longitudinal database for children in schools in England, linking pupil characteristics to school and college learning aims and attainment. It holds individual pupil-level attainment data for pupils in all schools and colleges who take part in the exams, and pupil and school characteristics (e.g., age, gender, ethnicity, special educational needs, eligibility for free school meals, etc.) sourced from the School Census for maintained schools only.

Students who achieved a GCSE Maths in each of the years in Table 1 below (June sessions only) were followed up for two years and data for their level 3 qualifications in the four exam sessions before the end of Key Stage 5 were included. For example, students who achieved a GCSE Maths in 2015 were followed up in 2016 and 2017 and their AS and A level results identified. Note that later cohorts could not be included because end-of-course exams were cancelled in 2020 and 2021 due to the COVID-19 pandemic.

The analyses were restricted to students who were 16 years old at the end of each academic year. This age restriction was made to have a set of "typical" candidates at the end of Key Stage 4.

Table 1: GCSE Maths cohorts included in the research.

GCSE exam year	A level completion	GCSE Maths	Number of students achieving the GCSE
2014	2016	Legacy (A*–G)	505 962
2015	2017	Legacy (A*–G)	544 984
2016	2018	Legacy (A*–G)	521 772
2017	2019	Reformed (9–1)	530 482

As shown in Table 1, the GCSE grades awarded in the period of study belonged to two different grading scales: A*–G for the legacy qualifications, and 9–1 for the reformed GCSEs. For some of the analyses in this study, the GCSE Maths grades pre- and post-reform were converted to a common numerical scale using the Department for Education's conversion values[4] for 2017 and 2018 performance table calculations (DfE, 2016).

Progression from GCSE Maths to A level Maths, A level Further Maths and Core Maths[5] was investigated.

Methods

Descriptive statistics were produced on the number and proportion of GCSE Maths students progressing to the qualifications mentioned above during pre-reform (2014–16) and post-reform (2017) years. Progression was investigated overall and by GCSE grade. Pre- and post-reform grade distributions were also produced for all qualifications above, overall and for each GCSE Maths grade. To further explore the effect of GCSE reform on progression to and performance in level 3 mathematics, while controlling for students' backgrounds, multilevel logistic regression analyses were carried out. The outcomes (dependent variables) of the regressions were as follows:

- progression to A level Maths, Core Maths, and A level Further Maths
- achievement of at least grade A in A level Maths, Core Maths, and A level Further Maths
- achievement of at least grade C in A level Maths, Core Maths, and A level Further Maths.

The independent variables in the regression models included: the year GCSE Maths was achieved (this is an indicator of pre-reform (2014 to 2016) or post-reform (2017)), GCSE grade (using the common GCSE grade scale as described above), gender, overall prior attainment, level of deprivation and school type.

4 GCSE 9–1 grades kept their face value (i.e., 9=9, 8=8, etc.), and A*–G grades were mapped as follows: A* = 8.5, A = 7, B = 5.5, C = 4, D = 3, E = 2, F = 1.5, G = 1.

5 Core Maths is a level 3 qualification aimed at students who have passed GCSE Maths at grade 4 or above, but who have not chosen to study AS/A level Maths. It helps students consolidate and extend their mathematical knowledge and provides them with transferable mathematical skills to support their other level 3 subjects (e.g., psychology, geography, business-related courses, sports, social sciences, ...) and their transition to employment and further study. For more details see, for example, https://www.ocr.org.uk/qualifications/core-maths/.

The level of attainment at Key Stage 4 (prior attainment) was measured by an average GCSE and equivalents point score per entry (for details on how this was calculated, see DfE (2017)). The average GCSE and equivalents point score per entry, which ranges from 0 to 9, was used to divide students into three approximately equally sized groups: low attainment, medium attainment and high attainment. In each year, these terciles were based on the full Key Stage 4 cohort of students.

The level of income-related deprivation of the students was measured by the Income Deprivation Affecting Children Index (IDACI). This index is based on the student's home postcode and describes the percentage of children in a very small geographical area (Lower Layer Super Output Area or LSOA) living in low income families (more details about the IDACI can be found here: https://www.gov.uk/government/publications/english-indices-of-deprivation-2015-technical-report). It varies between 0 and 1 and indicates the level of income deprivation in the area in which a student lives. It cannot, however, indicate the level of income deprivation affecting an individual student. This measure was used to divide students into three approximately equally sized groups: low deprivation (more affluent), medium deprivation and high deprivation. As above, in each year, these terciles were based on the full Key Stage 4 cohort of students.

We classified schools into five groups: comprehensive schools, secondary modern schools, independent schools, selective schools, and other. *Comprehensive* and *secondary modern* schools (which include free schools and academies) do not select their intake on the basis of academic achievement or the wealth of the parents of the students they accept. *Selective* schools are state-funded schools that admit students on the basis of some sort of selection criteria, usually academic. *Independent* schools are fee-charging private schools, independent from many of the regulations and conditions that apply to state-funded schools. *Other* schools included, for example, sixth form and further education colleges, special schools, pupil referral units, tutorial colleges, and training centres.

Note that some of the variables described above are collected as part of the annual school census, which is only compulsory for state-maintained schools (and optional for independent schools). This can lead to high levels of missing data, particularly among independent school students, for some variables (e.g., IDACI deprivation). Students with missing data in any of the independent variables were not included in the regression analyses.

With logistic regression models such as the ones fitted in this research, estimates are hard to interpret directly because they are log odds of the outcome (e.g., progressing to A level; achieving at least a grade A). But, in simple terms a positive parameter estimate (for a categorical variable) means that being in that category is associated with a higher probability compared to being in the reference category. Negative values mean a reduction in probability. A positive parameter estimate for a continuous variable means that the increase in that variable is associated with an increase in the probability of the outcome. To aid interpretation, alongside the tables with the results from the regression analyses, figures are presented showing the probability of the outcome for different GCSE Maths grades and broken down by the GCSE year.

To ensure confidentiality of the data, statistical disclosure controls have been applied to the results (tables and graphs). In particular, counts below 10 and percentages based on counts below 10 have either been suppressed or merged with other counts/percentages.

Results

Progression to level 3 qualifications in mathematics

Table 2 below shows the overall progression to level 3 qualifications in mathematics of students who achieved a GCSE Maths pre-reform (2014 to 2016) and post-reform (2017).

Progression to A level Maths increased post-reform. However, this increase could be the continuation of a trend already present pre-reform (as shown in Table 2, progression to A level Maths had been increasing year on year in the last three years prior to the GCSE reform). Progression to Core Maths and A level Further Maths also increased post-reform, but it is worth noting that progression to both qualifications continued to be low in absolute terms.

Table 2: Overall progression to level 3 qualifications in mathematics.

GCSE year	Progression	A level Maths		A level Further Maths		Core Maths	
		N	%	N	%	N	%
2014	No	465 271	92.0	499 823	98.8	504 848	99.8
	Yes	40 691	8.0	6139	1.2	1114	0.2
2015	No	492 946	90.5	536 819	98.5	542 088	99.5
	Yes	52 038	9.5	8165	1.5	2896	0.5
2016	No	465 586	89.2	513 115	98.3	517 297	99.1
	Yes	56 186	10.8	8657	1.7	4475	0.9
2017	No	470 651	88.7	521 135	98.2	525 400	99.0
	Yes	59 831	11.3	9347	1.8	5082	1.0

Figure 1 shows the progression to level 3 qualifications in mathematics broken down by achievement in GCSE Maths, pre- and post-reform. For A level Maths and A level Further Maths, progression increased post-reform for all students. The increase in progression rates was higher among those who achieved at least grade A/7 than for students with at least grade C/4. For Core Maths, although progression also increased post-reform for all students, the increase was slightly lower among students who achieved at least grade A/7 than among students who achieved at least grade C/4 in their GCSE Maths.

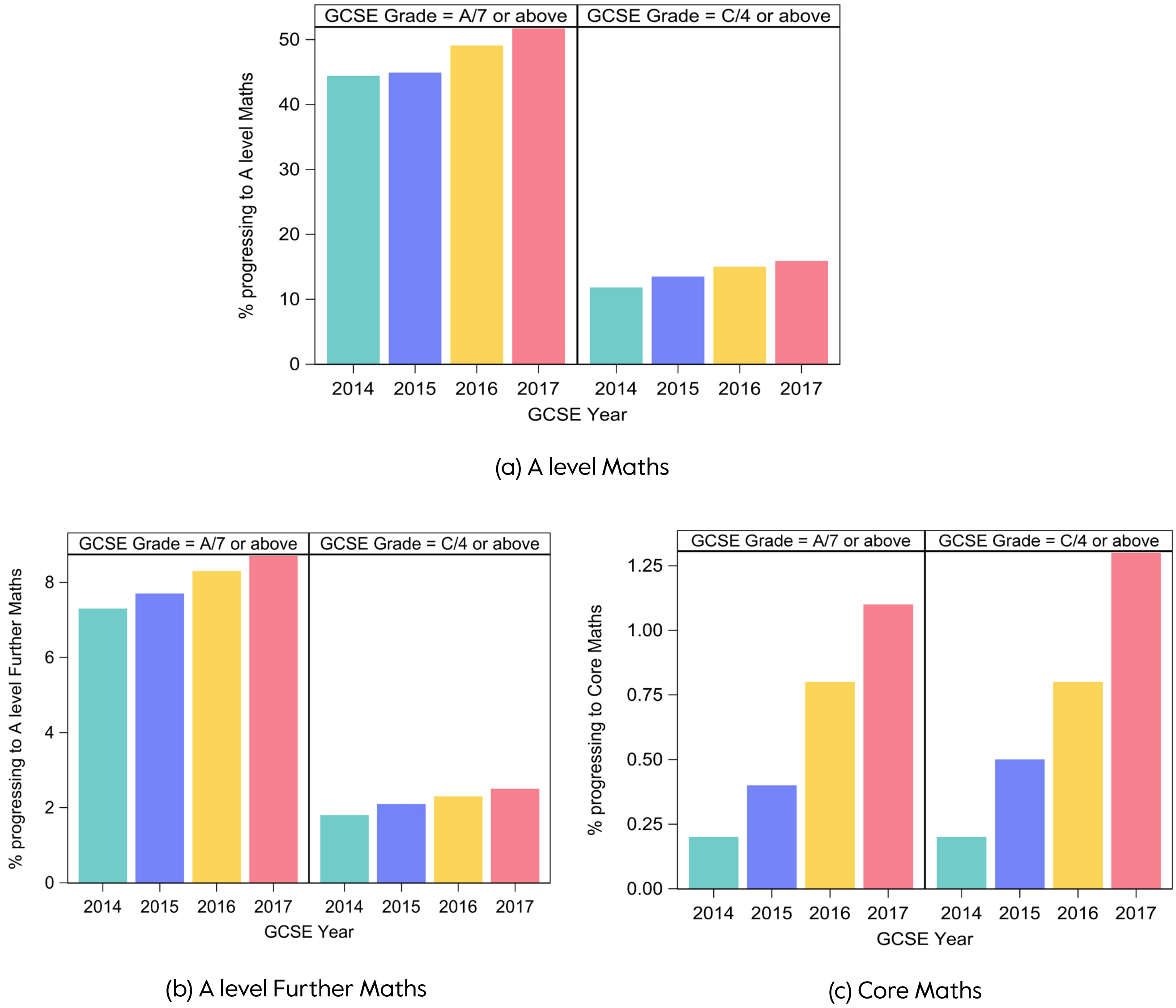


Figure 1: Progression to level 3 qualifications in mathematics, by GCSE year (pre-reform: 2014 to 2016; post-reform: 2017) and achievement of GCSE grade thresholds: A/7 or above and C/4 or above

To further explore the effect of GCSE reform on progression to A level Maths, while accounting for students' backgrounds, multilevel logistic regression analyses were carried out. The focus of the regression analyses was on the effect of the GCSE exam year (proxy for GCSE reform) and its interaction with the GCSE Maths grade. Students' background characteristics were included as controls. Table 3, which shows the results of the regression model looking at progression to A level Maths, indicates that the GCSE year was a statistically significant predictor of progression to A level Maths, and that its effect varied significantly by grade as shown by the interaction term included in the regression model (see the four bottom rows in Table 3).

To aid the interpretation of the results from the regression model, Figure 2 shows the probability of progressing to A level Maths according to GCSE exam year and GCSE Maths grade. The graph shows the probability values for a "reference candidate" (a female student, of medium prior attainment, medium level of deprivation, and in a comprehensive school). For students with different background characteristics, a similar picture is expected: the relationship between GCSE exam year, GCSE Maths grade, and progression to A level Maths does not change, although the actual probability values might be slightly higher/lower.

Figure 2 shows that the probability of progression post-reform (the solid red line) is below the probability of progression pre-reform for the low GCSE grades, but above for the high GCSE grades – so, towards the top of the GCSE distribution, the progression to A level becomes very slightly higher for students who achieved the GCSE in 2017 (post-reform). In particular, Figure 2 shows that a reference candidate with grade A/7 in GCSE Maths had a similar probability of progression pre- and post-reform: a probability of 0.12 to progress to A level Maths pre-reform (taking 2015 as an example, but very similar for the other pre-reform years) and a probability of 0.11 after the reform. However, the very top candidates had different probabilities of progression pre- and post-reform: a reference candidate with grade A* pre-reform (2015, A*=8.5) had a probability of progression of 0.56, while a reference candidate with grade 9 post-reform had a probability of 0.78.

Table 3: Progression to A level Maths, regression analysis results (N=1 761 038).

Variable		Estimate	Standard Error	P-value
Intercept		-16.361	0.073	<.0001
Gender	Male	0.951	0.008	<.0001
	[Female]	.	.	.
Deprivation	Medium	0.005	0.009	0.5890
	High	0.087	0.012	<.0001
	[Low]	.	.	.
Prior Attainment	Medium	2.827	0.050	<.0001
	High	3.432	0.048	<.0001
	[Low]	.	.	.
School Type	Independent	-1.028	0.594	0.0836
	Other	-0.259	0.143	0.0706
	Secondary Modern	-0.057	0.059	0.3357
	Selective	-0.160	0.045	0.0004
	[Comprehensive]	.	.	.
GCSE Maths Grade		1.641	0.007	<.0001
GCSE Exam Year	2014	1.041	0.075	<.0001
	2015	1.237	0.071	<.0001
	2016	1.557	0.069	<.0001
	[2017]	.	.	.
GCSE Maths Grade * GCSE Exam Year	2014	-0.151	0.011	<.0001
	2015	-0.166	0.010	<.0001
	2016	-0.190	0.010	<.0001
	[2017]	.	.	.

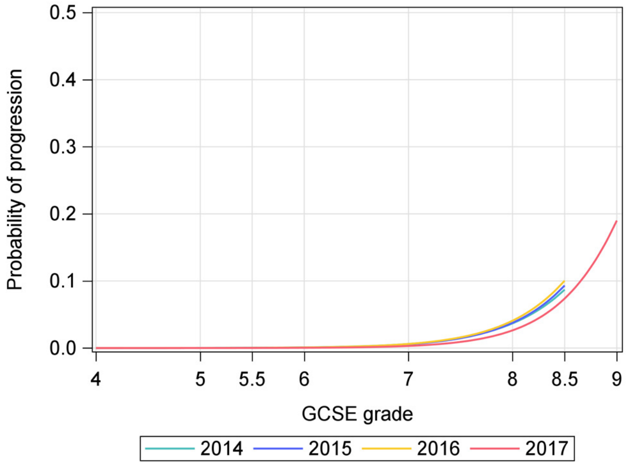
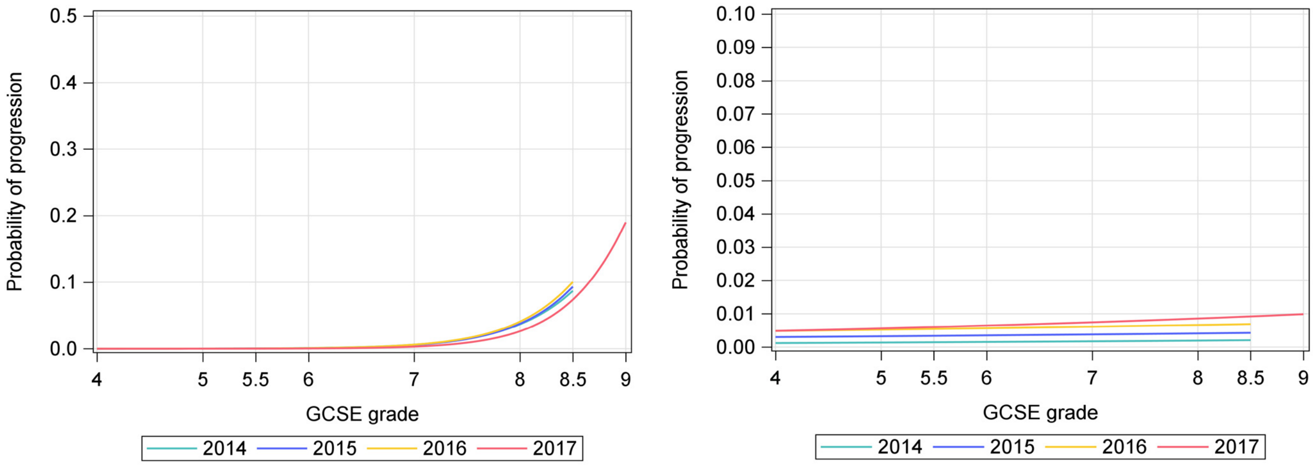


Figure 2: Probability of progression to A level Maths, by GCSE year and GCSE Maths grade (Gender=Female; Prior attainment=Medium; Deprivation=Medium; School type=Comprehensive)

The results of the regression models looking at progression to A level Further Maths and Core Maths are shown in Figure 3 (full outputs from the regression models are given in Table A1 in the Appendix). Note that, when looking at the probability graphs in Figure 3 the Y-axis scales differ.



(a) A level Further Maths

(b) Core Maths

Figure 3: Probability of progression, by GCSE year and GCSE Maths grade (Gender=Female; Prior attainment=Medium; Deprivation=Medium; School type=Comprehensive)

As in the model for progression to A level Maths, the results from the regression models confirmed that the year the GCSE was taken was a statistically significant predictor of progression to A level Further Maths and Core Maths, and its effect varied significantly by grade. Specifically, Figure 3(a) shows that the probability of progression to A level Further Maths post-reform was lower than the probability of progression pre-reform. On the contrary, Figure 3(b) shows that, in line with the results of the descriptive analyses, and although the rates of progression to Core Maths were very low pre- and post-reform, progression was slightly higher post-reform, independent of the grade achieved in GCSE Maths.

Performance in level 3 qualifications in mathematics

Figure 4(a) shows the A level Maths cumulative grade distribution for the cohorts of students progressing to A level Maths who achieved a GCSE in 2014 to 2016 (pre-reform) and in 2017 (post-reform).

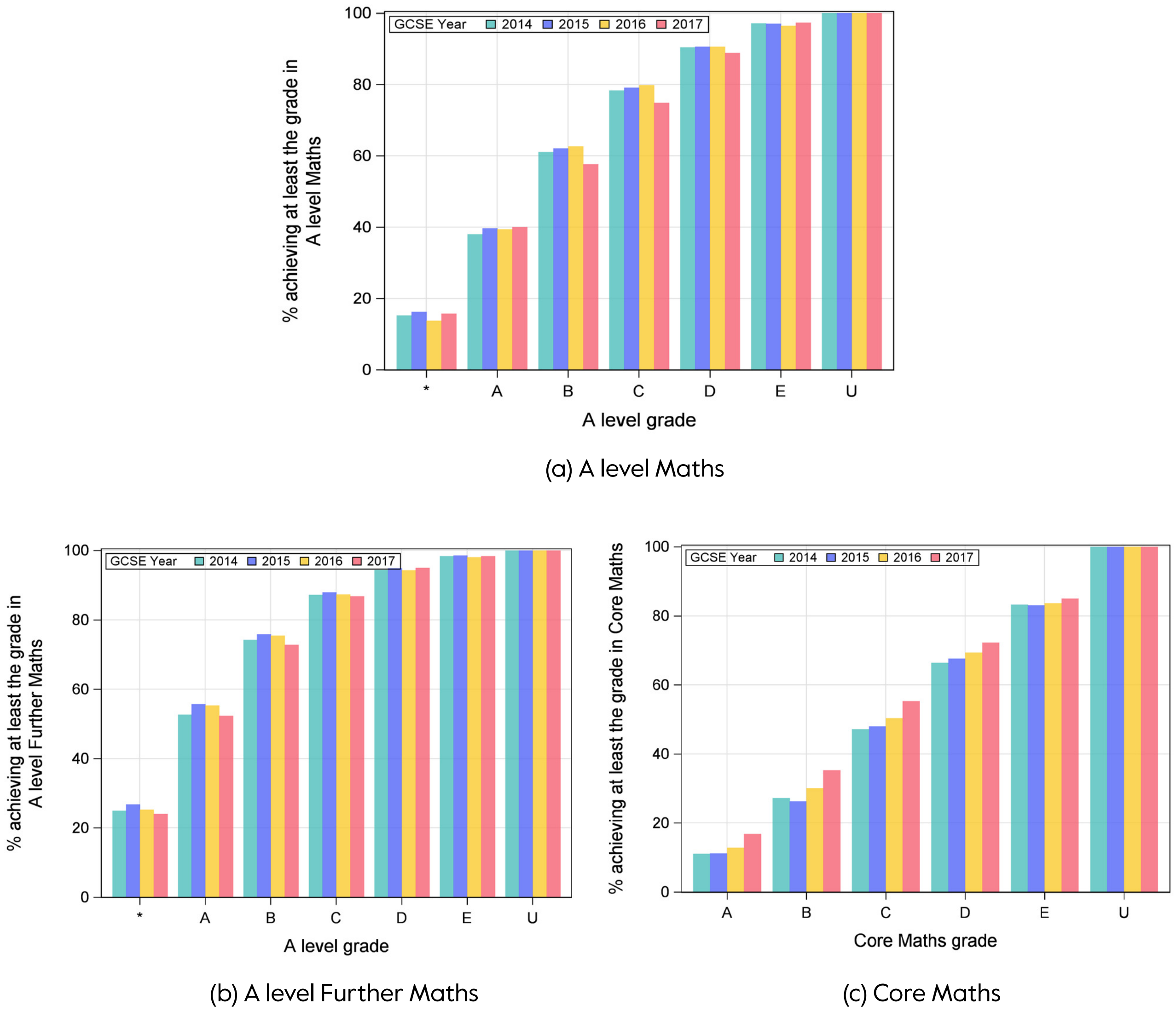


Figure 4: Cumulative grade distributions in level 3 qualifications in mathematics, by GCSE year (students progressing from GCSE Maths)

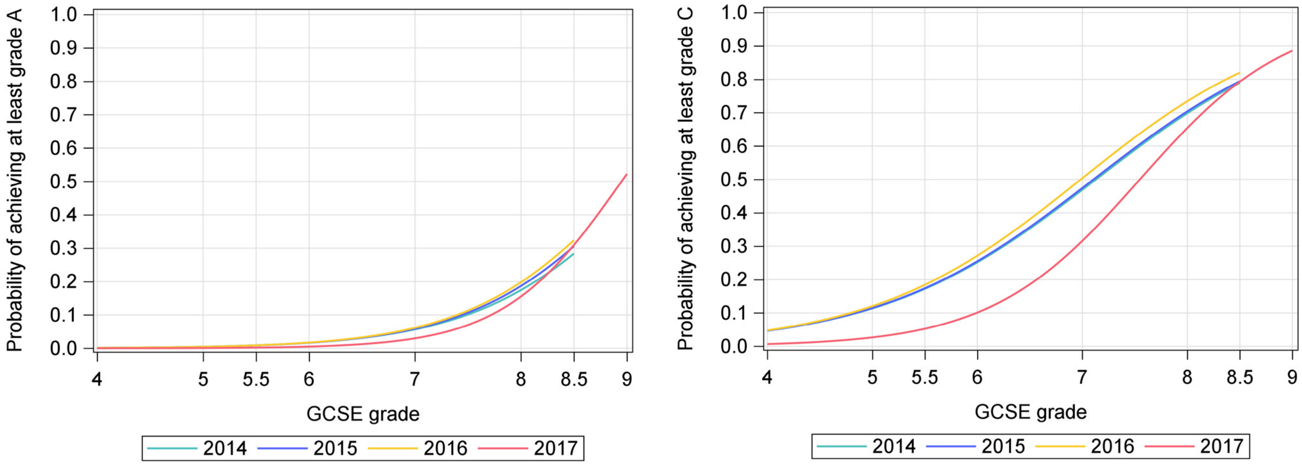
Compared to the last year pre-reform (2016), students who achieved a GCSE Maths post-reform (2017) were more likely to achieve an A* grade and at least grade A in their A level (although it was within the range for pre-reform years 2014 to 2016), but they were less likely to achieve grade B or above. Figure 4(b) shows that, compared to the pre-reform years, students who achieved GCSE Maths post-reform were less likely to get top grades (A*, at least grade A, at least grade B) in A level Further Maths. The picture for Core Maths (Figure 4(c)) was different: students who achieved a GCSE Maths post-reform performed better than students who achieved the GCSE pre-reform.

Table 4 shows the results of the regression analyses looking at the performance in A level Maths (i.e., achieving at least grade A; achieving at least grade C) pre- and post-reform. As for progression to A level Maths, the year the GCSE was taken was a statistically significant predictor of performance in A level Maths, and its effect varied significantly by grade, as shown by the interaction term included in the model (see bottom rows in Table 4).

Table 4: Achievement of grade thresholds in A level Maths, regression analysis results (N=176 398).

Variable		At least grade A			At least grade C		
		Estimate	Standard Error	P-value	Estimate	Standard Error	P-value
Intercept		-14.367	0.183	<.0001	-10.014	0.165	<.0001
Gender	Male	0.405	0.013	<.0001	0.227	0.015	<.0001
	[Female]	.	.	.	.	.	.
Deprivation	Medium	-0.134	0.014	<.0001	-0.156	0.016	<.0001
	High	-0.239	0.018	<.0001	-0.181	0.020	<.0001
	[Low]	.	.	.	.	.	.
Prior Attainment	Medium	-1.472	0.127	<.0001	-0.479	0.118	<.0001
	High	-0.555	0.122	<.0001	0.352	0.116	0.0025
	[Low]	.	.	.	.	.	.
School Type	Independent	0.869	1.194	0.4668	1.660	1.203	0.1677
	Other	0.224	0.213	0.2936	0.544	0.225	0.0155
	Secondary Modern	-0.431	0.066	<.0001	-0.376	0.065	<.0001
	Selective	0.252	0.038	<.0001	0.252	0.044	<.0001
	[Comprehensive]	.	.	.	.	.	.
GCSE Maths Grade		1.785	0.017	<.0001	1.411	0.016	<.0001
GCSE Exam Year	2014	4.364	0.197	<.0001	3.770	0.167	<.0001
	2015	4.079	0.184	<.0001	3.757	0.156	<.0001
	2016	3.992	0.176	<.0001	3.651	0.152	<.0001
	[2017]	.	.	.	.	.	.
GCSE Maths Grade * GCSE Exam Year	2014	-0.528	0.024	<.0001	-0.446	0.023	<.0001
	2015	-0.482	0.023	<.0001	-0.441	0.021	<.0001
	2016	-0.462	0.022	<.0001	-0.409	0.021	<.0001
	[2017]	.	.	.	.	.	.

Once we took into account the background of the students, including their prior attainment and their grade in GCSE Maths, both Table 4 and Figure 5 below show that the probability of achieving at least grade A at A level was lower post-reform (2017) than pre-reform (2014–16), apart from for the students who achieved the very top GCSE grades. In particular, a reference candidate with grade 7 in GCSE Maths had a higher probability of achieving at least a grade A at A level pre-reform than post-reform. The same patterns were found for the achievement of at least grade C.



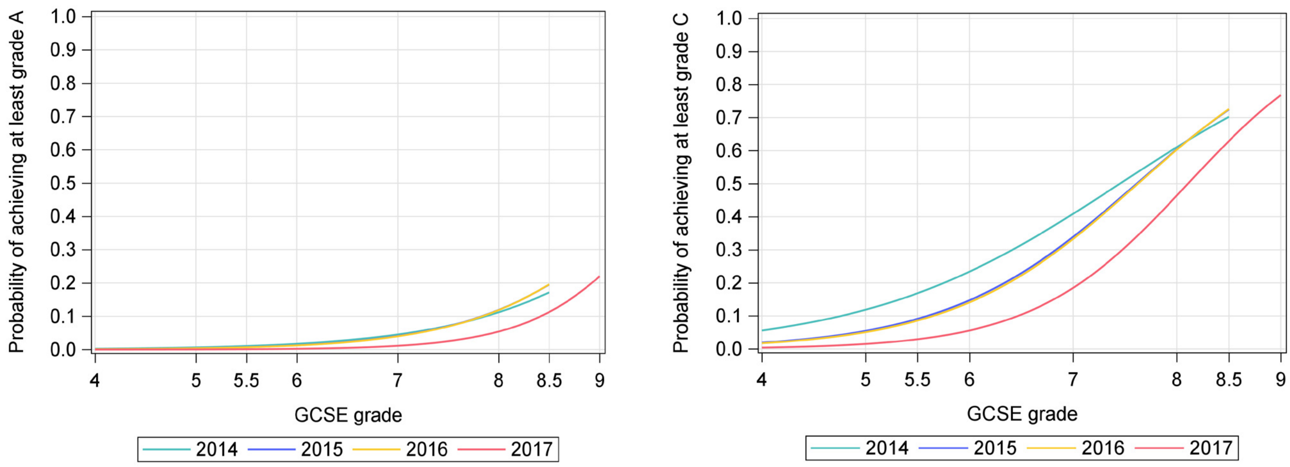
(a) Grade A or above

(b) Grade C or above

Figure 5: Probability of achieving a grade or above in A level Maths, by GCSE year and GCSE Maths grade (Gender=Female; Prior attainment=Medium; Deprivation=Medium; School type=Comprehensive)

The results of the regression models looking at performance in A level Further Maths and Core Maths are shown in Figure 6 and Figure 7, respectively (full outputs from the regression models are given in Table A2 and Table A3 in the Appendix).

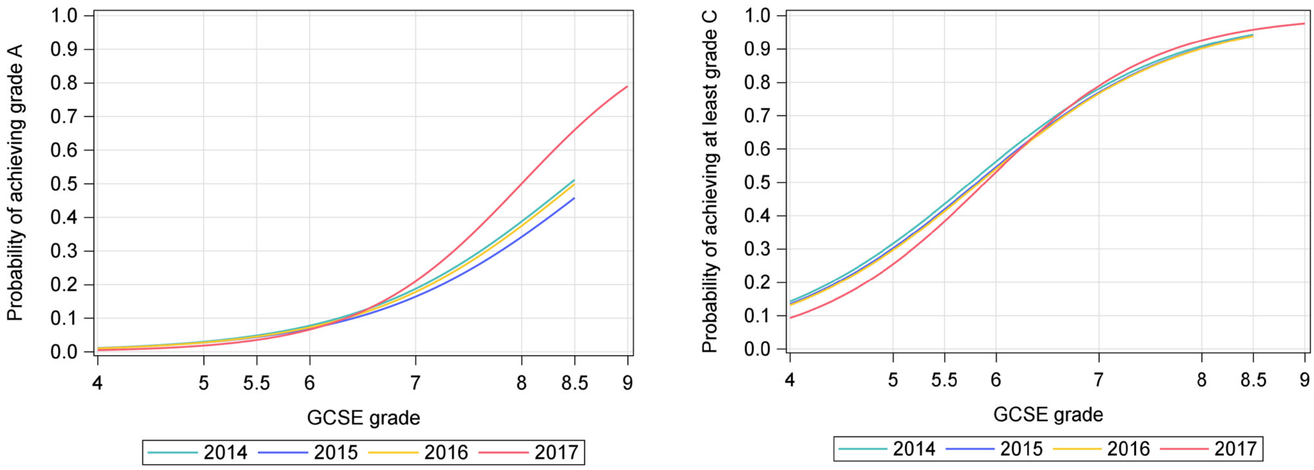
Once we took into account the background of the students, including their prior attainment and their grade in GCSE Maths, Figure 6 shows that the probability of achieving at least grade A or at least grade C in A level Further Maths was lower post-reform (2017) than pre-reform (2014–16), apart from for the students who achieved the very top GCSE grades. Performance in Core Maths was, however, generally higher post-reform (see the red lines for 2017 in Figure 7(a) and Figure 7(b)) than pre-reform.



(a) Grade A or above

(b) Grade C or above

Figure 6: Probability of achieving a grade or above in A level Further Maths, by GCSE year and GCSE Maths grade (Gender=Female; Prior attainment=Medium; Deprivation=Medium; School type=Comprehensive)



(a) Grade A

(b) Grade C or above

Figure 7: Probability of achieving a grade or above in Core Maths, by GCSE year and GCSE Maths grade (Gender=Female; Prior attainment=Medium; Deprivation=Medium; School type=Comprehensive)

Discussion and conclusions

This research has explored how well the GCSE in Maths prepared young people for further study in mathematics in the context of GCSE reform.

Contrary to fears of reduction in the uptake of A level Maths following the reform (e.g., Lee et al., 2018; Redmond et al., 2020) this research showed that progression generally increased post-reform. It should be noted, however, that this increase could be the continuation of a trend already present pre-reform (progression to A level Maths had been increasing year on year in the last three years prior to the GCSE reform). When controlling for students' backgrounds (including the grade achieved in GCSE Maths) the probability of progression post-reform was just below the probability of progression pre-reform for students with low GCSE grades. On the contrary, for students who achieved GCSE grades towards the top of grade distribution, the progression to A level was very slightly higher post-reform.

Performance in A level Maths was, in general, lower post-reform. In particular, the probability of achieving at least grade A or at least grade C in A level Maths was lower post-reform for students with any GCSE grade, apart from the students at the very top of the GCSE grade distribution. This contrasts with the perceptions of A level Maths teachers interviewed in research by Howard and Khan (2019) or Humphries et al. (2017), who commented that the reformed GCSE prepared students for the A level at least as well, if not better, than the legacy GCSE and that students sitting the reformed GCSE would be leaving Key Stage 4 with more mathematical knowledge than previous cohorts. However, it should be taken into account that students taking the reformed GCSE would have also taken the newly reformed A level Maths,[6] and it is well known that student performance tends to dip slightly in the first years of a new qualification (i.e., there is a sawtooth effect, as described, for example, in Cuff et al. (2019)). While the approach to awarding and grading A levels in this context (Newton, 2020) should have smoothed the sawtooth effect when looking at grade distributions, there could still be some evidence of relative under-performance. Furthermore, research showed that the reformed A level specifications were significantly more demanding than legacy specifications (Redmond et al., 2020), and there was concern from some teachers that while more able students may benefit from the more "aspirational" A level, lower performing students may be impacted negatively by the changes.

Progression to other level 3 qualifications in mathematics such as Core Maths or A level Further Maths increased post-reform (although it should be noted that progression to either of these qualifications was quite low both pre- and post-reform). In the case of Core Maths, as suggested by Mathieson et al. (2020), this increase could be seen as the result of the opportunity that this subject provides students for whom there was previously no option to study maths post-16. There were, however, differences in progression by the grade achieved in GCSE Maths: the increase in progression to Core Maths was slightly lower among students who achieved top grades (at least grade A/7) than among students who achieved lower grades (at least grade C/4), while the opposite pattern was found for progression to A level Further Maths. It is worth noting, however, that when accounting for the students' background characteristics, the probability of progression to A level Further Maths post-reform was lower than the probability of progression pre-reform.

Regarding performance in Core Maths and A level Further Maths, students who achieved a GCSE Maths post-reform were more likely to achieve top grades

6 Alongside GCSE reform, A levels have also been reformed. For example, students who sat the reformed GCSE Maths in 2017 (first year of assessments after the GCSE reform) were the first full cohort to sit the reformed Maths and Further Maths A levels in summer 2019 (A level Maths was available after one year of study in summer 2018; however, the entries in summer 2018 were small and were mainly younger students also studying A level Further Maths).

(grades A or B) in Core Maths, compared to students who achieved their GCSE Maths in pre-reform years. On the contrary, compared to the pre-reform years, students who achieved a GCSE Maths post-reform were less likely to achieve both grade A or above and grade C or above in A level Further Maths.

The new GCSE in mathematics also aimed to better prepare young people for further study in subjects with significant mathematical content (e.g., science subjects, economics, psychology). Howard and Khan (2019) reported that the reformed GCSE had positive implications beyond studying A level Maths and that the new GCSE would support students' progression to and performance in other subjects with mathematical content. Further to the research presented here, Vidal Rodeiro and Williamson (2022) investigated the impact of GCSE Maths reform on five maths-related A levels (biology, chemistry, physics, economics and psychology). Compared to pre-reform years, they found that overall progression was higher post-reform in all five subjects. Furthermore, performance in A level science subjects (biology, chemistry and physics) was very similar pre- and post-reform for students with the very top GCSE grades in mathematics, but it was lower post-reform for students with lower grades in GCSE Maths. However, in economics and psychology, performance was very similar pre- and post-reform.

The research discussed in this article is set in the context of recent reforms to GCSEs and A levels and, as with any reforms, changes take time to bed in. Given that this research focused on the first year after the reform (the new GCSE Maths was first assessed in 2017), it is possible that the results do not reflect how the reformed GCSE Maths will impact progression to and performance in level 3 qualifications in mathematics and subjects with mathematical content over the coming years. In the interim, however, the results of this research have raised important issues for the mathematics education community and for policy makers by increasing the understanding of how recent reforms to GCSE Maths have affected students, and contributing evidence on its impact on progression to post-16 study.

Overall, the findings indicate that some aims of the curriculum and assessment reform in upper secondary mathematics (in particular, increasing uptake of post-16 mathematics) may have been fulfilled. Going forward, it will be important to monitor the uptake of and performance in different post-16 mathematics qualifications (particularly by mid-attaining students), and continue to triangulate teacher perceptions with trends in attainment.

Acknowledgements

This work was produced using statistical data from the Office for National Statistics (ONS). The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets that may not exactly reproduce National Statistics aggregates.

References

ALCAB. (2014). *Report of the ALCAB panel on Mathematics and Further Mathematics*. The A Level Content Advisory Board.

Cuff, B. P., Meadows, M., & Black, B. (2019). An investigation into the Sawtooth Effect in secondary school assessments in England. *Assessment in Education: Principles, Policy & Practice*, *26*(3), 321–339.

Department for Education. (2013). *Mathematics: GCSE subject content and assessment objectives.*

Department for Education. (2016). *Progress 8: How Progress 8 and Attainment 8 measures are calculated.*

Department for Education. (2017). *16-18 Accountability Measures: Technical Guide for measures in 2017.*

Department for Education. (2019). *A level and other 16 to 18 results: 2017 to 2018 (revised).*

Department for Education. (2020). *A level and other 16 to 18 results: 2018 to 2019 (revised).*

Gill, T. (2018). *Uptake of GCE A level subjects 2017. Statistics Report no. 121.* Cambridge Assessment.

Gill, T., & Williamson, J. (2016). *Uptake of GCE A level subjects 2015. Statistics Report no. 109.* Cambridge Assessment.

Gove, M. (2013). *Ofqual policy steer letter: reforming Key Stage 4 qualifications.* [Letter from the Secretary of State for Education to Ofqual's Chief Regulator].

Grima, G., & Golding, J. (2019). *Reformed GCSE Mathematics qualifications: teachers' views of the impact on students starting A levels*. Ofqual Educational Assessment Seminar Scarman House, Warwick University.

Hernandez-Martinez, P., Williams, J., Black, L., Davis, P., Pampaka, M., & Wake, G. (2011). Students' views on their transition from school to college mathematics: rethinking 'transition' as an issue of identity. *Research in Mathematics Education*, *13*(2), 119–130.

Howard, E., & Khan, A. (2019). *GCSE reform in schools: The impact of GCSE reforms on students' preparedness for A level Maths and English literature.* Office of Qualifications and Examinations Regulation.

Humphries, S., Cotton, W., Khan, A., & Taylor, R. (2017). *GCSE mathematics: understanding schools' approaches to tiering*. Office of Qualifications and Examinations Regulation.

Lee, S., Lord, K., Dudzic, S., & Stripp, C. (2018). *Investigating the Impact of Curriculum and Funding Changes on Level 3 Mathematics Uptake*. Mathematics in Education and Industry.

Mathieson, R., Homer, M., Tasara, I., & Banner, I. (2020). 'Core Maths chooses you; you don't choose Core Maths'. The positioning of a new mathematics course within the post-16 curriculum in England. *The Curriculum Journal*, *31*(4), 704–721.

Newton, P. E. (2020). *Maintaining Standards: During normal times and when qualifications are reformed.* Office of Qualifications and Examinations Regulation.

Noyes, A., & Adkins, M. (2016). Reconsidering the rise in A-Level Mathematics participation. *Teaching Mathematics and its Applications: An International Journal of the IMA*, *35*(1), 1–13.

Noyes, A., & Sealey P. (2012). Investigating participation in Advanced level mathematics: a study of student drop-out. *Research Papers in Education*, *27*(1), 123–138.

Ofqual. (2017). *GCSE Subject Level Conditions and Requirements for Mathematics.* Office of Qualifications and Examinations Regulation.

Pearson Education. (2019). *GCSE Mathematics Qualification – UK: Regulated qualification efficacy report*. Pearson UK.

Redmond, B., Golding, J., & Grima, G. (2020). Teaching and learning for 'moving goal-posts': Reformed A Levels in mathematics. In R. Marks (Ed.), *Proceedings of the British Society for Research into Learning Mathematics, 40*(1).

Rigby, C. (2017). Exploring students' perceptions and experiences of the transition between GCSE and AS Level mathematics. *Research Papers in Education*, *32*(4), 501–517.

Rushton, N., & Wilson, F. (2014). *Teachers' and employers' views on the transition from GCSE mathematics to A-level mathematics or employment.* Paper presented at the British Educational Research Association conference, London.

Vidal Rodeiro, C., & Williamson, J. (2022). *The impact of GCSE Maths reform on progression to A level.* Cambridge University Press & Assessment.

Appendix

Table A1: Progression to level 3 qualifications in mathematics (A level Further Maths; Core Maths), regression analysis results (N = 1 761 038).

Variable		A level Further Maths			Core Maths		
		Estimate	Standard Error	P-value	Estimate	Standard Error	P-value
Intercept		-23.965	0.250	<.0001	-7.648	0.067	<.0001
Gender	Male	1.183	0.017	<.0001	0.021	<.0001	<.0001
	[Female]	.	.	.	.	.	.
Deprivation	Medium	0.020	0.017	0.2506	0.024	0.3610	0.3610
	High	0.026	0.022	0.2503	0.029	0.0063	0.0063
	[Low]	.	.	.	.	.	.
Prior Attainment	Medium	3.055	0.184	<.0001	0.039	<.0001	<.0001
	High	3.206	0.180	<.0001	0.048	<.0001	<.0001
	[Low]	.	.	.	.	.	.
School Type	Independent	0.310	0.893	0.7285	1.025	0.8374	0.8374
	Other	-0.252	0.278	0.3648	0.243	0.0200	0.0200
	Secondary Modern	-0.529	0.092	<.0001	0.155	0.1528	0.1528
	Selective	-0.177	0.050	0.0004	0.136	<.0001	<.0001
	[Comprehensive]	.	.	.	.	.	.
GCSE Maths Grade		2.160	0.020	<.0001	0.143	0.010	<.0001
GCSE Exam Year	2014	2.976	0.258	<.0001	0.116	<.0001	<.0001
	2015	2.150	0.254	<.0001	0.081	0.0065	0.0065
	2016	2.275	0.244	<.0001	0.072	0.0004	0.0004
	[2017]	.	.	.	.	.	.
GCSE Maths Grade * GCSE Exam Year	2014	-0.329	0.031	<.0001	0.021	0.2477	0.2477
	2015	-0.223	0.030	<.0001	0.014	<.0001	<.0001
	2016	-0.228	0.029	<.0001	0.013	<.0001	<.0001
	[2017]	.	.	.	.	.	.

Table A2: Achievement of grade thresholds in A level Further Maths, regression analysis results (N = 27 386).

Variable		At least grade A			At least grade C		
		Estimate	Standard Error	P-value	Estimate	Standard Error	P-value
Intercept		-14.868	0.716	<.0001	-10.739	0.644	<.0001
Gender	Male	0.162	0.033	<.0001	0.132	0.047	0.0052
	[Female]	.	.	.	.	.	.
Deprivation	Medium	-0.207	0.032	<.0001	-0.299	0.047	<.0001
	High	-0.400	0.040	<.0001	-0.500	0.054	<.0001
	[Low]	.	.	.	.	.	.
Prior Attainment	Medium	-0.578	0.550	0.2934	0.168	0.451	0.7094
	High	0.991	0.538	0.0653	1.398	0.445	0.0017
	[Low]	.	.	.	.	.	.
School Type	Independent	1.187	1.872	0.5258	-1.080	1.770	0.5417
	Other	-0.293	0.419	0.4834	0.057	0.536	0.9157
	Secondary Modern	-0.376	0.148	0.0110	-0.231	0.189	0.2215
	Selective	0.442	0.055	<.0001	0.440	0.082	<.0001
	[Comprehensive]	.	.	.	.	.	.
GCSE Maths Grade		1.599	0.054	<.0001	1.341	0.056	<.0001
GCSE Exam Year	2014	5.705	0.725	<.0001	4.776	0.691	<.0001
	2015	4.334	0.714	<.0001	2.581	0.666	0.0001
	2016	4.185	0.687	<.0001	2.384	0.639	0.0002
	[2017]	.	.	.	.	.	.
GCSE Maths Grade * GCSE Exam Year	2014	-0.614	0.085	<.0001	-0.523	0.084	<.0001
	2015	-0.434	0.084	<.0001	-0.252	0.081	0.0018
	2016	-0.416	0.080	<.0001	-0.228	0.078	0.0032
	[2017]	.	.	.	.	.	.

Table A3: Achievement of grade thresholds in Core Maths, regression analysis results (N = 12 140).

Variable		Grade A			At least grade C		
		Estimate	Standard Error	P-value	Estimate	Standard Error	P-value
Intercept		-11.133	0.466	<.0001	-7.725	0.284	<.0001
Gender	Male	0.297	0.070	<.0001	0.350	0.052	<.0001
	[Female]	.	.	.	.	.	.
Deprivation	Medium	-0.150	0.076	0.0476	-0.134	0.057	0.0188
	High	-0.427	0.092	<.0001	-0.299	0.064	<.0001
	[Low]	.	.	.	.	.	.
Prior Attainment	Medium	0.676	0.282	0.0165	0.787	0.112	<.0001
	High	1.525	0.282	<.0001	1.717	0.119	<.0001
	[Low]	.	.	.	.	.	.
School Type	Independent	-6.107	23.842	0.7978	4.548	26.927	0.8659
	Other	2.270	0.663	0.0006	0.708	0.563	0.2083
	Secondary Modern	-0.287	0.264	0.2778	-0.441	0.189	0.0196
	Selective	0.201	0.197	0.3084	0.397	0.187	0.0333
	[Comprehensive]	.	.	.	.	.	.
GCSE Maths Grade		1.326	0.058	<.0001	1.199	0.047	<.0001
GCSE Exam Year	2014	2.078	0.811	0.0104	1.189	0.535	0.0262
	2015	2.168	0.580	0.0002	1.128	0.391	0.004
	2016	1.959	0.504	0.0001	1.083	0.346	0.0018
	[2017]	.	.	.	.	.	.
GCSE Maths Grade * GCSE Exam Year	2014	-0.317	0.131	0.0153	-0.177	0.101	0.0800
	2015	-0.353	0.093	0.0001	-0.178	0.072	0.0137
	2016	-0.309	0.079	0.0001	-0.174	0.063	0.0058
	[2017]	.	.	.	.	.	.

An example of redeveloping checklists to support assessors who check draft exam papers for errors

Sylvia Vitello, Victoria Crisp and Jo Ireland (Research Division)

Introduction

When new exam papers are drafted, they need to go through a quality assurance process, just as we would expect for all important educational and non-educational products. Many aspects of question design contribute to ensuring that exam results accurately reflect a learner's relevant knowledge, skills and understanding and, thus, to ensuring that it is appropriate to use the assessment results in the intended ways. For example, features such as language accessibility, visual resources and context affect how well learners can show what they have learned (e.g., Crisp & Sweiry, 2006; Ahmed & Pollitt, 2007; Crisp, 2011; Crisp & Macinska, 2020). The most extreme problems with exam questions occur where a clear error appears in a paper. For example, this could be a factual inaccuracy which then renders a question unanswerable, or something on an exam paper that gives away the answer to a question elsewhere on the paper. It is of great importance that awarding bodies have robust procedures in place to ensure that questions are of high quality and errors are avoided. These procedures often involve a staged process through which exam papers are developed incorporating input from a number of assessors with expertise in the relevant subject.

Recently, Suto and Ireland (2021) reviewed the literature from the field of error detection and explored the psychological and system-level causes of errors in order to recommend a set of principles for how to minimise errors in exam papers. They highlighted various psychological causes of errors that are relevant to the context of exam paper construction. These include cognitive biases and limitations related to memory, attention and our tendency to use heuristics (i.e., imperfect, non-rational methods) over analytical approaches during tasks involving judgement and decision making. These human characteristics can cause us to make errors and to fail to detect ones made by others or ourselves. A checklist is one type of tool that can potentially help avoid errors by supporting appropriate checks during the stages of a process. Gawande (2010) explains how the use of checklists in error prevention or detection is supported by psychological theories of cognition and attention. Checklists can help overcome or mitigate psychological error factors by acting as aide-memoires that encourage the user to take a more systematic and analytic approach to the checking task. Another reason why checklists can help with certain cognitive limitations is that they specify the checker's task, making it clear to the checker what they should check for. Under-specification of a task or process has been argued to be a significant factor in the production of human error across industries (Reason, 2013).

This article is about some of the checklists that are used to support exam paper production processes at Cambridge University Press & Assessment. We discuss the approach that was taken to redevelop these checklists. The aim of this article is to draw attention to the complexity that is involved in designing checklists and, also, to provide a practical example of how research can be used strategically to inform key design decisions.

Project context and aim

In Cambridge University Press & Assessment, exam papers and other assessment materials are produced through a process of drafting, review and refinement involving a number of professionals with appropriate subject and assessment expertise. Focusing more specifically on OCR (one of our exam boards), during its assessment materials production process, exam papers (and other assessment materials) undergo a specific, carefully designed series of checks after a complete draft has been produced. These checks are aimed at ensuring the quality of the assessment materials, including detecting errors so that they can be corrected before the paper is sat by candidates. This article describes a research-informed project that involved redeveloping the checklists (and other related materials) to support OCR's assessors at this stage of the quality assurance process when a complete draft of the paper has been produced. The focus was on four of OCR's checking roles:

- Candidate Proxy – an assessor who works the exam paper as if they are a candidate.
- Assessment Marker – an assessor who marks the Candidate Proxy's exam script and reviews the alignment between the exam paper and the mark scheme.
- Assessment Analyst – an assessor who applies a question analysis technique to all of the questions in the paper, whereby they deconstruct the constituent words, phrases and parts of the question.
- Pre-exam Check – one of the final checks of exam papers, which is performed by an assessor[1] whose primary aim is to catch any serious errors that could affect the candidates' ability to answer questions.

The Candidate Proxy, Assessment Marker and Assessment Analyst all complete their checks around the same time. The results of these checks are then reviewed by the assessment manager within OCR and the paper is revised as needed. Some papers then undergo a proofreading check and a plagiarism check followed by another internal review by the assessment manager. The Pre-exam Check occurs after all these other stages.

1 The Pre-exam Check is sometimes performed by an external assessor or an internal assessment specialist with no prior involvement in the exam paper's development.

Before the redevelopment project started, OCR had already been providing their assessors with a type of checklist to use when performing these checking roles. These checklists took the form of a set of questions about the exam paper for assessors to consider and respond to when performing their checks. These checklist questions were embedded within a document known as the "report form". Assessors used this report form alongside conducting their checks and filled it in with details of their evaluation of the paper including the issues they identified. These completed forms were then reviewed by the assessment manager.

The main focus of the redevelopment project was to create a new report form for each of these roles, which would contain a new checklist that was strategically designed to ensure that all issues appropriate to a role would be checked. In addition to the checklists and report forms, the other materials that checkers would be provided with were reviewed and revised (specifically, instruction documents and training materials) to ensure that they also cohered with the new report forms. This article reports primarily on the checklists and report forms. The final checklist for the Assessment Analyst role is shown in the Appendix as an example.

Overview of the redevelopment approach

As this project was strongly linked to the operational running of the assessment materials production process and aimed to take a research-informed approach, a cross-department working group was established including researchers, assessment managers, a manager who oversaw a team of staff who co-ordinate the work of external assessors, and a project manager. This collaborative approach was critical for ensuring that, while drawing on the relevant research literature, decisions about the new checklists and redeveloped materials aligned with OCR's vision and intentions for the checking roles, and that the materials would be useable in practice.

The project was collaborative and iterative with the redevelopment usefully informed by input gathered at various stages of the process from a range of people with different roles in the question paper process. The main stages of work were:

- a mapping exercise in which a taxonomy of question paper error types was mapped against each of the four checks in order to set out what should be checked for each role
- initial design and drafting of checklists and report forms
- consulting internal staff involved in the assessment materials production process, followed by refining the materials as needed
- piloting the materials in several subjects with those who conduct the checks – assessors were asked to check an example exam paper using the revised checklist and other support materials and provided feedback, after which the materials were further refined
- consulting internal staff again on the changes and minor further revisions.

The next section of this article describes the mapping exercise that formed the foundation of our checklists. The two subsequent sections focus on design decisions relating to the checklist items and design decisions about the report forms within which the checklists would appear. These design decisions evolved throughout the course of the redevelopment project, with final decisions often resulting from an iterative consideration of information, discussions and evidence across different stages of the project (e.g., initial design, consultation, piloting and refinement process). For brevity, the sections on design decisions bring together themes that arose at any stage of the redevelopment process rather than separating out points based on the chronology of events.

Redeveloping the checklists and report forms

Mapping exercise

As the starting point for revising the checklists and report forms, OCR members of the project working group mapped out the types of error that each checking stage (role) was intended to identify. A taxonomy of 42 error types that had recently been developed by Suto et al. (2023) was used. This was derived from an analysis of Cambridge University Press & Assessment's records of assessment materials errors from across several years (2012 to 2018).

The mapping exercise was based on an approach recently developed by Suto et al. (2023) as a way to systematically and strategically evaluate existing checks of assessment materials. They showed how the approach can help assessment teams to understand, for example, whether all error types are being targeted across checking roles, how many error types are targeted by any individual role (i.e., checker workload), and how many times each error type is checked for (i.e., by one or multiple roles). Together, this can aid in building a strategic map of which error types *should* be checked by each role.

In our redevelopment project, the OCR members of the working group conducted a review of each check using the Task Descriptors,[2] recruitment criteria and the instruction documents given to assessors about the checking tasks, looking at which of the 42 error types were being targeted by each role according to these documents. This led into mapping out the intentions for the checking roles in terms of whether checkers in each role should be expected to look for and identify each type of error. The relevant working group members identified four categories to help them distinguish between different kinds of role intentions:

- Core focus – This error type is a core focus of this checking role. This means that the checker performing this role should conduct a thorough check for all errors of this type.
- Peripheral focus (high impact) – This error type is not a core part of this checking role and should not be a main focus for checkers. However, because of the nature of the role any *high* impact errors of this type should be

2 The Task Descriptor for each assessor role is a publicly available document containing a brief description of what is involved in the role. For example, the Task Descriptor for the Candidate Proxy can be found here: https://www.ocr.org.uk/Images/471234-candidate-proxy-task-descriptor.pdf.

identified if the checker performs their check correctly.

- Peripheral focus (lower impact) – This error type is not a core part of this checking role and should not be a main focus for checkers. However, because of the nature of the role it is possible (or even likely) that lower impact errors of this type would be identified if the checker performs their check correctly.
- Not expected – This error type is not intended to be detected by the checker performing this checking role.

These role intentions were recorded in a mapping grid of the kind shown in Table 1.

Table 1: An illustration of the kind of output produced from the mapping exercise (note: this example illustrates the concept and does not reflect the actual output for this project).

	Error type			
	1 – topic not on relevant syllabus	**2 – topic inappropriate for exam/ component**	**3 – item has a factual inaccuracy**	**4 – item is of an inappropriate level of demand**
Checking role 1				
Intention of role	Core focus	Core focus	Peripheral (high impact)	Peripheral (high impact)
On the checklist	Yes – explicit	Yes – explicit	Yes – explicit	Yes – implicit
Checking role 2				
Intention of role	Not expected	Peripheral (lower impact)	Core focus	Core focus
On the checklist	No	No	Yes – explicit	Yes – explicit

In addition, as can be seen in the example, the mapping grid was also used to record initial decisions about whether the error type should be addressed on the checklists being developed. For example, did it need an explicit checklist item, or could it potentially be an implicit part of a checklist item? These considerations were important because it would not be practical to ask all assessors to check for all error types in the Suto et al. (2023) taxonomy, given the high number of error types.

The mapping grid was regularly referred to during the design stage of the checklist development, as it helped the working group to make strategic and systematic decisions about what should be included in the checklists.

Design decisions about the checklist items

Many decisions had to be made about how to design the checklist items. This section draws attention to several different aspects of the checklists, focusing on ones where decisions were complex or had a particularly strong influence on other design decisions. The aspects discussed in this section relate to the type, length, content, phrasing and structure of the checklists.

The type of checklist

One fundamental decision that we needed to make was what type of checklist to have. In the literature, two main types of checklist stand out: Read-Do and Do-Confirm (Gawande, 2010), where the distinction concerns when the actions on the checklists are taken. Read-Do checklists are completed as part of the checking task, with the checklist items acting as prompts for actions and checkers marking the checklist items as complete as they go along. Do-Confirm checklists are completed after the tasks are done; checkers perform their tasks from memory and experience first and then use the checklist to confirm that all of the checklist items have been completed. Neither of these types of checklist seemed to fully reflect what checkers do in the exam paper context. Unlike in aviation or healthcare contexts, assessment checkers are not required to take direct action (i.e., they do not make any changes to the exam paper); instead, their task is to review and report on the state of the exam questions and provide recommendations. Therefore, it was important that the design of the checklists supported this different type of checklist.

A relatively simple way of achieving this was to reflect this task of reviewing and reporting in both the wording of the checklist items and in how checkers were asked to respond to the item. In line with Suto et al. (2023), we decided that checklist items for our exam paper checks should be phrased as questions. This communicates clearly what the checkers need to review the exam paper for and that they need to provide an answer to this question on the checklist. An example of the form of the checklist item is shown below:

> "Are all the answer spaces appropriate in terms of both type (e.g., lines, graph paper) and size?"

As that example shows, we also decided to phrase each checklist item as a "yes/no" question, specifically where "yes" meant that there were no problems of this kind with the exam materials. The aim of structuring all checklist items in this way was to make it easy and quick for those using the checkers' completed checklists (e.g., assessment managers) to see if any problems had been identified. Another option of "not sure or unable to say" was also provided to encourage checkers to record issues that they felt might be problems even if they were unsure.

Another debate during development was whether or not to ask checkers to record that they had checked every checklist item for each individual exam question or question part (e.g., by completing each cell of a grid showing each exam question part). This had to be considered carefully early in the redevelopment project, as it had implications for various fundamental aspects of the checklist (e.g., checklist length, structure, phrasing). Potential advantages of this "question-by-question" method were considered, which included that it could help draw the checkers' attention to each checklist item for each exam question part, in line with the "point and call" method[3] (Hikida et al., 2015), and that it

3 "Point and call" checks are used in various industrial contexts in Japan and involve use of a checklist to prompt pointing at the item to be checked and calling out its state. The method has been found to reduce error rates (Haga, Akatsuka & Shiroto, 1996, as cited in Hikida et al., 2015).

could provide more detailed data on where errors or issues occurred. Potential disadvantages were also discussed, including concerns that assessors may perceive this as added administration, and that it could result in a large matrix which might discourage assessors from actively engaging with it, especially for papers with a large number of questions and question parts (e.g., mathematics). The latter could lead to some assessors simply ticking boxes by default without actually checking each aspect carefully, negating the purpose of such a grid. The possibility of using a "question-by-question" strategy was also complicated by the fact that some checklist items were elements to be checked at the level of the whole paper rather than at the level of the question or question part. Due to these complications, it was decided not to require a record of the checking of each question or question part for each checklist item. However, it was made clear in all question-level checklist items and in accompanying instructions that all questions should be checked.

The length of each checklist

Another general factor that we considered early in the design stage was how long the checklist should be (i.e., how many checklist items). It was important to start discussing checklist length early in the process because it would affect other key design decisions such as the amount of content (i.e., how many errors and issues) that could be covered in the checklists and how to express this content in checklist items (e.g., should we have many specific items or fewer broader items?). Ultimately the lengths of the final checklists were the result of carefully considering and balancing different factors.

In the literature on safety industries' use of checklists, Gawande (2010) promotes short checklists, explaining that "a rule of thumb some use is to keep it between five and nine items, which is the limit of working memory" (p. 123) and to focus on the "killer items" – the most critical checks. It is important that checklist designers tailor guidelines to individual situations, assessing the impacts (positive and negative) that deviations from the guidelines may have on the checkers and their capability to complete their checks and any concurrent tasks they perform.

For our exam paper context, it was decided that our checklists could benefit from being longer and more comprehensive for a number of reasons relating to the context and purpose of these checklists, which differ from those in safety-critical industries. In particular, it was considered important to use the checklists as a means to ensure clarity around the remit of the checking roles, as under-specification of checking roles is a key factor in increasing the risk of error (Reason, 2013). Being flexible with regard to checklist length was deemed to be reasonable because of two other features of exam paper checks. The exam paper checks were not as strictly constrained by time limits as some of the checks for which the safety-industry advice was designed. In addition, the question paper checks were to be presented in a written document rather than having to be recalled from memory, which meant that the number of items on the checklist did not need to be constrained by memory limitations. Nevertheless, in designing the checklists, much attention was still paid to what was a reasonable number of points to expect the assessor to check for, bearing in mind the nature of their task. For example, the

Assessment Analyst's task involves careful analysis of the text and how different parts of this relate to one another. Therefore, it was considered undesirable to distract them from their focus on this task with a large number of checklist items. The final numbers of checklist items for each role ranged from 7 to 16.

Content covered across each checklist

Many design decisions were affected by views on what content the checklist should target (i.e., which paper errors). We considered many factors in our decision-making about the checklist content, including: the remit and scope of the checking role; the importance of the error for the checking role and assessment manager; and the purpose and usefulness of including a particular content point in the checklist.

A lot of this key information had been set out as part of the mapping exercise, and, therefore, the mapping grid was our starting point. As described earlier, the mapping grid distinguished between different error types in terms of their importance and function in the checking role. This facilitated the prioritisation of content, which was important given the discussions around checklist length. It was decided that, as a minimum, the checklist should include each error type that had been identified as a "core focus" of the checking role.

The mapping exercise only included the error types that had been identified by Suto et al. (2023). Therefore, it was also important to check with OCR colleagues (such as the assessment managers) whether it would be useful to include other checks in the checklist. Based on these discussions, a small number of checks were requested that did not originate from Suto et al.'s (2023) list of error types. For example, for the Candidate Proxy, Assessment Analyst and Assessment Marker roles, the assessment managers felt it was important for the first checklist item to relate directly to the key nature and purpose of the checker's task. Accordingly, for the Candidate Proxy and Assessment Analyst, the first checklist item focused on question answerability:

> "Are all the questions answerable (i.e., candidates who have studied the full course should be able to make a sensible attempt at answering each question)?"

For the Assessment Marker, the first item asked about alignment between the Candidate Proxy's answers and the question and mark scheme:

> "For each question, does the Candidate Proxy's response align with the question and the mark scheme?"

Content covered by individual checklist items

The decisions about content discussed in the previous section were about what to cover across each checklist as a whole. Decisions also needed to be made about how much content individual checklist items should cover. In practice, discussions about these decisions were often intertwined but we separate them here for readability.

One of the most important considerations with regard to the content of individual checklist items was how focused we needed each checklist item to be. Should items focus on one error type only? Again, this decision has many consequences. It would inevitably affect checklist length. If each checklist item were focused on one error type, then we would need more checklist items than if checklist items were broader in scope. However, the broader the checklist item, the less helpful it may be for the checkers, and for the assessment managers using the outputs of the checks.

In the final checklists, some checklist items were a direct representation of one error type (based on the mapping grid). This level of specificity was often chosen in cases where it was considered particularly important for checkers to focus on an error type or for assessment managers to have confirmation that the exam paper had been checked for a specific issue. For example, for error type 3 in Suto et al.'s (2023) taxonomy, "Item content factually inaccurate", a checklist item was written asking:

> "Is the subject content of all the questions factually accurate?"

In other cases, where error types were closely related, more than one error type was addressed by one checklist item. For example, error types 21 and 22 in Suto et al.'s (2023) taxonomy relate to inconsistency across the different parts of a question and inconsistency between different questions, respectively. For checking roles that were expected to check for both of those error types, one checklist item was written combining both aspects:

> "Does the paper avoid inconsistencies within and between questions (e.g., terminology, subject content)?"

Phrasing of the checklist items

The phrasing of checklist items was carefully considered at multiple times during the design stage given the importance of clearly communicating the check to the checker. As described earlier, one of these decisions was to phrase the checklist items as questions because it made the checking task and output of the task (i.e., yes, no, not sure or unable to say, not applicable) clear to the checker and to those who would need to review the completed checklists.

Another consideration was making sure that the phrasing clearly described the aspect of the exam question or paper that the checkers needed to review and make a judgement about. Guidance about preparing checklists argues that wording should be kept simple and exact, using words familiar within the context (Gawande, 2010). The project working group tried to follow this principle in the design of the checklists, and as part of consultation efforts, a staff member with relevant training conducted a review of the materials to ensure simplicity and conciseness in the language used. Phrasing decisions were not always straightforward, as there was sometimes a trade-off between simplicity and specificity. Although checklist guidelines recommend conciseness (Gawande, 2010), the guidelines and wider psychology literature on human factors in error also emphasise the risk of ambiguity and under-specification (Reason, 2013). One strategy that we used was to include additional information within parentheses for checklist items where there was potential ambiguity, for example:

> "Are all the questions answerable (i.e., candidates who have studied the full course should be able to make a sensible attempt at answering each question)?"
>
> "Are all the answer spaces appropriate in terms of both type (e.g., lines, graph paper) and size?"

For each checklist, the content and wording for each of the checklist items for that role was drafted and refined through discussion with members of the project working group, and later through wider consultation and feedback from piloting.

Structure and layout of the checklists

Given the relatively high number of checklist items for most roles (7 to 16), consideration was given to whether the checklist items within a checklist could be grouped into sections, as proposed by Degani and Wiener (1993). Some checklist items were factors to be checked in each individual exam question (e.g., whether each question is clear, unambiguous and will not cause confusion), while others were elements to be checked at the paper level (e.g., whether the exam paper avoids testing exactly the same content in more than one place within the paper). For the Assessment Marker's checklist, there were also some marking-related checks (e.g., whether the mark scheme rewards candidates for what the questions ask for). These three areas (question-level, paper-level and marking-related) provided a logical way of dividing the checklists into sections to make them more manageable for assessors to use.

Design decisions about the report forms

In the checking processes already in place before the redevelopment project, checkers completed a report form alongside conducting their checks. This contained: a table for assessor and question paper details; basic instructions for using the form; issues to check for; and a comments table for recording details of issues found. For the redevelopment, new report forms were drafted drawing on the structure and information in the pre-existing report forms. The table for assessor and question paper details and the basic instructions were revised as needed, with a general aim of ensuring consistency between the basic elements of the report forms for different roles, unless there was a good reason for there to be differences.

The checklist items replaced the points to check for listed in the previous forms. The checklist items were presented in tables with a column containing the checklist items and a separate column within which checkers should respond Y for "yes", N for "no", ? for "not sure or unable to say", or N/A for "not applicable". Where relevant, the checklist items were separated into different tables for the sections on question-level, paper-level and marking-related checks.

The comments tables were also updated, ensuring consistency between roles, where possible. A key design focus was to ensure that completed forms would provide useful information to assessment managers. In the pilot versions of the report forms for the Candidate Proxy, Assessment Marker and Assessment Analyst, the comments table included a column for recommendations to ensure that suggestions for improvement were provided alongside comments on potential issues. This was not included in the pilot report form for the Pre-exam Check, as it was initially felt to be unnecessary given the nature of the errors most likely to be identified at this late stage, which should automatically suggest the required correction. However, the experience of the pilot suggested that adding the recommendations column to the table would be useful to ensure clear information is provided by assessors to the assessment managers. Another decision to aid useability was to include a column in the comments tables for all checking roles where checkers are asked to record the checklist item to which each comment relates. The intention of this was to allow assessment managers to easily confirm that checkers had written in the comments table for each checklist item where they identified a possible issue. It was also considered to be a way to support data analysis of the kinds of issues identified during checks.

Summary and reflections

The project described in this article provides a case study of using research and guidance about checklist design to support a systematic redevelopment of the materials to support exam paper checking processes, intended to ensure well-designed assessments. For each checking role, checklists were carefully designed and integrated into forms for assessors to use when carrying out checks. Associated instructions and training materials were updated to cohere with the checklists and report forms. The establishment of a cross-department working group and wider consultation and piloting were valuable in drawing on broad expertise and ensuring usability of materials for those involved.

The materials developed in this project are now in use by assessors and assessment managers. Future review of their usefulness is planned so that they can be refined, if needed, to optimise checking processes. It is hoped that further work can explore how the data from the report forms could be used to support the analysis of errors that occur in live exam papers and, indeed, of errors that could have appeared in live exam papers but were successfully identified during checks. Feeding back to assessors on errors and "near misses" also has potential benefits. Such ongoing monitoring and feedback can help to continually improve processes and assessor expertise, thus ensuring exam paper quality, something of crucial importance to the accurate and fair assessment of learners.

Acknowledgements

We would like to thank the many OCR colleagues who contributed to the project, in particular Frances Wilson, Stephanie Wyre, Naomi Rowe, Kate Elliott, Helen O'Leary, Jane Bowen, and all the assessors who gave us feedback on the checking materials.

References

Ahmed, A., & Pollitt, A. (2007). Improving the quality of contextualized questions: an experimental investigation of focus. *Assessment in Education: Principles, Policy and Practice, 14*(2), 201–232.

Clay-Williams, R., & Colligan, L. (2015). Back to basics: checklists in aviation and healthcare. *BMJ Quality & Safety, 24*(7), 428–431.

Crisp, V. (2011). Exploring features that affect the difficulty and functioning of science exam questions for those with reading difficulties. *Irish Educational Studies, 30*(3), 323–343.

Crisp, V., & Macinska, S. (2020). Accessibility in GCSE Science exams – students' perspectives. *Research Matters: A Cambridge Assessment publication, 29*, 2–10.

Crisp, V., & Sweiry, E. (2006). Can a picture ruin a thousand words? The effects of visual resources in exam questions. *Educational Research, 48*(2), 139–154.

Degani, A., & Wiener, E. L. (1993). Cockpit checklists: concepts, design, and use. *Human Factors, 35*(2), 345–359.

Gawande, A. (2010). *The checklist manifesto: How to get things right.* Metropolitan Books.

Hikida, K., Matsuzaki, N., Yamamoto, S., Sakane, Y., Murata, S., Ogawa, M., Kusunoki, M., & Yoshimura, K. (2015). *The human error reduction effect of point and call checks on maritime training*. 7th International Conference on Emerging Trends in Engineering & Technology, pp. 157–159.

Reason, J. (2013). *A life in error: from little slips to big disasters*. Routledge.

Suto, I., & Ireland, J. (2021). Principles for minimizing errors in examination papers and other educational assessment instruments. *International Journal of Assessment Tools in Education, 8*(2), 310–325.

Suto, I., Williamson, J., Ireland, J., & Macinska, S. (2023). On reducing errors in assessment instruments. *Research Papers in Education, 38*(3), 357–377.

Appendix

Assessment Analyst checklist

1	Are all the questions answerable (i.e., candidates who have studied the full course should be able to make a sensible attempt at answering each question)?	
2	Are all the questions clear, unambiguous and without risk of confusing candidates?	
3	Is the subject content of all the questions factually accurate?*	
4	Are all the necessary visuals and resources provided (i.e., those in the question paper such as graphs, tables, images, text extracts, sources or equations, and those that are separate such as pre-release materials, inserts, etc.)?	
5	Are all the visuals and resources clear, complete, factually accurate and consistent with all other aspects of the question?	
6	Do all multiple choice and other selected response questions have the right number of response options and the right number of correct responses (e.g., none of the distractors could be interpreted as correct even if knowledge beyond the specification is used)?	
7	Does the paper avoid inconsistencies within questions (e.g., terminology, subject content)?	

*Check facts that are critical to answering the question at the appropriate level. For example, there is no need to check the accuracy of data from an original source/original research but it is important to check the attribution of sources where relevant and critical to answering the question (e.g., for history papers).

An analysis of the relationship between Secondary Checkpoint and IGCSE results

Tim Gill (Research Division)

Introduction

The Cambridge Lower Secondary programme offered by Cambridge Assessment International Education ("Cambridge") is usually studied by learners aged 11–14. It is part of the Cambridge Pathway, which is made up of four stages: Primary (for those aged 5+), Lower Secondary (11+), Upper Secondary (14+) and Advanced (16+). The idea behind the Cambridge Pathway is that each stage should enable learners to build on their development in the previous stage in a seamless manner.

For the Lower Secondary programme, assessment of performance is optional, but many schools choose to offer their students the Secondary Checkpoint assessments at the end of the programme (usually at age 14). The assessments are available in four main subjects: English (first or second language), Mathematics, Science, and Global Perspectives. At the end of Upper Secondary (aged 16), many learners take Cambridge IGCSEs, which are offered in many different subjects.

Given that there is a high level of coherence between the curricula in the two stages (Lower Secondary and Upper Secondary) and that learners following the Cambridge Pathway build on their development in the previous stage, performance in Secondary Checkpoint should be a good indicator of performance at IGCSE.

In this article we investigate the relationship between Secondary Checkpoint and IGCSE results. The focus is on two separate aspects:

- predictive validity of Secondary Checkpoint; that is, how well performance in the assessment predicts performance in IGCSE
- impact of taking Secondary Checkpoint on students' subsequent performance (at IGCSE).

Predictive validity studies

The predictive validity of an assessment is defined as how well scores in the assessment predict scores in a future assessment. There are several reasons why we might want to check the predictive validity of assessments. Firstly, the mark or grade given to students should say something about their ability in the subject.

This is important for selection purposes. Schools and universities sometimes need to select a limited number of students from a large pool and want to be confident that they are selecting the students who are most likely to succeed in the next stage of education.

Examples of predictive validity studies in relation to selection include selection of university students (e.g., Muzyamba et al., 2012; Vulperhorst et al., 2018; Shaw & Vidal Rodeiro, 2019; McManus et al., 2021) and selection of pupils to attend selective schools (e.g., Hall, 2015; Brown & Fong, 2019).

The predictive validity of grades given to students is also important for other reasons, for example setting targets for students, allocating students to sets, and identifying students who are behind their peers and therefore need extra help. Examples of predictive validity studies in these areas can be found in Strand (2006), who investigated how well the results of Key Stage 2 (KS2) tests in England predicted performance in Key Stage 3 tests taken three years later, and in Deary et al. (2007) who explored the predictive validity of the Cognitive Abilities Test[1] (CAT) in relation to GCSEs.

Predictive validity studies have also been used in the context of standard maintaining. For example, the system of standard maintaining in England depends to a high degree on the prior attainment of students,[2] so it is important to ensure that this attainment has some predictive power.[3] These studies include investigations of the use of KS2 scores to predict GCSE performance (Benton & Sutch, 2014; Treadaway, 2013) and the use of GCSE mean scores to predict AS and A level performance (Benton & Lin, 2011; Benton, 2015).

Most predictive validity studies report on simple correlations between results in the initial assessment and results in the assessment being predicted. Other studies (e.g., Wyness et al., 2022) attempt to improve predictions by including additional information on student and/or school characteristics. However, for many studies, this data is not available and so it is only possible to calculate correlations.

In terms of the results from predictive validity studies, the literature is dominated by the predictive validity of upper secondary school grades in relation to university performance. Previous research on the validity of school grades at one stage of school to predict grades at a later stage is harder to find. The research by Strand (2006) found correlations between the performance in KS2 and KS3 tests (a gap of three years) ranging from 0.52 to 0.81, depending on the subject. Similarly, Treadaway (2013) reported correlations between 0.71 and 0.90 depending on which specific measures of KS2 and KS3 test performance were used to calculate the correlations.

Treadaway (2013) also reported correlations between KS2 test performance and various measures of performance at KS4 (a gap of five years) ranging from 0.67

1 The CAT is a test of reasoning ability used in many schools to assess student potential.

2 Known as the "comparable outcomes" approach (Ofqual, 2011).

3 It is important to note that in the context of standard maintaining there is no interest in predicting grades for individuals: it is about allowing for differences in cohort ability.

to 0.73. Benton and Sutch (2014) found that the median correlation between KS2 mean test level and GCSE grade across a large number of subjects was 0.50. Sammons et al. (2014) found correlations of KS3 test scores with GCSE grades (a gap of three years) of 0.80 for English and 0.87 for maths.

The size of these correlations suggests that there was still a significant amount of variation in outcomes not accounted for by the prior attainment measure. A correlation of 0.70 means that over 50 per cent of the variation is unaccounted for.[4] However, it is worth saying that some of this variation is desirable because it means that students can make different amounts of progress. It also acknowledges the fact that there are many other factors which have an impact on future performance. If grades at one level perfectly predicted grades at the next level, then there would be no need for a second assessment!

Impact of taking a particular qualification

This article also investigates whether taking a particular qualification provides better preparation than others for future study, and therefore leads to better outcomes in the future. This is likely to be the case when the present and future qualifications are provided by the same organisation, such as Secondary Checkpoint and IGCSE qualifications provided by Cambridge. Given that IGCSE curricula are designed to cohere with Checkpoint, we might expect that students taking Checkpoint would have an advantage at IGCSE compared with students not taking Checkpoint. This approach is laid out in the Cambridge Pathway (Cambridge Assessment International Education, 2019), which claims that, at each stage, students build on their learning at the previous stage in a so-called "spiral" approach (Ireland & Mouthaan, 2020).

There is little prior research looking at the impact of taking a particular qualification, and most of this relates to the impact of taking pre-university qualifications on university outcomes. For example, in England, academic qualifications tend to lead to better outcomes than vocational qualifications (Bailey & Bekhradnia, 2007; Gill, 2018). Further, taking the Extended Project Qualification (EPQ) in addition to A levels is associated with better degree outcomes (Gill, 2022). Similarly, Shaw and Bailey (2011) found that, in the US, achieving the Advanced International Certificate of Education (AICE) Diploma offered by Cambridge was associated with significantly higher scores in the first year at university than taking the International Baccalaureate (IB).

Aim of the current research

The purpose of the research presented here was twofold. First, to investigate whether the results of Secondary Checkpoint assessments provide good predictions of IGCSE grades. Secondly, investigating whether taking Checkpoint gives candidates an advantage at IGCSE. The research questions were:

- To what extent do the results of Secondary Checkpoint assessments predict IGCSE grades?
- Did schools which offered Checkpoint tend to do better in their IGCSE results than similar schools which did not?

4 The variance accounted for is calculated by squaring the correlation coefficient, 0.7 x 0.7 = 0.49, leaving 51 per cent of the variance unaccounted for.

Predictive validity of Secondary Checkpoint

Data and methods

For the predictive validity analysis, we used results from students taking Secondary Checkpoint in 2017, matched to their IGCSE results in either 2018 or 2019. These years were chosen as they were the most recent years prior to the COVID-19 pandemic, which disrupted learning and assessments worldwide. Although the usual time gap between Secondary Checkpoint and IGCSE assessments is two years, there were also a substantial number of candidates with a gap of just one year.

The data for this work was downloaded from internal databases. This included the Checkpoint raw scores, IGCSE grades, and candidates' details, including gender, date of birth, and country.

Checkpoint raw scores are standardised so that they are comparable between different exam sessions. Standardised scores are between 0 and 6 and are rounded to the nearest 0.1, which is then reported to the candidate. We used the rounded score in all the analyses of predictive validity.

IGCSE grades ranged from A* to G, with those failing to get a grade classified as "U".[5] These were converted to numbers (A* = 8, A = 7, etc., down to U = 0) to enable the predictive validity analysis.

Table 1 shows, for each Checkpoint subject we considered, the IGCSE subjects which we used results from to assess the predictive validity. Where more than one IGCSE syllabus exists for a particular subject, we only used data from the syllabus with the most entries. As well as the individual Checkpoint subjects, we also calculated the predictive validity of a mean Checkpoint score, using the average Checkpoint score of English (or English as a Second Language (E2L)), mathematics and science. This analysis was restricted to candidates who took all three subjects.

Table 1: Progression from Checkpoint to IGCSE.

Checkpoint subject	IGCSE subject					
	English	E2L	Maths	Biology	Chemistry	Physics
English	✓		✓	✓	✓	✓
E2L		✓	✓	✓	✓	✓
Maths	✓	✓	✓			
Science				✓	✓	✓
Mean	✓	✓	✓	✓	✓	✓

We undertook two main analyses of the relationship between Checkpoint score and IGCSE grade: first, correlations between Checkpoint score and IGCSE grade, and second, mean IGCSE grades for candidates achieving each Checkpoint score band.

5 IGCSE qualifications with grades 9 to 1 were also available, but entries to these were much lower, so we decided to include only A* to G qualifications.

Matching process

To undertake the predictive validity analysis, we first needed to identify candidates who sat both Checkpoint and IGCSE assessments. This involved matching candidates using their name, gender and date of birth. Two different methods were employed for this process. The first was identifying candidates who had an exact match of name, gender and date of birth (although with an allowance for first names and surnames to be swapped over and for "middle" names to be present at one stage and missing at the other). The second method matched exactly on gender and date of birth but allowed for small differences in the name. For this, we used the SPEDIS function in SAS (Gershteyn, 2000). The results from both methods were combined and any duplicates were removed.

It is a limitation of this research that we do not know how many candidates there were who took both Checkpoint and IGCSE but were not found by the matching process. We only had the progression data for those candidates who were matched.

Results

Between approximately 20 per cent and 50 per cent of Checkpoint candidates were able to be matched to a result in an IGCSE, depending on the subject. A much smaller percentage of IGCSE candidates were matched (under 10 per cent in all subjects). This was expected, because in many countries students only start taking Cambridge International qualifications in upper secondary (i.e., at IGCSE level).

Table 2 shows the Pearson correlation coefficients between the rounded score on the Secondary Checkpoint subject and the grade achieved in various IGCSE subjects. It also shows the number of matched candidates used to calculate the correlations.

The results reveal that, as expected, the highest correlations were within the same subject. All the within-subject Pearson correlations were 0.69 or above, with the highest correlation between Checkpoint Science and IGCSE Biology (0.78). Correlations between different subjects were lower, between 0.41 and 0.57. This demonstrates that Checkpoint has a high degree of predictive validity, particularly within the same subject.

Table 2: Correlations between Secondary Checkpoint score and IGCSE grade (all subjects).

Checkpoint subject	IGCSE subject	Candidates matched	Correlation (Pearson)
English	English	7458	0.69
	Maths	9494	0.49
	Biology	6727	0.57
	Chemistry	6539	0.48
	Physics	6516	0.50
E2L	E2L	4282	0.74
	Maths	4104	0.41
	Biology	3298	0.56
	Chemistry	3336	0.46
	Physics	3453	0.49
Maths	Maths	13 311	0.75
	English	7201	0.52
	E2L	8069	0.41
Science	Biology	9607	0.78
	Chemistry	9491	0.71
	Physics	9587	0.73
Mean (of English, Maths & Science)	English	6253	0.68
	E2L	3561	0.69
	Maths	8463	0.74
	Biology	6129	0.78
	Chemistry	6000	0.69
	Physics	5974	0.73

Figure 1 compares the mean IGCSE grade for candidates achieving each Checkpoint score band, for the different IGCSE subjects. For example, for candidates achieving a Checkpoint score band of 3.0–3.9, the mean IGCSE grade in English was around 5.5 (halfway between grades C and B) compared with around 3.5 (halfway between grades D and E) in E2L. Figure 2 presents the same data for the relationship between mean Checkpoint score band and mean IGCSE grade.

In both figures, all the lines slope upwards, which demonstrates the predictive validity of Checkpoint, as mean IGCSE was higher for each higher Checkpoint score band.

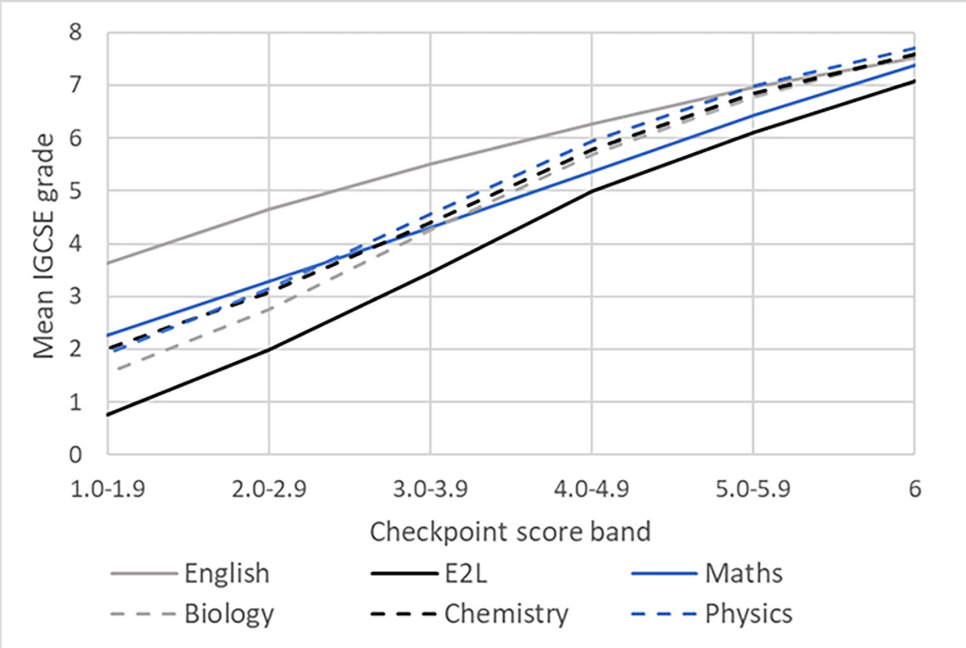


Figure 1: Mean IGCSE grade for each Checkpoint score band, by subject

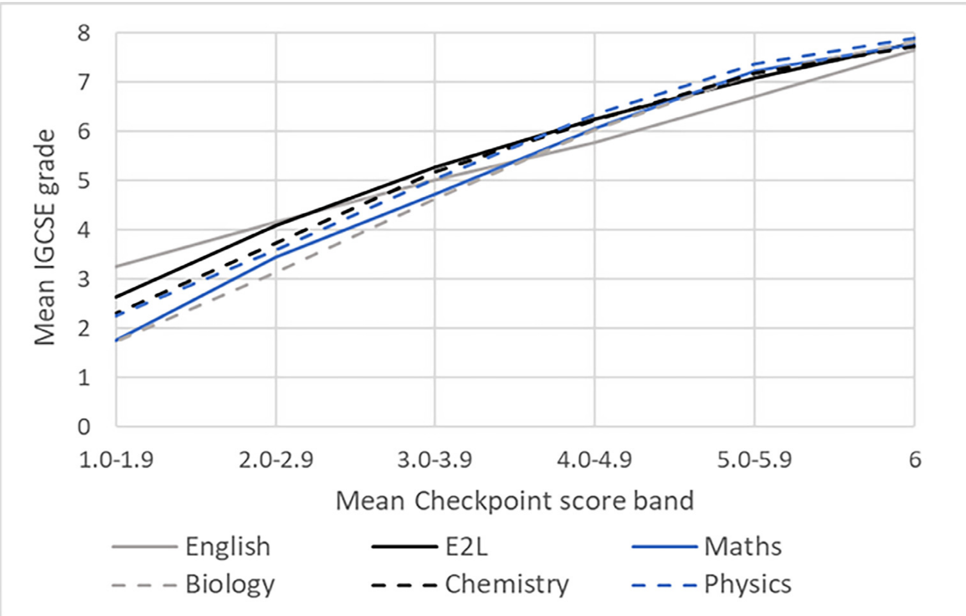


Figure 2: Mean IGCSE grade for each mean Checkpoint score band, by subject

Figure 1 shows some differences between the subjects in terms of the mean IGCSE grade for each Checkpoint score band (in the same subject). For candidates achieving Checkpoint scores of 4.0–4.9 or higher, the mean IGCSE grade was lower for E2L and Maths than for the other subjects. For those achieving a score band of 3.0–3.9 or lower, the mean IGCSE grade was highest for English and lowest for E2L. Some of the differences were substantial. For example, for candidates achieving a Checkpoint score band 2.0–2.9, the mean IGCSE grade was around 2 (grade F) in E2L and about 4.5 (between grades C and D) in English.

In Figure 2, there was very little difference in the mean IGCSE grade between subjects for candidates achieving a mean Checkpoint score of 6. For score bands 5.0–5.9 and 4.0–4.9, the mean IGCSE grade was lower for English than for other subjects. However, for score bands of 2.0–2.9 or below the mean grade was higher for English (and also for E2L). The flatter slope for English indicates a lower correlation.

Impact of adopting Checkpoint on IGCSE results

Data and methods

In this second section of the article, we investigated whether schools which offered Checkpoint tended to do better in their IGCSE results than similar schools which did not. To do this, we looked at schools which had recently adopted Secondary Checkpoint and investigated whether they improved their IGCSE results in subsequent years. We chose to use a differences-in-differences approach for this analysis. This technique is appropriate for assessing the effect of a reform or the introduction of a new programme or policy (see, for example, Abramovsky et al., 2011; Belot & Vandenberghe, 2014). The outcome variable in these types of models is the difference in an outcome measure before and after the reform or programme is introduced. Comparisons can then be made, in terms of this difference, between those exposed to the new reform/programme and those not exposed.

For this research, the "reform" was the adoption of Secondary Checkpoint. We categorise the centres adopting Checkpoint as the "treatment" group, and all other centres (that is, those not offering Checkpoint) as the "control" group. The variable of interest was the difference in the mean IGCSE grade (across all subjects in a centre) before and after the time when the centres adopted Checkpoint. We used the difference in mean IGCSE over two separate periods of two years (between 2017 and 2019 and between 2016 and 2018). Then we identified centres which did not offer Checkpoint two years before those dates (i.e., either 2014 or 2015), and split these centres into those adopting Checkpoint in one of the next two years and those not doing so. For the centres adopting Checkpoint we then had a mean IGCSE based on candidates who did not take Checkpoint (2016 or 2017 IGCSEs) and a mean IGCSE based on candidates who did (2018 or 2019 IGCSEs).

These years were chosen as the most recent results not to be disrupted by the COVID-19 pandemic. The time gap of two years was thought to be appropriate in this context because it was enough time for some schools to have a significant proportion of their candidates taking Checkpoint, but it was short enough that most centres would not have had many other changes likely to affect performance. The data from the two sets of years was combined, in order to increase the number of centres in the treatment group, which otherwise would have been too low for a meaningful analysis.

The general form of the model was:

$$\Delta y_j = (y2_j - y1_j) = \beta_0 + \beta_1 X1_j + \beta_2 X2_j + \cdots + \beta_m Xm_j + u_j$$

where Δy_j is the change in mean IGCSE between year 1 (either 2016 or 2017) and year 2 (either 2018 or 2019) in school j, $X1$ to Xm are the independent variables (including whether the centre adopted Checkpoint, and some contextual variables), β_1 to βm are the regression coefficients and u_j is the residual.

As well as the indication of whether the centre adopted Checkpoint, three other contextual variables were included in the models, because we thought they might have a significant effect on the outcome variable (change in IGCSE performance). These were the mean IGCSE in the centre in year 1, the country in which the centre was located, and the pair of years that the data came from (either 2016–18 or 2017–19).

Undertaking the analysis at centre level meant there were some drawbacks. It was not simply a case of selecting all centres which adopted Checkpoint and seeing if they improved their IGCSE results, for two main reasons. We needed to consider: 1) the proportion of IGCSE candidates in each centre who actually took Checkpoint qualifications, and 2) the number of different Checkpoint subjects they took on average. Any impact of taking Checkpoint on IGCSE results at a centre level was likely to be much less if only a small proportion of candidates took Checkpoint, or if they only took one subject.

To account for the first of these two issues, we only counted centres as being in the treatment group when the proportion of their IGCSE candidates who were matched to Checkpoint results was at least 50 per cent. Centres in the control group were those with zero candidates taking Checkpoint. Centres with between 0 per cent and 50 per cent of matched candidates were excluded from the analysis. To calculate this percentage, we needed to match candidates between Checkpoint and IGCSE, using names, gender, and date of birth. As with the predictive validity analysis described earlier in this article, this was a shortcoming, and there may have been some schools where the percentage of IGCSE students who took Checkpoint was 50 per cent or greater, but the matching process only picked up fewer than 50 per cent.

To take account of the second issue, we then categorised centres by the mean number of Checkpoint subjects taken by their students (in year 1). Centres in the treatment group were categorised into two groups: those where Checkpoint candidates took an average of fewer than 2 Checkpoint subjects ("treatment 1"); and those where Checkpoint candidates took an average of 2 or more Checkpoint subjects ("treatment 2"). We decided to run two sets of regression models: firstly, with a binary variable indicating whether the school adopted Checkpoint; and second, with a three-way grouping of centres into control, treatment 1 or treatment 2.

Additionally, we did not include centres where the number of IGCSE candidates was very different between year 1 and year 2, because this might have had a big impact on results in the centre. Therefore, we excluded any centres where the number of candidates in one year was more than 1.5 times the number in the other year. Finally, we removed centres with fewer than 30 IGCSE candidates in either year, because results in small centres were more likely to be volatile between years.

Results

There were 1035 centres which did not offer any Checkpoint qualifications in the base years (and fulfilled the criteria described in the previous section), 35 of which started offering it in the next two years with the remaining 1000 continuing not to offer it. The summary statistics for the difference in mean IGCSE in the two groups of centres are in Table 3.

Table 3: Summary statistics for the difference in mean IGCSE, by 2-level treatment group (y2-y1[6]).

Centre group	N	Mean	S.D.	Min	Max
Control	1000	-0.013	0.512	-1.978	2.345
Treatment	35	0.244	0.523	-0.659	1.634
All	**1035**	**-0.004**	**0.514**	**-1.978**	**2.345**

This shows that, overall, there was almost no difference in mean IGCSE between year 1 and year 2. However, there was a substantial difference in the mean IGCSE difference between the control and treatment groups. On average, schools adopting Checkpoint improved their mean IGCSE performance by a quarter of a grade, equivalent to an increase in one grade for every fourth IGCSE. Centres not adopting Checkpoint had almost no difference in their mean IGCSE.

Figure 3 plots the mean IGCSE in year 1 against the mean IGCSE in year 2 for each centre, with different symbols for centres adopting Checkpoint. The figure also shows lines of best fit for the two groups of centres. The first thing to note is that the dots are mostly clustered around the line of equality (not shown), meaning that most centres only had small changes in their mean IGCSE between year 1 and year 2.

The line of best fit is higher for Checkpoint centres (in black colour), which is consistent with these centres improving more than non-Checkpoint centres, on average. It can also be seen that the difference between both lines of best fit was slightly smaller for centres with higher values of mean IGCSE in year 1.

6 Y1 = 2016 and Y2 = 2018, or Y1 = 2017 and Y2 = 2019.

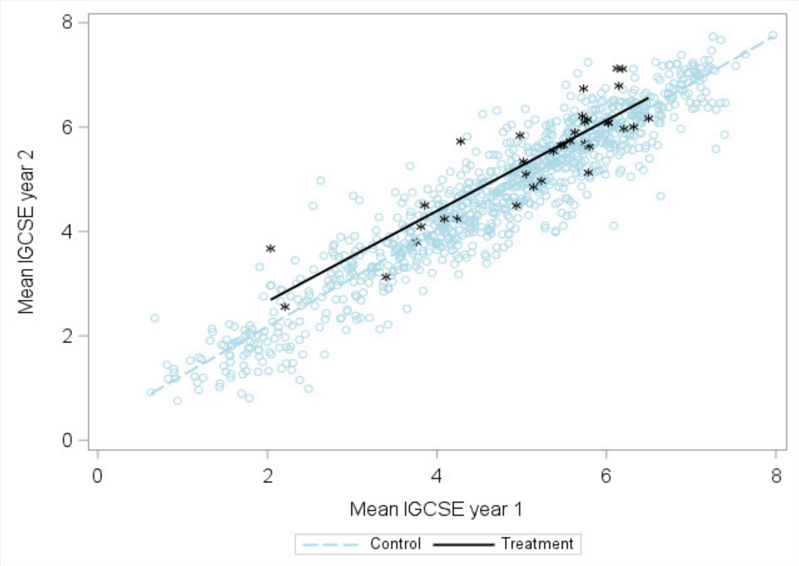


Figure 3: Mean IGCSE in Y1 and Y2, split by treatment group

Next, as discussed in the methods section, the treatment group of centres was split into two, based on the mean number of Checkpoint subjects taken by candidates in the centre. There were 9 centres where the mean number of Checkpoint subjects was less than 2 (treatment 1), and 26 where it was 2 or more (treatment 2). Table 4 shows summary statistics for the difference in mean IGCSE in the three groups.

Table 4: Summary statistics for difference in mean IGCSE, by 2-level treatment group (y2-y1[7]).

Centre group	N	Mean	S.D.	Min	Max
Control	1000	-0.013	0.512	-1.978	2.345
Treatment 1	9	0.017	0.281	-0.447	0.363
Treatment 2	26	0.322	0.567	-0.659	1.634
All	**1035**	**-0.004**	**0.514**	**-1.978**	**2.345**

This shows that the centres with a mean number of Checkpoint subjects lower than 2 ("treatment 1") had almost no improvement in mean IGCSE, whereas centres with a mean of 2 or more ("treatment 2") improved by 0.322 of a grade. This suggests that adopting Checkpoint was more beneficial for centres with candidates taking at least 2 Checkpoint subjects.

Figure 4 plots the mean IGCSE in year 1 against year 2, with different symbols for centres in each treatment group. This shows that the line of best fit is highest for centres in treatment 2 and the lines of best fit for treatment 1 and the control group are very similar. Furthermore, this figure shows that the increase in IGCSE performance for schools in treatment 2 was lower at higher values of year 1 mean.

7 Y1 = 2016 and Y2 = 2018, or Y1 = 2017 and Y2 = 2019

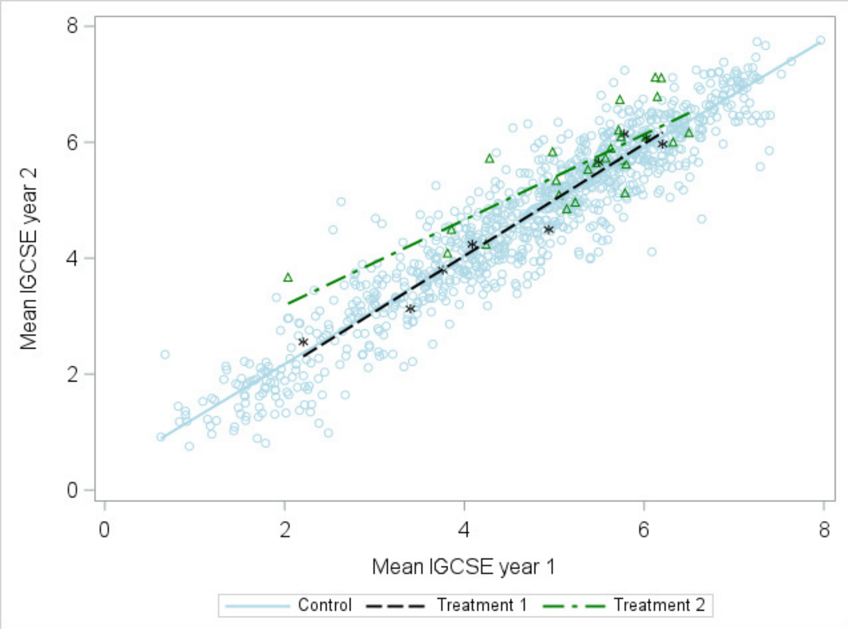


Figure 4: Mean IGCSE in Y1 and Y2, split by three-level treatment group

Regression analysis

Table 5 shows the results (excluding country effects) of the regression model with a binary indicator of whether the centre adopted Checkpoint. Table 6 shows the results for a model with a three-level treatment variable (control, treatment 1, treatment 2).

These results show that schools adopting Checkpoint had a significantly greater improvement in mean IGCSE than schools not doing so. Overall, this advantage was around one fifth of a grade. However, Table 6 reveals that the significant difference was only present in schools where the mean number of Checkpoint subjects taken by candidates was 2 or more (treatment 2), where it amounted to more than a quarter of a grade. This is equivalent to one grade in every fourth IGCSE taken in a centre.

The other significant variables in both models were the mean IGCSE in year 1 and the country. An increase in mean IGCSE in year 1 was associated with a significantly worse outcome (by 0.18 of a grade for a one grade increase). In other words, centres with high mean IGCSE in year 1 were less likely to have improved their mean IGCSE by year 2. There were significant differences between countries, but these are not reported here, because of the large number of different countries.

Figures 3 and 4 suggested that the advantage for Checkpoint centres might be less at higher values of mean IGCSE in year 1. To explore this further, we included an interaction term between mean IGCSE in year 1 and treatment group in the models, but this was not significant.

Table 5: Regression coefficients difference in mean IGCSE (binary indicator of treatment group).

Variable		Estimate	Std. error	t-value	Pr(>|z|)
(Intercept)		0.742	0.122	6.09	<0.001
Centre group	Control				
	Treatment	0.203	0.089	2.28	0.023
Mean IGCSE in year 1		-0.179	0.016	-11.31	<0.001
Country		**			

Table 6: Regression coefficients difference in mean IGCSE (three-level indicator of treatment group).

Variable		Estimate	Std. error	t-value	Pr(>|z|)
(Intercept)		0.742	0.122	6.09	<0.001
Centre group	Control				
	Treatment 1	-0.054	0.188	-0.29	0.772
	Treatment 2	0.275	0.100	2.74	0.006
Mean IGCSE in year 1		-0.179	0.016	-11.31	<0.001
Country		**			

Discussion

Predictive validity of Checkpoint

The results of the predictive validity analysis showed that there was a strong association between Checkpoint scores and IGCSE grades, particularly in the same subject. Correlations within subject varied between 0.69 and 0.78, which were very similar to figures from previous research (in a UK context) looking at the relationship between scores at different educational stages (e.g., Strand, 2006; Sutch, 2013; Sammons et al., 2014; Carroll & Gill, 2023). These figures suggest that Checkpoint scores provide useful information on how well students are likely to do in their IGCSEs. This information can be used by schools to help with target setting, streaming, and identifying students who may need extra help.

The results of this research also showed that candidates with the same (high) Checkpoint score achieved the highest grades in the science subjects at IGCSE, and the lowest grades in E2L. Similarly, we found that, for the same (high) mean Checkpoint score, it was easier to achieve a high IGCSE grade in Mathematics or the Sciences than in English or E2L. There were some large differences between subjects, which could be an indication that standards were not entirely aligned between subjects, either at Checkpoint or at IGCSE. However, we should be careful not to assume that this was definitely the case, as this pattern could also be due to the lower correlation between mean Checkpoint score and IGCSE grade in English and E2L.

Impact of adopting Checkpoint on IGCSE results

Schools adopting Checkpoint had a significant improvement in their mean IGCSE, compared to schools not adopting Checkpoint. The difference amounted to a quarter of a grade (for centres where the mean number of Checkpoint subjects taken was at least 2). While this is not a large effect, it is of practical significance. For some students it could mean the difference between progressing to A levels or not.

This is perhaps not a surprising conclusion given that Checkpoint and IGCSEs are both offered by the same awarding organisation and IGCSE curricula are designed to cohere with Checkpoint. There may be similar effects for schools that choose to adopt consecutive qualifications from other awarding organisations.

Some caution is required when interpreting these findings. We have evidence of an effect of offering Checkpoint on IGCSE performance at the school level, but we do not know whether there were any other factors which we were unable to account for, but which were important in determining the difference in mean IGCSE. Unfortunately, we had very little information on each school (we only had the number of students at the school, their mean IGCSE in year 1 and the country). Many other factors can affect how a school performs, including the ability of the students, the effectiveness of the school leadership team and the teachers and the resources available to them. We have had to assume, for the purpose of this analysis, that these factors did not change over the two-year period that we looked at. Further research on this topic would be interesting if it was possible to acquire more information about schools and include this in the analysis.

References

Abramovsky, L., Battistin, E., Fitzsimons, E., Goodman, A., & Simpson, H. (2011). Providing employers with incentives to train low-skilled workers: evidence from the UK employer training pilots. *Journal of Labor Economics, 29*(1), 153–193.

Bailey, N., & B. Bekhradnia. (2007). *The academic experience and outcomes of students with vocational Level 3 qualifications*. Oxford: HEPI.

Belot, M., & Vandenberghe, V. (2014). Evaluating the 'threat' effects of grade repetition: exploiting the 2001 reform by the French-Speaking Community of Belgium. *Education Economics*, *22*(1), 73–89.

Benton, T. (2015). Can we do better than using 'mean GCSE grade' to predict future outcomes? An evaluation of Generalised Boosting Models. *Oxford Review of Education*, *41*(5), 587–607.

Benton, T., & Lin, Y. (2011). *Investigating the relationship between A level results and prior attainment at GCSE*. Coventry: Ofqual.

Benton, T., & Sutch, T. (2014). *Analysis of the use of Key Stage 2 data in GCSE predictions*. Coventry: Ofqual.

Brown, A., & Fong, S. (2019). How valid are 11-plus tests? Evidence from Kent. *British Educational Research Journal, 45*(6), 1235–1254.

Cambridge Assessment International Education. (2019). *An international education from Cambridge*. Cambridge, UK: Cambridge Assessment International Education.

Carroll, M., & Gill, T. (2023). *Progression from GCSE to A Level, 2018 – 2020. Statistics Report Series No. 129.* Cambridge University Press & Assessment.

Deary, I., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*(1), 13–21.

Gershteyn, Y. (2000). Use of SPEDIS function in finding specific values.

Gill, T. (2018). Preparing students for university study: a statistical comparison of different post-16 qualifications, *Research Papers in Education*, *33*(3), 301–319.

Gill, T. (2022). *Are students who take the Extended Project Qualification better prepared for higher education study?* Cambridge University Press & Assessment.

Hall, M. T. (2015). An examination into the validity of secondary school entrance scores in predicting the academic success of secondary aged students. *Current Issues in Education, 18*(1).

Ireland, J., & Mouthaan, M. (2020). Perspectives on curriculum design: comparing the spiral and the network models. *Research Matters: A Cambridge Assessment publication*, *30*, 7–12.

McManus, I. C., Woolf, K., Harrison, D., Tiffin, P. A., Paton, L. W., Cheung, K. Y. F., & Smith, D. T. (2021). Predictive validity of A-level grades and teacher-predicted grades in UK medical school applicants: a retrospective analysis of administrative data in a time of COVID-19. *BMJ Open, 11*(12).

Muzyamba, M. C., Goode, N., Kilyon, M., & Brodbelt, D. (2012). Predictors of success in a UK veterinary medical undergraduate course. *Journal of Veterinary Medical Education*, *39*(4), 380–8.

Ofqual. (2011). *Maintaining standards in GCSEs and A levels in summer 2011*.

Sammons, P., Sylva, K., Melhuish, E., Siraj, I., Taggart, B., Toth, K., & Smees, R. (2014). *Influences on students' GCSE attainment and progress at age 16*. London, UK: Institute of Education, University of London.

Shaw, S., & Bailey, C. (2011). Success in the US: are Cambridge International assessments good preparation for university study? *Journal of College Admission*, *213*.

Shaw, S., & Vidal Rodeiro, C. (2019). The value of predictive validity studies and the need for 'fit-for-purpose' data to inform postsecondary admissions policies and decision making. *Strategic Enrollment Management Quarterly, 6*(4), 23–39.

Strand, S. (2006). Comparing the predictive validity of reasoning tests and national end of Key Stage 2 tests: which tests are the 'best'? *British Educational Research Journal*, *32*(2), 209–225.

Sutch, T. (2013). *Progression from GCSE to AS and A level, 2010*. *Statistics Report Series No. 69*. Cambridge, UK: Cambridge Assessment.

Treadaway, M. (2013). *An analysis of Key Stage 2 reliability and validity. FFT Research Paper No. 2*. FFT Education Ltd.

Vulperhorst, J., Lutz, C., de Klein, R., & van Tartwijk, J. (2018). Disentangling the predictive validity of high school grades for academic success in university. *Assessment & Evaluation in Higher Education*, *43*(3), 399–414.

Wyness, G., Macmillan, L., Anders, J., & Dilnot, C. (2022). Grade expectations: how well can past performance predict future grades? *Education Economics, 31*(4).

Synchronous hybrid teaching: how easy is it for schools to implement?

Filio Constantinou (Research Division)

Introduction

In recent years, there have been various calls to make teaching and learning spaces more flexible, to allow them to better cater for the needs of the increasingly diverse student population (see e.g., Raes, 2022; Wang et al., 2018). One such flexible space is that created by the concurrent combination of in-person and online instruction, allowing both on-site and remote students to attend lessons simultaneously. This merging of modalities, known as "synchronous hybrid learning" (Raes et al., 2020), "blended hybrid model" (Bartlett, 2022), "concurrent classroom" (Ladd, 2020), "synchromodal teaching" (Bell et al., 2014) or "dual-mode teaching" (Centre for the Enhancement of Teaching and Learning, 2020), is more commonly encountered in higher education. Its introduction may represent, to some extent, an attempt to respond to the declining number of students enrolling in traditional, in-person university programmes following an increase in the offering of distance-learning ones. This hybrid form of instruction, or "synchronous hybrid teaching" (henceforth SHT), provides higher flexibility and, as such, can be particularly attractive to learners who are given the option to attend lectures either remotely or in person depending on their personal and/or professional circumstances and commitments (Bower et al., 2014; Gosper et al., 2010). However, despite its appealing nature, SHT is still an "emerging practice" in higher education, with research in this area being "in its infancy" (Raes et al., 2020, p. 286).

Research in SHT is even more limited in the context of primary and secondary education, where SHT is a much less frequently occurring or discussed mode of teaching. The first time SHT surfaced as an instructional possibility in primary and secondary schools internationally was the COVID-19 pandemic. During the pandemic, infected students and their close school contacts were required to stay at home for a period of time to help reduce the spread of the virus. To minimise the disruption to these students' learning, many schools around the world attempted to implement SHT. Following this experience, SHT has been viewed by many as one strategy that schools could adopt post-pandemic to make learning more flexible and more inclusive (see e.g., International House World Organisation, 2020; Joshi, 2023; Weller, 2021). For example, with SHT, ill students, students with mobility issues, as well as international students needing to spend some time in their home country, would still be able to attend classes. This would, in turn, help to mitigate the learning loss they would otherwise experience.

While the potential benefits of employing SHT are unquestionable, the ease with which this mode of teaching can be implemented in primary and secondary education is not sufficiently understood. To help examine this issue, this study interviewed a number of primary and secondary teachers about their experience of using SHT in schools in different parts of Europe during the COVID-19 pandemic. The study illuminates various challenges confronted by teachers and students during SHT, thereby identifying a number of important obstacles that need to be overcome for SHT to be smoothly implemented in schools and for it to function as an effective instructional strategy.

Method

This research was part of a larger mixed-methods project which sought to record and understand teachers' experiences of planning and delivering lessons during the COVID-19 pandemic. The project involved a questionnaire completed by teachers from around the world, and follow-up in-depth interviews with 13 of the questionnaire respondents, all based in different parts of Europe (for more details about the project's methodology and findings, see Carroll & Constantinou, 2022, 2023). The present study drew upon the experiences of the interviewees, specifically 12 of them (as one did not engage in SHT). As shown in Table 1, the 12 interviewees represented a diverse group: they were based in different countries, worked in different education sectors (primary and secondary), had different roles within their school, taught different subjects, and their teaching experience ranged from 6 to 35 years. It is worth noting that all teachers interviewed were based in schools in Europe, while the majority of them worked in the private sector. This may have restricted, to some extent, the range of experiences captured through the interviews. It is likely that SHT outside of Europe and in the state sector may have manifested itself somewhat differently.

The interviews were semi-structured and lasted approximately 90 minutes each. They were conducted online in June and July 2021, and were used to collect more in-depth information about how teachers experienced the COVID-19 disruption. During the interviews, the teachers were invited to describe and reflect on their experiences of teaching during the pandemic, including those concerning SHT. Written informed consent was obtained from all interviewees (see BERA, 2018).

The interview transcripts were analysed thematically (Braun & Clarke, 2021) in MAXQDA (VERBI Software, 2021). The analysis centred around the challenges involved in implementing SHT. Overall, four different types of challenges were identified, all of which are exemplified below through relevant interview extracts.

Table 1: Interview participants (N=12).

Characteristics		N
School location	UK	6
	Cyprus	1
	Italy	2
	Romania	1
	Spain	1
	Switzerland	1
Education sector	Primary	2
	Secondary	10
School type	State-funded	3
	Private	9
Gender	Female	7
	Male	5
Position in the school	Teacher with a leadership role (e.g., head of department)	7
	Teacher without a leadership role	5
Subject area*	Creative subjects (e.g., art, design and technology, music)	2
	Humanities and Social Sciences (e.g., English language, literature, history)	5
	Science and mathematics	3

*This category concerns only the secondary teachers (the primary teachers taught all subjects).

Findings

The interviewees' experience of SHT seemed to be overall negative, hence the focus of the analysis on the challenges involved in using this instructional mode. In general, the participants found SHT particularly demanding and did not think it had led to high-quality learning for all students. In fact, they described it as the worst of the three types of teaching they employed during the pandemic (the other two being fully remote teaching, and socially distanced fully in-person teaching):

> "It's the worst of the choices, hybrid is the worst..."
>
> "Full online teaching is definitely preferable to hybrid which is a nightmare."
>
> "Hybrid teaching is much more difficult than all one [fully remote teaching] or all the other [fully in-person teaching]."
>
> "Not have hybrid – you either are at school or not at school. That's it. You can't have this two-way thing, it's just horrible."

The analysis identified four different types of challenges experienced by teachers during SHT: (a) co-ordination challenges, (b) administrative challenges, (c) interaction challenges, and (d) engagement challenges. These are described and exemplified below.

Co-ordination challenges

Teachers described SHT as a "juggling act". Operating simultaneously in two different instructional modalities felt like being "pulled in two different directions":

> "And I don't think any teacher, including myself, really succeeded or thrived in those conditions, because it really was a juggling act in terms of you had maybe 18 people in the classroom demanding your attention, and then you had to keep the 2 or 3 online included as well. So, it was a really poor second best in terms of delivering teaching to the learners."

> "Teaching entirely online obviously is far from ideal, but you can focus entirely on one thing. Hybrid teaching was where I felt most pulled in two different directions."

Co-ordinating on-site and remote learning activity proved particularly challenging in practical subjects like music, as well as in subjects entailing a strong conversational component like English:

> "A lot of the time that we've been open we've been teaching classes with half the class present in the room and also teaching with pupils online, which is really difficult, especially in a practical subject like music, to try and get anything meaningful happening in both places. That is the biggest challenge."

> "I've never had more than 75 per cent in the classroom for this whole academic year. I've always had at least 25 per cent simultaneously learning remotely, so that's been very difficult in terms of – you know, as an English teacher, in terms of classroom discussion, debate, it's been quite hard to manage."

Administrative challenges

Teaching concurrently on-site and remote students seemed to place considerable administrative demands on teachers. Examples of administrative tasks that had to be carried out by teachers using SHT included setting up the technology in the classroom at the beginning of the lesson, uploading relevant resources on the online platform for the benefit of the remote students, and marking work which had been submitted online:

> "And it was just practical elements. Our lessons are very short. Our single lessons are only 35 minutes. So, in terms of getting in in the

> beginning of the lesson to a classroom, setting up all the technology, logging on to Teams, taking the register, administratively at the beginning of lessons, it's very, very time-consuming, so that was quite tough."

> "If they're hybrid, that's really, really hard, because you have to upload all the work for the online students, you also have to download the work for the online students, you have to mark on the screen, and all that sort of stuff."

The communication with the remote students before and after the lesson, often involving multiple email exchanges, also proved administratively demanding and time-consuming:

> "It's complicated, and also the extra work to make sure that you send all the work to her [remote student] in advance, then receiving millions of emails, of course, because she's not with you face-to-face to ask you those questions. So, when you send the PowerPoint and the activities, all the millions of emails, 'I don't understand this activity' – normal things that when you are in the classroom one-to-one, it's easier."

Interaction challenges

Another aspect of teaching and learning that was negatively impacted during SHT concerned the quality and quantity of interactions during the lesson. The interactions most affected were those between on-site and remote students, as well as those between the teacher and remote students. The analysis identified three factors, or types of constraints, that seemed to restrict and undermine these interactions: (a) sound-related constraints, (b) visual constraints, and (c) cognitive constraints.

Sound-related constraints

Technical difficulties often caused delays in the verbal interaction between on-site and remote students. These delays resulted in remote students not having as many opportunities as their on-site classmates to speak and to actively participate in the lesson:

> "They [remote students] often had quite patchy wi-fi connections, so if you asked them questions there was a time lag. Getting them involved was really difficult and often as a teacher you ended up just saying 'Look, just listen and follow us as best as you can because it's just not working in terms of including you in the lessons, no matter how much we try it'. So, that was my general experience of having learners both at home and in the classroom."

Apart from the issue of time lag, there was also the problem of poor sound quality which deprived remote students of further participation opportunities such as that of "reading out loud" in English lessons:

> "Things like reading, in English, we do a lot of reading out loud so it's difficult for them to do reading out loud, because again, the [on-site] students can't necessarily hear them very well. So, I think it was quite an isolating experience for those students, but I think it was probably better than them not joining at all."

However, it was not just the on-site students who had difficulty hearing their remote classmates. The reverse problem also occurred. Remote students often had difficulty hearing what was being said in the classroom by their peers and/or the teacher:

> "The sound quality isn't always that good. I mean, I've got a very loud voice, I'm a teacher, you know, but the discussions that we have – they [remote students] can't necessarily hear the other students' answers, especially if those students are quite quiet."

> "When I listen back to my own recordings, the quality is awful. One of the main things that I really noticed is that in terms of the sound, if – and I've got quite a good laptop with quite a good microphone, but if I move really at all, even if I stand up or if I move at all around the classroom even to write on the board, the quality of the sound is very, very poor. So, it means that the remote learners can't really hear the teacher speaking, if the teacher is not sitting directly in front of the computer at all times."

Remote students' inability to hear well on-site students' verbal contributions resulted in teachers repeating them, thereby compromising the flow and quality of classroom discussions:

> "It led to quite an unnatural way of working on my part, where we teachers found ourselves almost having to repeat everything that the students in class said, so that the online ones could hear."

> "It was very, very unnatural, and I think it [having to repeat things] has certainly curtailed the discussions and the quality of discussion that we would usually have."

Visual constraints

The quality of SHT was further compromised by various visual constraints. For example, remote students often struggled to read what was written on the whiteboard, or had difficulty seeing the teacher and generally what was taking place in the classroom:

> "If you're writing on the whiteboard, it's not always that clear using the camera, what exactly you're writing."

> "I think the quality of understanding and of seeing what's happening in the classroom was poor for the remote learners."

To help alleviate the problem of visual accessibility faced by remote students, some teachers sat in front of their laptop and avoided moving around the classroom. However, this rendered the lesson more static and less interactive, while attaching a "lecture-style" character to it:

> "When we had learners online and learners in the classroom at the same time, you were kind of shackled to the computer screen as well because you had people who were watching you through the camera onscreen, so it was difficult to move away from the screen at the same time. So, all of those things meant that all of the best practice in teaching quickly reverted to teaching from the front in almost a lecture-style approach."

> "What it took away was the animation of me moving around, I'm just sitting in front of a laptop and using the laptop as a remote device ... So, it makes for a very static lesson."

Another visual constraint related to teachers' difficulty, or inability, to see remote students clearly and use visual cues, such as students' body language, to assess their level of engagement in the lesson:

> "The most difficult part of hybrid teaching is not being able to just do the informal assessment, of reading [remote] students' faces, their body language, their level of attention, their level of engagement. You can't read that when it's hybrid, it's harder."

Cognitive constraints

SHT can prove particularly attentionally demanding for teachers, as it requires simultaneously attending to, and managing, two different groups of students which are not equally visible and have distinct circumstances and needs. The considerable cognitive challenge that this process entails sometimes resulted in teachers losing sight of the remote students who were often fewer in number and therefore easier to be overlooked or to be "forgotten". This tended to restrict even further the interaction between remote students and the teacher:

> "I think completely online teaching is better than hybrid, because your whole concentration is fully on that. With hybrid teaching, a majority of kids are in the class. They are there right in front of you, right? Then you have these three or four kids who are online. There's nothing wrong with them, they're just online. I do the same [art] demonstration to the whole class, so both sides can see it. However, I'm then going to go straight to the people in my class, I'm going to ask people 'Any questions? No? OK. All right, anybody else? Any questions?' 'Yes, we've got this thing' 'OK'. Dealing with them, dealing with that. Then I'm going to go around and have a look to see what they're doing. But the people online, unless they put their camera on, because there might be a problem or they can't do the thing, by the time I've got around

24 people or whatever it is, there's not much time, 45 minutes, there's not much time to really deal with them. So, I would say my teaching in hybrid of the people online isn't great. Usually what happens is I end up saying at the end of the lesson 'Hi, guys, sorry about you guys, I forgot about you'."

> "I spoke to one of the kids I have the other day who's been online all year and he's like 'Oh, Miss, I like being online because I can get up late in the morning and I can do what I want and the teachers sort of forget about me a bit and that's quite nice because then I don't have to work so hard,' and things like that."

> "The challenge was to remember that I have to connect with the [remote] student, because you arrive to the lesson and you start, and you forget about that student, that was challenging."

Engagement challenges

Another set of challenges that tended to compromise the quality of SHT pertained to student engagement. As many of the interviewees noted, remote students seemed to be less engaged in the lesson than their on-site counterparts:

> "And the other one, I think, is about motivation, because I feel that students being at home are less motivated, less engaged, less willing to take part in the lessons."

In some cases, this lower motivation and engagement may have resulted from practical obstacles, such as attending the lesson from a different time zone and/or while surrounded by family members or other potential sources of distraction:

> "Because we have a lot of international students, many of them were still at home and hadn't travelled back for particular reasons. So, there have been some pupils that have learnt online all of this time since we've been open again. But from Malaysia, Singapore, like they're learning at different times of night, so a lot of them they were showing me out the window of their house and it's like 11 o'clock at night and they're in a lesson here in the middle of the day, so it's been really tricky for students learning remotely."

> "I think it's not the same when you are at home and with your brother, your sister, or something."

Some of the remote students engaged in what one interviewee described as "ghosting", exploiting probably the invisibility granted to them by the remote nature of their attendance. Ghosting involved logging into the online session to give the teacher the impression of attendance and then engaging in a different activity that was unrelated to the lesson:

> "We had quite a lot of ghosting – you know, where they logged on in the beginning and then they went and did something else, and then they logged off at the end of the lesson – and that was quite hard to track … We call it ghosting yeah, where you kind of sign on and then you go off and watch Netflix, and then you come back 40 minutes later and you go 'I'm still here.' That was quite hard to police."

Discussion

SHT, which was employed by many schools around the world during the COVID-19 pandemic to mitigate learning loss for students who had to stay at home, drew attention to its affordances and to the possibility of it being used as an instructional strategy post-pandemic to render lessons more inclusive. To explore the feasibility of this proposal, this study drew upon the experiences of primary and secondary teachers who used SHT during the pandemic while working in schools in different parts of Europe. Its aim was to develop a better understanding of SHT and to gain insight into schools' readiness to implement it. Overall, the findings of the study suggest that SHT, albeit providing students with more flexibility, is a demanding mode of teaching, one involving four different types of challenges: (a) co-ordination challenges, (b) administrative challenges, (c) interaction challenges, and (d) engagement challenges.

Synchronous hybrid teaching: a socio-technical process

SHT emerged as a socio-technical process, one shaped by the interplay between the social and the technological environment. While the social environment, including teachers, students and their characteristics (e.g., teachers' and students' competence in using the technology), certainly affected how technology was used in hybrid lessons, it was the technological infrastructure with its various inherent limitations as well as failures that seems to have been more impactful, affecting the social environment in decisive ways. The study provided various examples of how "the technological" influenced "the social", such as:

- Technical difficulties caused by malfunctioning or inadequate technological infrastructure (e.g., sound delays, poor sound quality) prevented remote students from fully and actively participating in the lesson, while also curtailing classroom discussion.
- The lack of additional cameras in the classroom forced teachers to sit in front of their laptop to ensure that remote students could see them, which in turn rendered the lesson more static.
- Setting up the technology at the beginning of the lesson, uploading all relevant learning resources and managing the email communication with remote students after the lesson, increased teacher workload.
- Co-ordinating interactions and activities across two different media increased cognitive load, resulting in teachers experiencing "hyper-zoom" or "hyper-focus" (Raes et al., 2020; Zydney et al., 2019).
- The nature of remote, computer-mediated communication enabled "ghosting", while also depriving teachers of the opportunity to access remote students' body language and use it to assess their level of engagement in the lesson.

Many of the challenges described by teachers seem to be similar not only to those encountered in SHT in higher education (see e.g., Bower et al., 2014; Raes et al., 2020), but also to those experienced in hybrid meetings in the workplace (see e.g., Saatçi et al., 2019). These commonalities arguably suggest that "synchronous hybridity" constitutes a distinctive communication modality which is socio-technical in nature and is characterised by its own set of challenges.

Is synchronous hybrid teaching a genuinely inclusive mode of instruction?

One of the most appealing features of SHT is the flexibility it can provide. With SHT, students who would otherwise miss school (e.g., students with certain health conditions, mobility issues and/or family circumstances) are still able to attend lessons (provided, of course, that they have access to at least an electronic device and an internet connection, which might not be the case in less affluent contexts).

Even though it can afford flexibility, SHT may often struggle to operate as an effective inclusion strategy. This is due to its inability to consistently provide the two groups of students involved, that is, on-site and remote ones, with comparable learning opportunities and experiences. In particular, the findings of the study point to an asymmetrical relationship between on-site and remote students, with the latter having fewer opportunities to actively participate in the lesson than the former. As the interviewees noted, remote students had overall fewer opportunities to contribute to classroom discussions, had more difficulty following the lesson (e.g., difficulty seeing what was written on the whiteboard; difficulty hearing what was being said in the classroom), received less attention from the teacher who was "pulled in two different directions" and was more likely to attend to the needs of on-site students, received less personalised feedback during the lesson as a result of being less visible to the teacher, and were exposed to more distractions (e.g., surrounded by family members) and temptations (e.g., "ghosting"). These asymmetries are likely to have resulted in remote students being more excluded than included in the lesson, while also impeding the development of a sense of community, or "co-presence" (Bower et al., 2014), between remote and on-site students.

While these phenomena undermined the quality of SHT during the pandemic, they should not be viewed as inherent features of SHT or as insurmountable obstacles. SHT during the pandemic was employed as an emergency solution and, as such, its implementation was not accompanied by appropriate planning and/or the necessary infrastructure. For SHT to deliver the desired learning outcomes for all students, there needs to be adequate investment in both the technological and social infrastructure of the teaching and learning space. Investing in the technological infrastructure would involve, for example, equipping the classroom with professional microphones, cameras, speakers and monitors, as well as ensuring that both the school's and the remote students' internet connection is sufficiently fast and reliable. Improving the social infrastructure, on the other hand, would require, for instance: providing teachers with appropriate training to help them cope with the technological and pedagogical demands of SHT;

familiarising remote students with the functionality of the online platform and introducing them to techniques they can employ to tackle possible technical issues; and developing routines and norms that can allow learning to continue when technical difficulties arise. To differentiate SHT carried out during the pandemic from more carefully planned and delivered forms of SHT found in many higher education institutions worldwide, it is suggested that the former is referred to as "emergency SHT" and not merely as "SHT".

Future directions

Through identifying four different types of challenges involved in delivering SHT in schools, this study hopes not only to have further illuminated SHT but also to have pointed to useful directions for improving SHT in the future and rendering it a truly inclusive mode of instruction. However, due to its small scale, this investigation needs to be complemented by further research. Future research into SHT in primary and secondary education could examine, for example: (a) the experiences of a larger number of teachers from a wider range of educational settings across the world to help develop a broader and more nuanced understanding of the phenomenon of SHT, (b) the experiences of students, both remote and on-site ones, to help provide a more holistic picture of the shortcomings as well as affordances of SHT, and (c) the effectiveness of different strategies for delivering SHT.

References

Bartlett, L. (2022). Specifying hybrid models of teachers' work during COVID-19. *Educational Researcher, 51*(2), 152–155.

Bell, J., Sawaya, S., & Cain, W. (2014). Synchromodal classes: designing for shared learning experiences between face-to-face and online students. *International Journal of Designs for Learning, 5*(1).

BERA. (2018). *Ethical guidelines for educational research* (4th ed.). London: BERA.

Bower, M., Dalgarno, B., Kennedy, G. E., Lee, M. J., & Kenney, J. (2014). Blended synchronous learning: A handbook for educators. Office for Learning and Teaching, Australian Department of Education.

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology, 18*(3), 328–352.

Carroll, M., & Constantinou, F. (2022). Learning loss in the Covid-19 pandemic: teachers' views on the nature and extent of loss. *Research Matters: A Cambridge University Press & Assessment publication, 34*, 6–25.

Carroll, M., & Constantinou, F. (2023). Teachers' experiences of teaching during the Covid-19 pandemic. Cambridge University Press & Assessment.

Centre for the Enhancement of Teaching and Learning. (2020). *Dual-mode teaching: introduction*. The University of Hong Kong.

Gosper, M., McNeill, M., Phillips, R., Preston, G., Woo, K., & Green, D. (2010). Web-based lecture technologies and learning and teaching: a study of change in four Australian universities. *Research in Learning Technology, 18*(3), 251–263.

International House World Organisation. (2020, August 13). *International House schools offering hybrid classes*.

Joshi, K. (2023, April 19). Is hybrid learning the future of education? Top insights from leaders. *Emeritus*.

Ladd, T. (2020, June 19). Optimizing concurrent classrooms: Teaching students in the room and online simultaneously. *Forbes*.

Raes, A., Detienne, L., Windey, I., and Depaepe, F. (2020). A systematic literature review on synchronous hybrid learning: gaps identified. *Learning Environments Research, 23*, 269–290.

Raes, A. (2022). Exploring student and teacher experiences in hybrid learning environments: does presence matter? *Postdigital Science and Education, 4*, 138–159.

Saatçi, B., Rädle, R., Rintel, S., O'Hara, K., and Nylandsted Klokmose, C. (2019). Hybrid meetings in the modern workplace: stories of success and failure. In H. Nakanishi, H. Egi, I.-A. Chounta, H. Takada, S. Ichimura, & U. Hoppe (Eds.), *Collaboration Technologies and Social Computing* [CRIWG+CollabTech 2019] (pp. 45–61). Springer.

VERBI Software. (2021). *MAXQDA 2022* [computer software]. Germany: VERBI Software. https://www.maxqda.com/

Wang, Q., Huang, C., & Quek, C. L. (2018). Students' perspectives on the design and implementation of a blended synchronous learning environment. *Australasian Journal of Educational Technology, 34*(1), 1–13.

Weller, S. (2021, April 10). *Why I'm happy working as a hybrid teacher*. TES.

Zydney, J. M., McKimmy, P., Lindberg, R., & Schmidt, M. (2019). Here or there instruction: Lessons learned in implementing innovative approaches to blended synchronous learning. *TechTrends, 63*, 123–132.

Research News

Lisa Bowett (Research Division)

The following reports and articles have been published since *Research Matters*, Issue 35:

Research and statistics reports on our website

Constantinou, F., & Carroll, M. (2023). Teacher-student interactions in emergency remote teaching contexts: Navigating uncharted waters? *Learning, Culture and Social Interaction, (43).*

Johnson, M., & Coleman, V. (2023). Teaching in uncertain times: Exploring links between the pandemic, assessment workload, and teacher wellbeing in England. *Research in Education.*

Johnson, M., & Coleman, V. (2023). Teachers' research diaries – reflection and reconnection in times of social isolation. *International Journal of Research & Method in Education.*

Lim, C. H. J. & Gill, T. (2023). *Provision of GCE A level subjects in 2018. Statistics Report Series No. 130.* Cambridge University Press & Assessment.

Lim, C. H. J., & Gill, T. (2023). *Provision of GCE A level subjects in 2019*. Statistics Report Series No. 131

Conference presentations

Brady, J., Majewska, D., & Greatorex, J. (2023, July 13-17). *Wanted dead or alive: canonical authors for literature in English curricula*. Paper presented at the 11th European Conference on Education, London, UK and online.

Crisp, V., & Elliott, G. (2023, July 13-17). *The elusive perfect question: an evaluation of the quality of past and recent exam questions in science*. The 11th European Conference on Education, London, UK and online.

Hughes, S., & Green, C. (2023, July 3-5). *Teaching and learning using digital technology: What are the affordances and how do they align with the affordances of digital?* Paper presented at Edulearn 15th annual International Conference on Education and New Learning Technologies, Palma de Mallorca, Spain.

Williamson, J., & Vidal Rodeiro, C. (2023, August 23-25). *The Impact of Curriculum and Assessment Reform in Secondary Education on Progression to Mathematics Post-16.* Paper presented at the European Conference on Educational Research, Glasgow, UK.

Williamson, J. (2023, June 16). *Performance in GCSE mathematics topics pre- and post-reform.* Paper presented at British Society for Research into Learning Mathematics (BSRLM) day conference, Manchester, UK.

The annual conference of the British Educational Research Association (BERA), took place at Aston University, Birmingham, 12 to 14 September 2023. Our researchers presented a total of 5 papers:

Constantinou, F. *Turning a curse into a blessing: teachers' reflections on how education has benefited from the COVID-19 disruption.*

Greatorex, J., Barnett, A., Beverley, P., Brady, J., Hingorani, P., Johnson, L., Morrish D., Nelson, S., Roberts, J., & Sidhu, A. *An analysis of cultural representations of India and the United Kingdom in English Language and Literature textbooks.*

Johnson, M., & Majewska, D. *Formal, informal and non-formal learning: Key differences and implications for research.*

Kreijkes, P. *Differential effects of subject-based and integrated curriculum approaches on students' learning experiences and outcomes: A review of reviews.*

Vidal Rodeiro, C. L. & Williamson, J. *Tracking the "June 2020 cohort": did the cancellation of exams in England hinder progression to post-16 study?*

Blogs, panels and podcasts

The following blogs and podcasts have been published since Research Matters, Issue 35:

Gill, T. (2023, June 13). Students taking an EPQ have better higher education outcomes.

Johnson, M. (2023). *School of Hard Knocks: How to Build Your Curriculum*. The Schools & Academies Show, ExCel Exhibition Centre, London, 17th May.

Kriejkes, P. (2023, June 28). What competencies do students need when working with data?

Majewska, D. (2023, July 21). What are the secrets to successful science education (if any)? Lessons from high-performing education systems. *Cambridge Partnership for Education.*

Oates, T., Suto, I., & Sultanova, G. (2023, September 11). Holistic Assessment: Theory & Practice – SIG AEA-Europe (online seminar).

Oyinloye, B. (Host). (2023, March 22). Innovative Research Methodologies: Solicited diaries with Martin Johnson [Audio podcast episode] In *BERA UK Podcast*. https://www.bera.ac.uk/media/innovative1

Powell, L. (Host). (2023, May 10). Big fish, little pond: how class rank affects A level choices with Dr Joanna Williamson [Audio podcast episode] In *Psychology in the Classroom.* https://changingstatesofmind.com/podcast

Sharing our research

We aim to make our research as widely available as possible. Listed below are links to the places where you can find our research online:

Journal papers and book chapters: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/journal-papers-and-book-chapters/

Research Matters (in full and as PDFs of individual articles): https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

Conference papers: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/conference-papers/

Research reports: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-reports/

Data Bytes: https://www.cambridgeassessment.org.uk/our-research/data-bytes/

Statistics reports: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/statistical-reports/

Blogs: https://www.cambridgeassessment.org.uk/blogs/

Insights (a platform for sharing our views and research on the big education topics that impact assessment around the globe): https://www.cambridgeassessment.org.uk/insights/

Our YouTube channel: https://www.youtube.com/channel/UCNnk0pi7n4Amd_2afMUoKGw contains Research Bytes (short presentations and commentary based on recent conference presentations), our online live debates #CamEdLive, and podcasts.

You can also learn more about our recent activities from Facebook, Instagram, LinkedIn and Twitter.

Contents / Issue 36 / Autumn 2023

Cambridge University Press & Assessment
Shaftesbury Road
Cambridge
CB2 8EA
United Kingdom

ResearchDivision@cambridge.org
www.cambridge.org