RESEARCH AND ANALYSIS

A review of standards in GCSE computer science

ofqual

Author

- Tim Stratton

With thanks to

- Charlotte Draper
- Rachel Taylor
- Ian Stockford

Contents

Executive Summary

Computer science is a relatively new GCSE that was first awarded in 2012. It has since gone through a series of changes, in terms of the content and assessment structure of the qualification, and in terms of the size and makeup of the cohort taking the qualification. Due to these changes, and following representations from stakeholders, Ofqual undertook a substantial programme of research to consider grading standards in GCSE computer science over time.

GCSE computer science represents an unusual scenario due to the number of changes which have occurred within the short lifespan of the qualification. These have included changes to the assessment structure, reform, the COVID-19 pandemic and reported high levels of malpractice in the pre-reform specifications. Entries have grown substantially from being only offered by a small number of schools and colleges to being much more widely available since being included in the Ebacc school performance measure and with the discontinuation of GCSE Information and Communication Technology (ICT). Substantial changes to the design of a qualification, the context within which it is operating, and the nature of its entry can introduce challenges in effectively maintaining standards over time.

This report includes details of the programme of research that Ofqual has undertaken to consider grading standards over time in GCSE computer science. There are 2 main strands of work: the first strand used a range of methodologies and analyses to consider whether there is any evidence that standards have not been consistently maintained over time, and the second aimed to consider the possible impact of any changes to the current grading standard, by reviewing examples of student work from summer 2023.

Strand 1 of this programme of work utilised a series of analytical approaches, both judgemental and statistical, to evaluate the grading standard of GCSE computer science over time. These analyses indicated that there has been a small reduction in the likelihood of students receiving at least a grade 7 (grade A pre-reform) or grade 4 (grade C pre-reform) between 2014 and 2019. No consistent effect is seen at grade 1 (grade G pre-reform). Through these analyses we aimed to exclude possible valid reasons for a change in outcomes, such as a change in the ability of the cohort, change in the familiarity of centres with the qualification and assessment and changes in the type of school or the cohort taking the qualification. Therefore, this change in outcomes likely suggests a small unintended change in the qualification standard. Analyses indicate this change in standard may have occurred primarily between 2014 and 2017, when there was a large increase in entry to the qualification, including many new centres offering computer science.

The second strand of work aimed to consider the possible impact a change in standards would have on the performance necessary to achieve a grade 7 or grade

4 in the most recent assessments. A group of 8 subject experts reviewed examples of students' work at various mark points. The findings indicated that the experts believed a small change in the performance standard necessary for students to attain a grade 7 or grade 4 would have a limited impact on the skills and knowledge demonstrated by students at these grades. However, a larger change would risk potentially undermining the value of the qualification. Experts also provided various qualitative insights into the standard of the qualification.

Taking into account the range of evidence, there is a compelling case that standards may not have been consistently maintained through the period from 2014 to 2019, with the standard being set slightly more severely during that period. This change in standards appears to have been the result of a gradual change over a series of years. These small incremental changes are unlikely to have been detectable by senior examiners in any individual year, but cumulatively have resulted in a more substantive change. This is not as a consequence of a failure of awarding organisations to have provided sufficient oversight and care through the awarding process but is a consequence of the changes to the qualification and the context in which it was operating during this period of time.

Introduction

Purpose

Computer science is a relatively new GCSE that was first awarded in 2012. It has since gone through a series of changes, in terms of the content and assessment structure of the qualification, and in terms of the size and makeup of the cohort taking the qualification. Due to these changes, and following representations from stakeholders, Ofqual undertook a substantial programme of research to consider grading standards in GCSE computer science over time, with a view to considering whether or not standards have been effectively maintained.

A brief history of GCSE computer science qualifications and assessment

A GCSE in ‘computing’ was first offered by the awarding organisation (AO) OCR, with a pilot award in 2011 and the first full award in 2012. The qualification was designed to develop students’ understanding of the inner working and programming of computer systems, distinct from the end-user focus of the existing GCSE in ICT (Dallaway, 2015).

Assessments in this first specification consisted of a single exam counting for 40% of the qualification grade and 2 controlled assessments, conducted in the classroom, worth 30% each. Each controlled assessment lasted around 20 hours and the final piece of work was generated under controlled conditions, that is, direct teacher supervision. These assessments were marked internally by schools and colleges (referred to as centres throughout) and moderated by OCR.

In 2013, the Department for Education (DfE) published a national curriculum for computing, covering key stage 1 to key stage 4 and in 2014 ‘computer science’ was added to the Ebacc school performance measure, in the sciences category. This aimed to incentivise schools to provide computer science education (Brown et al. 2014). In 2014, a GCSE in computer science was offered by 2 more AOs, WJEC and AQA, and OCR revised their specification to align with the new computer science requirements for inclusion in the Ebacc. The qualification was made available from a fourth AO, Pearson, for first assessment from 2015.

During the period from 2014 to 2017, the structure of the qualifications offered by the different AOs was similar to that initially offered by OCR, although there were some differences between the AOs. All the qualifications consisted of an exam and one or more controlled assessments that made up 25% to 60% of the total qualification (see Table 1 for details). Controlled assessments were all marked internally by teachers and externally moderated by the AOs. All AOs had a single exam paper, however WJEC also included a 2-hour onscreen externally marked problem-solving assessment making up 30% of the qualification in addition to the controlled assessment.

Table 1. Assessment structure for the different computer science specifications available before and after GCSE reform, including the percentage contribution of each assessment to qualification outcomes.

AO	OCR	AQA	WJEC	Pearson
First assessed	2012	2014	2014	2015
Pre-reform structure (until 2017)	40% Written exam (Computer systems and programming) 30% Controlled assessment 1 (Practical investigation) 30% Controlled assessment 2 (Programming project)	40% Written exam (Computing fundamentals) 60% Controlled assessment (Practical programming)	45% Written exam (Understanding computer science) 30% Onscreen assessment (Solving problems using computers) 25% Controlled assessment (Developing computing solutions)	75% Written exam (Principles of computer science) 25% Controlled assessment (Practical programming)
Post-reform structure (2018 onwards)	50% Written exam 1 (Computer systems) 50% Written exam 2 (Computational thinking, algorithms and programming)	50% Written exam 1 (Computational thinking and problem solving) 50% Written exam 2 (Written assessment)	50% Written exam (Understanding computer science) 50% Onscreen exam (Computer programming)	50% Written exam (Principles of computer science) 50% Onscreen exam (Application of computational thinking)

Note: The post-reform structure represents the assessment structure of the qualification following the removal of the NEA. See text for details.

All GCSE subjects were reformed for first teaching between 2015 and 2018. Reformed GCSEs are graded on a 9 to 1 grading scale, rather than the pre-reform A* to G scale. Reformed GCSE computer science specifications were based on core subject content defined by the DfE (DfE, 2015) and were available for first teaching in 2016. At the same time, GCSE ICT was discontinued (Ofqual, 2015a), which had until that point sat alongside GCSE computer science. The assessment requirements for the reformed qualifications were more specific, and therefore all AOs' assessments followed the same structure. Assessment of the post-reform qualifications was intended to consist of assessment by exam, contributing 80% of marks, and a non-examination assessment (NEA) intended to take 20 hours in total under tightly-controlled conditions, making up 20% of the marks. The NEA was again permitted to be marked internally, but externally moderated.

The NEA task for the first year of reformed GCSE computer science was released by exam boards in September 2017 and due to be completed by March 2018, with first exams of the reformed qualifications sat in summer 2018. However, shortly after the NEA was released, reports of widespread malpractice, including solutions to the assessment being available online, led to the rapid withdrawal of the NEA (Ofqual, 2017). Following a public consultation, Ofqual stipulated temporary interim assessment arrangements. Centres were still required to conduct the 20-hour assessment, but it no longer counted towards a student's overall grade. The exam boards updated the weighting of their exam papers so they counted for 50% of the marks each (see Table 1). These arrangements were intended to remain in place until 2021, while Ofqual consulted on long-term changes to the assessments (Ofqual, 2019).

Alongside reforms, in 2018, the National Centre for Computing Education (NCCE) was established to help train computer science teachers. The NCCE provides lesson plans and resources, as well as training programmes for teachers. By 2018 nearly 80% of year 11 pupils were in a school offering GCSE computer science (Kemp & Berry, 2019).

In 2020 and 2021, formal exams were cancelled for all GCSE and A level qualifications and were replaced with a system of teacher assessment, due to impacts of the COVID-19 pandemic. There was a return to exams in 2022 and at this point, following a further Ofqual consultation, the GCSE computer science exams were updated to include questions assessing students' knowledge and understanding of programming skills, in lieu of the NEA (Ofqual, 2019). This has remained the case for 2023 and 2024.

In summary, despite being a relatively new GCSE that was first awarded in 2012, computer science has undergone many changes in terms of its content and assessment structure. The context within which it is operating has also changed, as has the size and makeup of the cohort taking the qualification.

Setting and maintaining grading standards

Determining if standards have been effectively maintained in a qualification is a challenging task. Outcomes from a qualification can change year on year for many reasons. However, in most cases these are legitimate increases or decreases in outcomes, which do not necessarily reflect a change in the standard of the qualification.

When seeking to maintain standards outside of any times of change, the aim is to ensure that results across successive years of the same qualification can be interpreted in the same way, in terms of what it tells us about student attainment in the subject. Typically, we would say that standards have been maintained if students receiving that same grade in different years show equivalent levels of attainment. By attainment we mean the level of skills or knowledge that students have developed through their course of study. When all else is stable, we would expect this attainment to be evidenced through students' performance in their assessments. Therefore, during stable periods we would expect the quality of students' work produced in exams, or other assessments, at each grade boundary (that is, the 'performance standard') to be highly similar between exam series. The aim of the awarding process is to set grade boundaries that make that the case.

Identifying grade boundaries that maintain standards is not straightforward as assessments change from year to year, both in terms of the content covered from the qualification's specification and because of changes to the difficulty of the assessment. Although assessment writers aim to write assessments that are of similar difficulty each exam series, this is highly challenging to achieve in practice, therefore no two exams are likely to be of exactly equal difficulty. Consequently, the grade boundaries are unlikely to be the same from year to year. If the assessment is more demanding in one year, then we would expect the grade boundaries to be lower to compensate. There is an added level of complexity in that GCSE assessments are 'compensatory'. This means that students can gain marks in different areas of the assessment but receive the same total marks, potentially showing very different profiles in terms of their skills and knowledge. Therefore, to support examiners in their judgements, statistical evidence is used to help identify the direction and size of any changes in assessment demand. The details of these 2 types of evidence and how they are used in tandem is discussed below.

While there are complexities to the maintenance of standards, once a qualification is well established, the aim of maintaining the performance standard over time is relatively simple to conceptualise, as outlined above. This is less so during times of change, when assessments or qualification content is updated, such as during reform. When qualifications change, it is less meaningful to consider whether or not performance is maintained in the new reformed version, compared with the previous version for 2 reasons. First, the content of the qualification and the way that content is assessed is likely to have changed substantially meaning like-for-like comparisons are not possible. Second, there is evidence that student performance might be impacted during such changes. Previous data has shown that students' performance in assessments is typically weaker in the first year after reform, and this is usually attributed to teachers being less familiar with new content or features of the updated assessments (Cuff et al., 2019). Performance then gradually improves over the following few years as teachers become more familiar with the reformed qualifications. This pattern of a dip in performance followed by gradual improvement is referred to as the Sawtooth Effect (more can be read about it in Newton, 2020). During these periods, it may not be meaningful for examiners to seek to identify similar levels of performance between pre and post reform qualifications, and it may not be fair to do so as students risk being disadvantaged if they happen to be in the first, or early, cohort taking a newly reformed qualification.

Therefore, during periods of reform in England statistical evidence is typically prioritised and judgemental evidence provides a more supporting role, to ensure that students are not disadvantaged. This approach seeks to reward students with the same level of underlying attainment similarly either side of the reforms, not disadvantaging those whose performance in the assessments may have been lower due to a lack of familiarity with the assessments post-reform. The assumption is made that if the makeup of the cohort has not substantially changed then we would not expect the outcomes to substantially change year on year, at the cohort level. The principle behind the use of statistical evidence is therefore that outcomes on the new assessments should be comparable to outcomes if the same cohort had taken the qualification in another year (Cresswell, 2003). However, this means that the quality of work produced by students during these periods may be weaker than that receiving the same grade during stable periods.

Operationalising the setting and maintenance of standards

Standards in GCSE assessments in England are maintained through the setting of grade boundaries. Grade boundaries represent the lowest mark where students demonstrate the performance necessary to receive each grade. Pre-reform at GCSE, awarding focussed on the key judgemental boundaries of A, C and F. To support the maintenance of standards across the transition, grades A, C and G were referenced to grades 7, 4 and 1 post-reform, which became the new judgemental boundaries. The intermediate boundaries are calculated arithmetically, equally spaced between the judgemental boundaries. Examiners use a range of evidence to help guide their decision making when recommending grade boundaries.

Examiners typically scrutinise examples of student work to identify the mark where students demonstrate the same level of performance as those at the grade boundary in the previous year. To achieve this, ‘archive evidence’ representing students’ work at each grade boundary from previous years is used to encapsulate the expected level of performance. Examiners review the quality of student work compared with the archive evidence, to identify the grade boundaries that most closely carries forward the performance standard from the previous year.

As discussed in the previous section, examiners decisions are supported by statistical evidence. One key source of statistical evidence is prior-attainment-based predictions. Predictions take into account the prior attainment of each cohort and provide an indication of what outcomes might be expected to look like if the cohort in the current year is similar to that in a previous reference year, in terms of all features which may affect outcomes except for prior attainment. They achieve this by carrying forward the value-added relationship for a qualification from a reference year, that is, the relationship between the cohort’s performance in a previous set of qualifications and the current assessment. For GCSEs this is typically the relationship between KS2 assessment results and GCSE results. Therefore, if the ability of the cohort as measured by their prior attainment is similar to the reference year, then the predicted outcomes will be similar to those in the reference year. However, if the ability of the cohort taking the subject has increased or decreased then the predictions will change accordingly. AOs use predictions to identify the grade boundaries which most closely maintain the relationship between prior attainment results of the cohort and results in the subject in question over time. The boundaries suggested by the predictions are then used to guide examiner judgements to set grade boundaries. Using statistical predictions in this way also helps support the alignment of standards between different AOs.

Predictions are based on a subset of ‘matched candidates’, those who are of target age group (for GCSEs this is those who would be 16 on 31 August of the year they took their exams), who have available prior attainment data (KS2 results). GCSE predictions also typically exclude students at selective or independent centres, as research has shown that they tend to have a different value-added relationship between prior attainment and current outcomes than students at other centres.

The reliability of statistical evidence will vary depending on the size and stability of the cohort taking the qualification. When the number of students taking a qualification is small, the statistical evidence is likely to be weaker, so AOs will put more weight on other sources of evidence, such as examiner judgement. Similarly, if there have been substantial changes in the cohort taking a qualification, such as large increases or decreases in entry, or changes to the types of students or centres taking a qualification, then statistical predictions may be less reliable representation of the performance of the current cohort.

In the early years following reform to the GCSE, greater weight was placed on statistical predictions. As discussed previously, the intention of this was to avoid students being disadvantaged in the early years post-reform, when performance may be lower due to teachers' unfamiliarity with the new content and assessments. However, awarding teams continued to scrutinise examples of student work to confirm that the quality of students' work at the grade boundaries was acceptable.

Setting and maintaining standards in GCSE computer science, 2012-2023

In the first award of GCSE computer science in 2012 it was necessary to set the standard for this new qualification. The first award was largely judgemental, but statistical evidence was used to support awarders' judgements. Statistical predictions were produced from a selection of related GCSE subjects (namely, ICT, physics and maths) to provide an indication of what outcomes might look like in the first award of computer science, taking into account the ability of the cohort. This was then used to inform examiner scrutiny of the quality of work students produced in the assessments to determine grade boundaries.

Following the first award in 2012, until 2017, grade boundaries continued to be set based on a balance of statistical evidence and examiner judgement with the aim of maintaining the performance standard. In most years, statistical predictions were produced based on the previous year's outcomes, to guide examiners in making their judgements. In the early years of the qualification, AOs would also have been aware that these were new assessments and a new specification, with which teachers would have been somewhat unfamiliar.

Reformed GCSE computer science assessments were first awarded in 2018. As for all reformed GCSEs, statistical evidence was prioritised during the reform period (2018 and 2019). As described above, this was to ensure that students were not disadvantaged due to any dips in performance during the transition years and to overcome the challenges to the use of examiner judgement during this period.

During the COVID-19 pandemic (2020 and 2021), normal assessment arrangements were suspended, and grades were awarded based on teacher assessments. Normal exam arrangements for GCSEs returned in summer 2022. However, in 2022 grade boundaries were set in such a way that outcomes were broadly mid-way between results in 2021 and 2019 as part of the 2-year return to pre-pandemic standards.

Summer 2023 then represented the first year that grading returned to pre-pandemic grading standards. To facilitate this, standard setting in GCSE computer science in summer 2023 was guided by predictions so that overall results would be similar to outcomes in 2019. This approach was taken to carry forward the grading standard from before the pandemic, but with protection built into the grading process to recognise the disruption that students had faced. This allowed for the fact that exam performance may have been a little lower than before the pandemic, similar to the approach taken during reform. However, examiners were asked in awarding in 2023 to review students' work at the grade boundaries and confirm that students were demonstrating an acceptable level of performance. Therefore 2023 provides a good representation of the current performance standard.

Structure of this report

The preceding sections of this report have outlined the history of GCSE computer science, the principles that underpin the setting and maintenance of standards, and how that is operationalised through the process of awarding. This aims to support understanding of the analytical approaches that are documented through the main sections of this report.

There are 2 main strands of work, the first strand used a range of methodologies and analyses to consider whether there is any evidence that standards have not been consistently maintained over time, and the second strand aimed to consider the possible impact of any changes to the current standard, by reviewing examples of student work from summer 2023.

The methodology, results and interim findings relating to each analysis are reported in the sections that follow, before the overall findings are discussed and conclusions drawn.

Throughout this report, reference will be made to grade A/7, C/4 or G/1 to describe effects at those grades that span across pre and post reform versions of the qualification. When referring to the percentage of students receiving each grade we mean the cumulative percentage, that is the percentage of students receiving either the grade in question or a higher grade.

Strand 1 - Standards over time 2012-2019

Aims

The aim of this strand is to look back on standards in GCSE computer science historically, focussing on the period from when assessments were first sat (2012) to the last year before the pandemic (2019). This is principally to identify if there have been any unexpected changes in standards in this qualification over time. If any changes in standard are identified the aim is to try to understand the cause of these changes and the size of the impact on student outcomes.

Structure of strand 1

To achieve the above aim we have taken a variety of approaches to consider standards in the qualification over time, both purely quantitative and more qualitative approaches. The following sections of this report will outline each of these methods in turn, detailing the aims, methodology and key findings of the individual approaches. Each of these methods allows us to control for different potentially confounding factors, however each method also comes with its own limitations and assumptions, which we outline in each section. We will then draw together the findings from these individual analyses.

The first section (analyses 1 and 2) includes contextual background information to changes in the qualification. This includes descriptive information of how the qualification and cohort has changed over time, and also an overview of how the AOs approached standard maintenance and setting grade boundaries in each year.

The second section (analyses 3 to 7) contains a range of statistical methods that look at the relationship between outcomes in the qualification and other measures of student attainment over time.

To overcome the limitations of purely statistical approaches, the final section (analyses 8) includes a more judgemental approach to considering standards over time. Here subject experts are used to review the demand of the assessments in each year and draw comparisons of the quality of students’ work necessary to receive key grades over time.

Caution needs to be taken when directly comparing between outputs from the different analyses as they each have their own assumptions and in some cases are calculated using a slightly different subset of the population.

Data

The key dataset used for the majority of the analyses presented in this report is the National Pupil Database (NPD). This is a dataset maintained by the DfE and contains details of students' assessment results, along with a large number of other student and centre characteristics. Data was taken from NPD years 2011 to 2019 and filtered to students who had taken GCSE computer science for the primary analyses. Data from GCSE maths, physics and English language are also used in various analyses for comparison. Prior attainment data was available in the NPD for the majority of students in each year based on their key stage 2 (KS2) national curriculum assessment results in maths and English.

Data was filtered to only 16-year-old students from England, who had a valid GCSE grade. Results data were combined with Schools Census data to provide student characteristics, these included: centre type attended, gender, ethnic group, language spoken, special education needs (SEN) status and free school meal (FSM) eligibility. Table 2 shows the number of students entered for GCSE computer science in each year, along with the percentage of students with available census data and prior attainment data. It is notable that the availability of prior attainment data is lower in 2015 due to boycotts of KS2 assessments in 2010. The majority of the analyses focus on years 2014 to 2019. Data prior to 2014 is presented where possible but needs to be treated with caution as entries were small.

Data on statistical predictions used by AOs comes from datasets regularly shared with Ofqual as part of routine monitoring of results in each year. Additional data was also collected from the 2 AOs with the largest entry to computer science, OCR and AQA. This included documentation of decision-making during grade boundary setting in each year and some of the supporting information used.

Where additional datasets were used, or additional data processing was carried out, this is detailed in the relevant section.

Table 2. Summary of student numbers and match rates across data sets in each year

Year	Total computer science Students	% with census data	% with prior attainment data
2011 (pilot)	92	97.8	96.7
2012	1,745	92.3	90.7
2013	4,179	95.9	92.7
2014	16,011	96.7	92.2

2015	33,773	96.6	69.1
2016	61,751	96.9	92.6
2017	67,374	96.8	92.5
2018	71,111	96.2	91.6
2019	75,165	95.2	91.6

Strand 1. Analysis 1. Cohort changes and outcomes over time

Aim

Changes to the cohort taking an assessment can make it more challenging to effectively maintain standards over time in a qualification. The aim of this first section is to identify any changes to the cohort entered for GCSE computer science over time. This will identify whether such changes might indicate a case for further exploration and provide context for any further analysis.

Cohort size

Figure 1 shows the number of students entering GCSE computer science between 2011 and 2019, both overall and broken down by individual AO. There are 2 notable things from this figure. The first is that OCR, the AO who were first to offer the GCSE, have continued to have the majority of entries over time. Second, the number of students taking the qualification increased rapidly between 2014 and 2016, before the increase in entries slowed down.

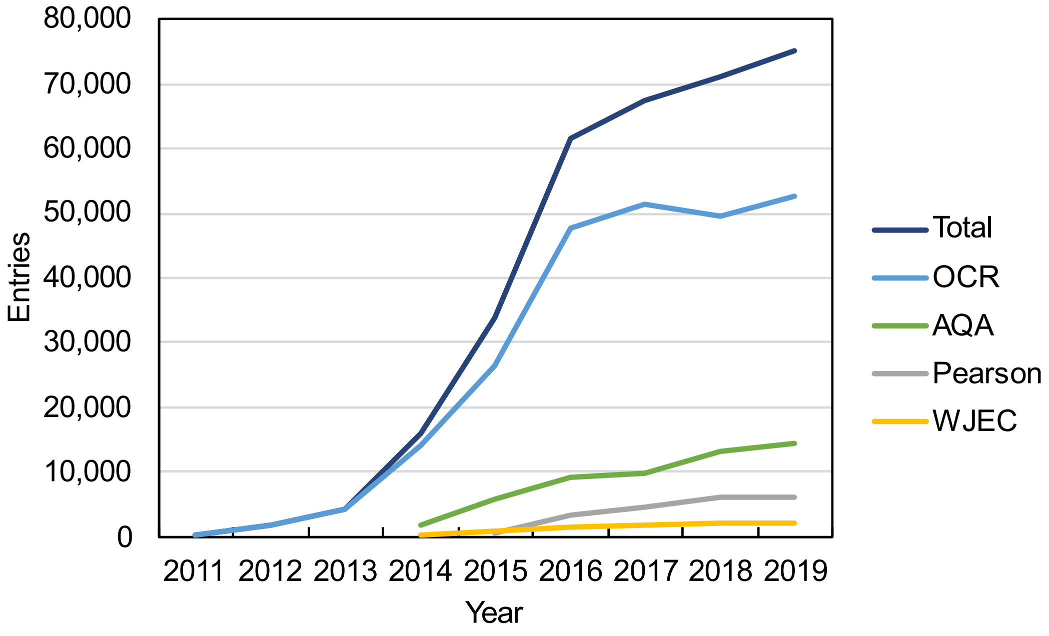

Figure 1. Entry numbers to GCSE computer science over time, overall and broken down by AO.

Centre characteristics

Figure 2 shows the different types of centres entering students to GCSE computer science over time. Centres are categorised into independent centres, selective centres, maintained schools (including secondary schools, academies and free schools), and colleges. It is notable that in the first couple of years of the qualification being available there was a much larger proportion of students from independent and selective centres. However, following 2014, the proportion of students from different centre types stayed broadly stable.

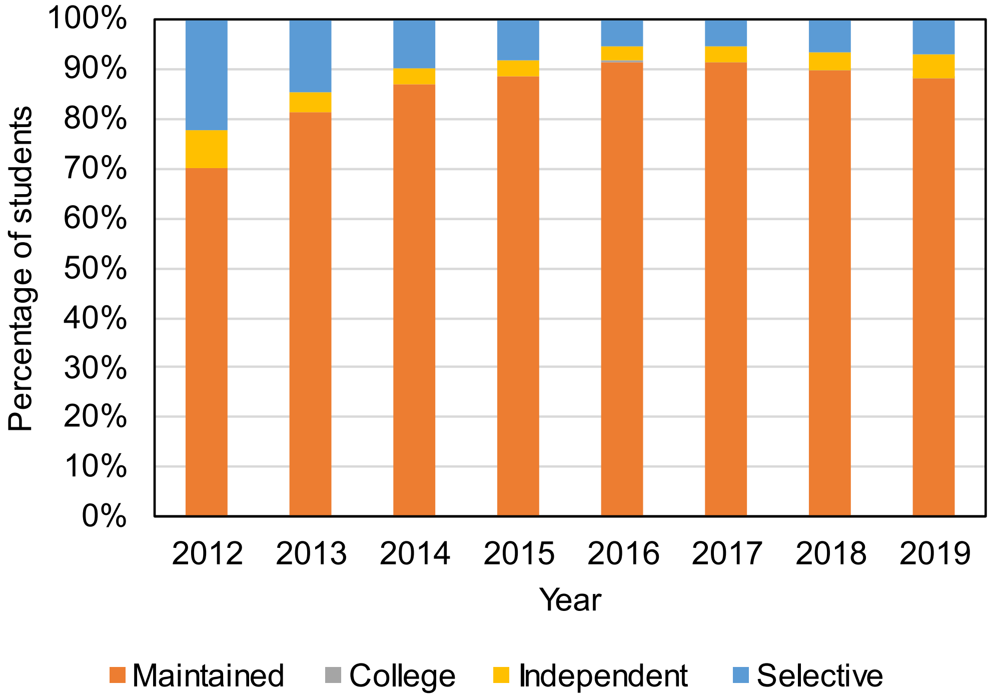

Figure 2. Proportion of entry in each year from different centre types.

Figure 3 shows the number of students at ‘new’ centres entering the qualification each year. By ‘new’ centres we mean centres that had never previously entered students for the qualification. As can be seen from Figure 3, a large proportion of students that were taking the qualification were from ‘new’ centres until around 2016. Table 3 summarises the number of centres and the average entry per centre in each year. It is notable that as the entry size increased the majority of this increase was through new centres offering the qualification, rather than existing centres increasing the number of students they entered. Average entry size per centre did gradually increase between 2012 and 2016 before stabilising from 2016 onwards, which may suggest there was some change in the cohorts within centres. The standard deviations also indicate that there is a large amount of variation in entry size between centres.

These changes to the cohort are worth noting in the context of evidence showing that when centres are unfamiliar with offering a qualification, students at these centres can perform less well in the assessments (Newton, 2020). In years where a large number of students entering the qualification were at new centres, that could lead to the performance of the cohort being weaker than might otherwise have been expected.

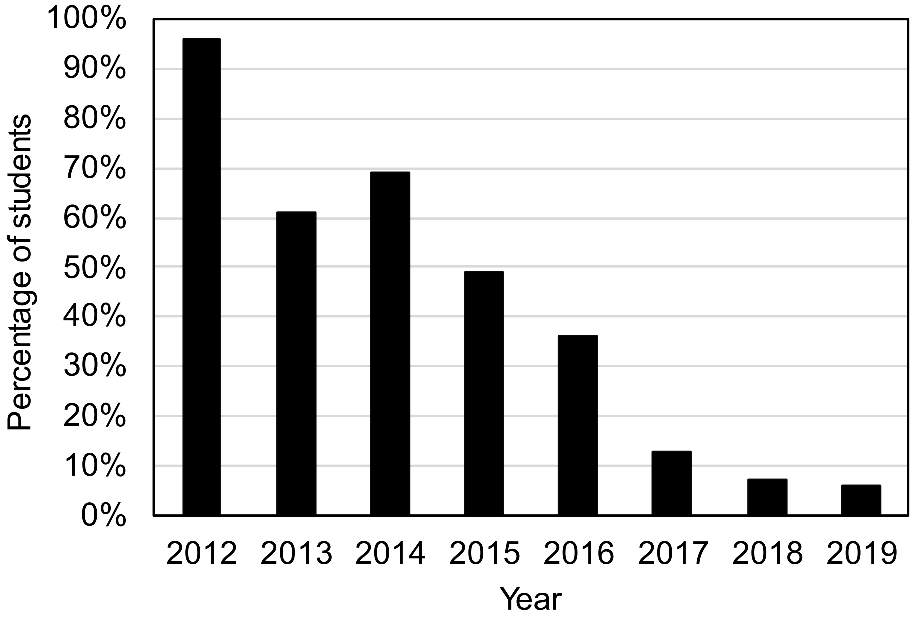

Figure 3. Percentage of students entering in each year from centres that were entering students for the first time.

Table 3. Number of centres and average number of students per centre over time.

Year	N Centres	Mean entry per centre	SD entry per centre
2012	97	18.0	10.7
2013	210	19.9	14.2
2014	724	22.1	15.0
2015	1437	23.5	15.8
2016	2340	26.4	19.7
2017	2652	25.4	17.4

2018	2845	25.0	16.7
2019	2922	25.7	17.5

Student characteristics

Next, we look at the characteristics of the students taking computer science over time. Figure 4 shows the average standardised prior attainment score of students over time. This is students' attainment in KS2 assessments 5 years before taking the GCSE. KS2 score is presented here on a standardised scale between 0 and 100 with a mean of 50 across all GCSE students. From Figure 4 it can be seen that the prior attainment of students taking computer science decreased fairly rapidly between 2012 and 2014, stabilised in 2015, before dropping again in 2016 and gradually increasing until 2019. The relationship between prior attainment and GCSE outcomes is strong for many GCSE subjects (Benton & Sutch, 2014) and so can give a good indication of expected outcomes, where other factors remain stable. Crucially, it is also used in the generation of predictions which are used to help set grade boundaries (see the section *Operationalising the setting and maintenance of standards*).

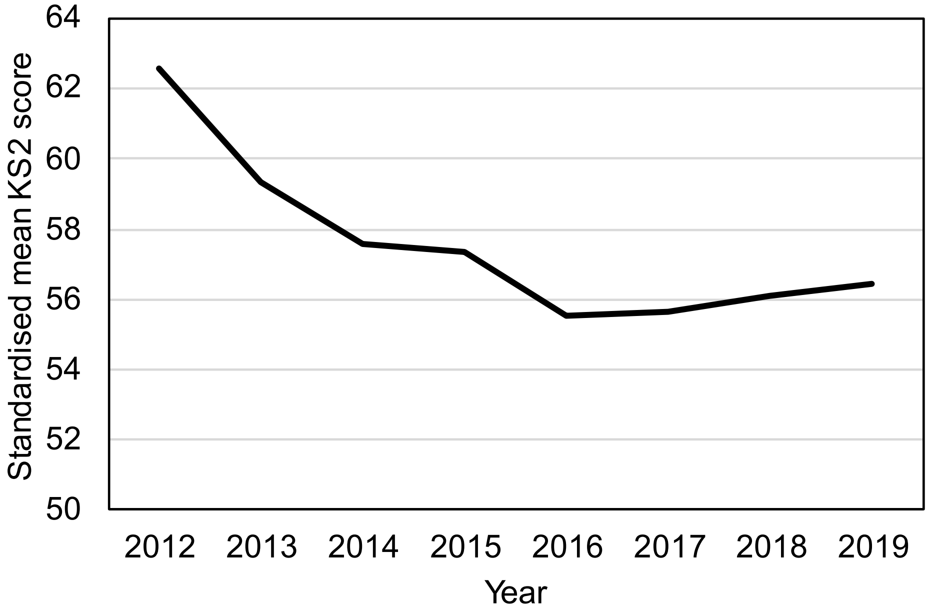

Figure 4. Mean prior attainment score over time for GCSE computer science students.

Table 4 shows the proportion of students taking computer science with different characteristics. This shows that the candidature has gradually changed over time. Most notably the proportion of students taking the subject with English as a foreign language (EFL) and those with special educational needs (SEN) has gradually increased since 2013. The candidature has also become more diverse, with a lower proportion of white students and a growing proportion of female students taking the subject. The largest step change in most of the characteristics was between 2015 and 2016, when entries also increased substantially.

Table 4. Characteristics of GCSE computer science cohort over time.

Year	% FSM	% EFL	% SEN	% White	% Female
2012	5.2%	13.5%	10.3%	79.0%	13.5%
2013	8.4%	14.5%	8.5%	77.9%	14.5%
2014	9.9%	14.9%	9.6%	78.0%	15.4%
2015	9.2%	15.6%	9.0%	78.1%	16.2%
2016	10.3%	17.0%	9.3%	77.3%	20.5%
2017	9.8%	17.2%	9.3%	76.8%	20.2%
2018	9.7%	18.9%	9.6%	74.1%	20.4%
2019	10.7%	19.9%	9.7%	71.8%	21.6%

Outcomes

Finally, we look at outcomes in the qualification over time. This is intentionally presented after the above analysis of other changes over time, as outcomes can change for a number of legitimate reasons that may be related to some of the above changes in entry patterns.

Figure 5 shows the cumulative percentage of students attaining at least a grade C/4 and A/7 over time. While outcomes have generally fallen over time at both grades, there is a particularly notable shift between 2015 and 2016. This coincides with the large increase in entries and some of the changes in candidature noted above. It also coincides with the fall in average prior attainment of the cohort, which could represent a legitimate fall in outcomes.

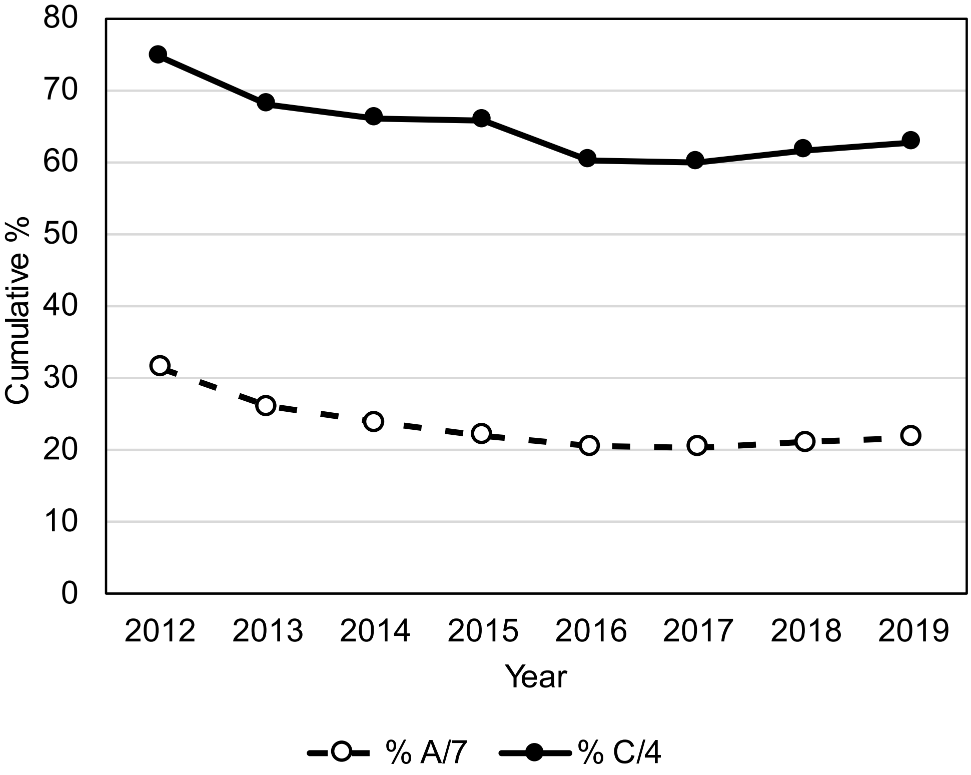

Figure 5. Cumulative percentage outcomes of students receiving at least a grade A/7 and C/4 over time.

Summary

The descriptive analyses reported here provide context for the analysis and discussion that follows. These analyses confirm that the qualification has seen a change in the size and make-up of the cohort taking the subject, along with a change in the outcomes in GCSE computer science. These analyses have shown that outcomes have declined in GCSE computer science over time, most notably between 2012 and 2016.

As noted, changes in cohorts over time make the maintenance of standards more challenging. The analyses that follow seek to identify whether or not the changes in outcomes that have been observed reflect genuine changes in the attainment of the GCSE computer science cohort over time, or whether they may be attributable to a change in standards over that period.

From the above descriptive analysis of student characteristics there are 3 potential legitimate reasons for a change in outcomes:

1. Outcomes could decline because the cohort taking GCSE computer science became weaker over time. Evidence from the prior attainment data suggests that this may be the case.

2. Students at centres that are delivering the qualification for the first time may perform worse in the assessments, potentially due to teacher unfamiliarity with the course content and the assessments. This could see outcomes for those centres being lower than would be the case when their familiarity increases. In years with a large number of new centres, this could contribute to overall dips in outcomes, if the boundaries based on the predictions suggested a quality of student work that could not be supported by the examiners.

3. Cohorts in later years may be functionally different than those in earlier years. As the number of centres increases, cohorts at those newer centres may have typically lower outcomes (relative to their prior attainment) than students at centres taking the qualification in earlier years. This could be due to demographic differences or to factors such as centre resources or teacher expertise differing between early uptake and later uptake centres.

Each of the above factors could lead to legitimate changes in outcomes in the qualification. Throughout the rest of the report, we aim to control and compensate for one or more of these factors in the analyses to understand what may be contributing to a change in outcomes. If changes in outcomes cannot be attributed to the above factors this could indicate an unintended change in standards over time.

Strand 1. Analysis 2. Predictions, grade boundaries and awarding documents

Introduction

In this section we review data from the AOs offering GCSE computer science over time that results from, or contributes to, decision making regarding grade boundary setting in each year. The aim is to identify whether there might be indicators of a potential change in standards, or risks to the maintenance of standards.

Outcomes relative to predictions

Grade boundaries in GCSE computer science were set using a balance of statistical and judgemental evidence. Each year statistical predictions were created based on a reference year, from which the relationship between prior attainment and outcomes is carried forward, as described in the section *Operationalising the setting and maintenance of standards*.

For GCSE computer science, each year, predictions were based on outcomes in the previous year, except for in 2016 when predictions were based on 2014. The reason for updating the reference year for predictions is typically to better reflect the cohort taking the assessment if the cohort make-up is changing over time, as was the case for computer science. The reference year may also be updated if entries have increased in small entry subjects, as larger samples usually provide a more reliable prediction. In 2016 the ‘reference year’ was not updated for computer science as, due to KS2 assessment boycotts in 2010, 2015 had fewer matched candidates.

Figure 6 and Figure 7 summarise the difference between predicted outcomes and matched candidate outcomes for grades A/7 and C/4, respectively. Data is combined across all AOs offering the qualification in each year, weighted by their total entry.

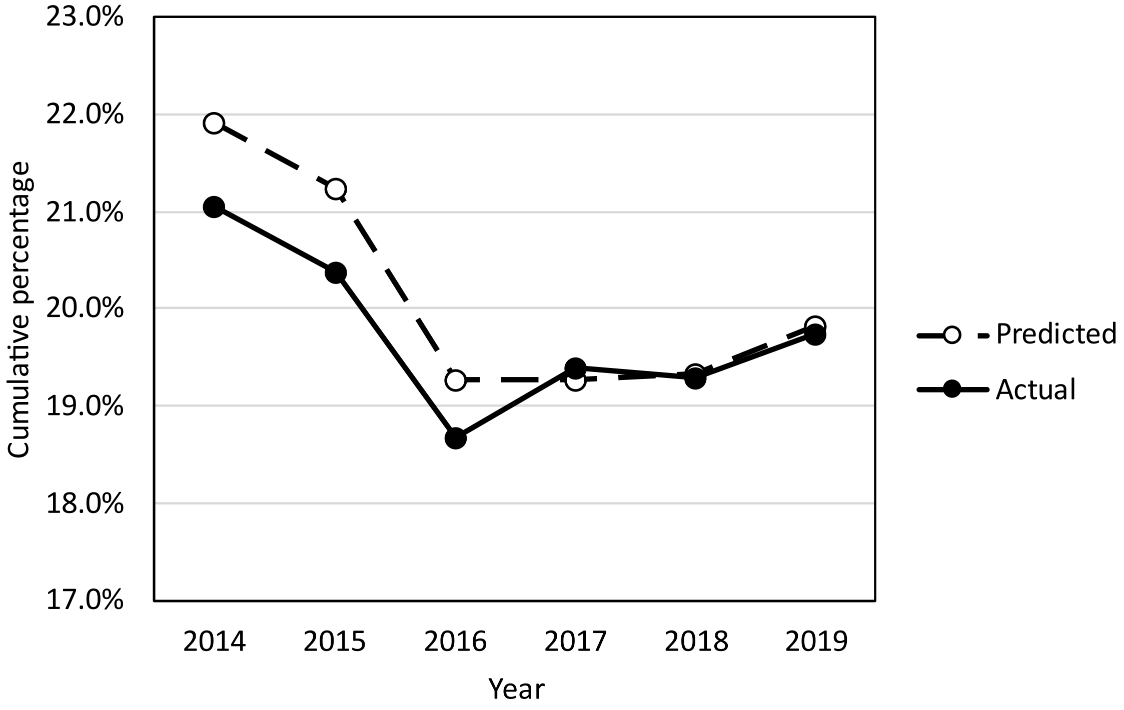

Figure 6. Cumulative actual outcomes and predicted outcomes for matched candidates for GCSE computer science at grade A/7.

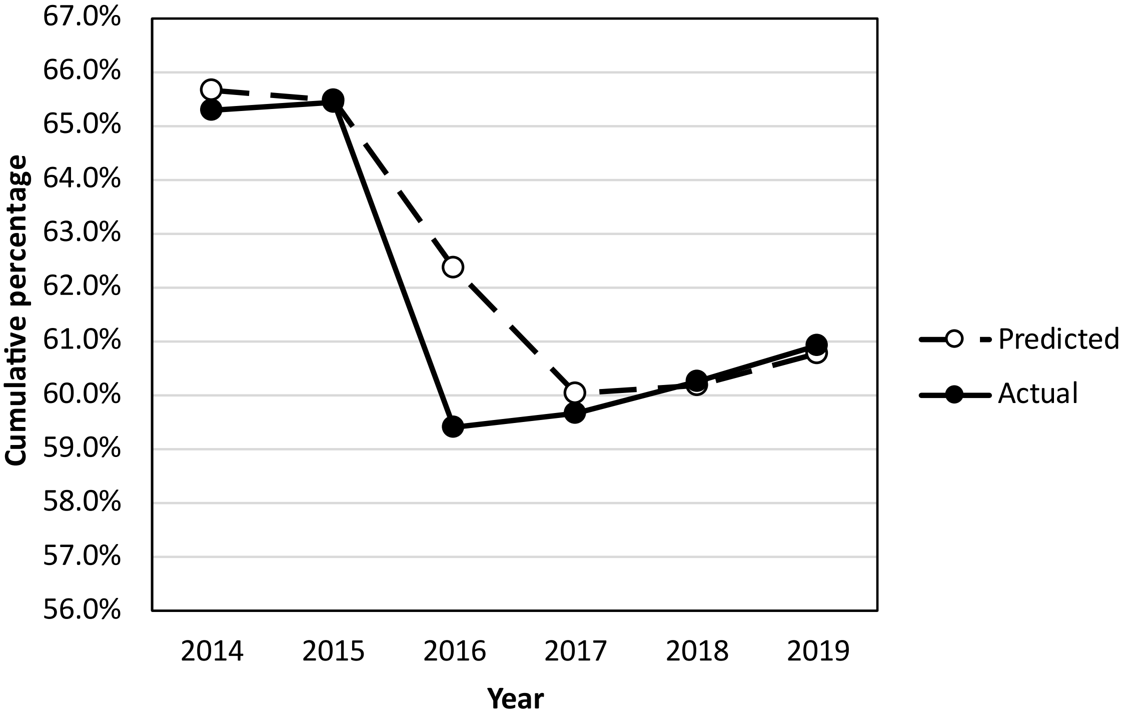

Figure 7. Cumulative actual outcomes and predicted outcomes for matched candidates for GCSE computer science at grade C/4.

Figure 6 and Figure 7 show that between 2014 and 2016, outcomes at grade A/7 were slightly below predictions, although within a one percentage point (pp) difference. Given that predictions are likely to be less reliable when based on small entry numbers, 1pp does not represent a large difference and awarders may legitimately put more weight on other evidence when statistics are less reliable. At grade C/4 outcomes were close to predictions in all years except 2016, when they were overall around 3pp below prediction.

Information from awarding documents indicates that where outcomes were below predictions it was typically because examiners judged the quality of work to be too low at the grade boundary indicated by the prediction and so a higher boundary was recommended than those suggested by the predictions. Figure 8 shows the grade boundaries set over time for the AO with the largest entry (OCR) in both their examined and controlled assessments.

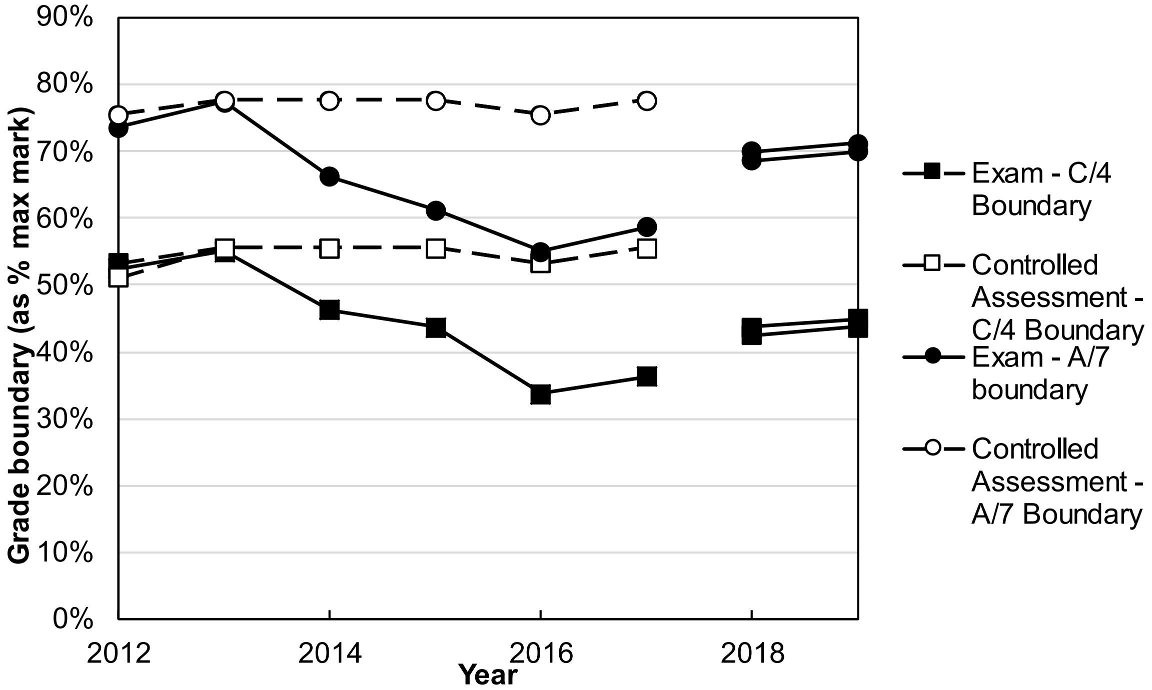

Figure 8. Grade boundaries over time for OCR assessments.

Malpractice

Documentation highlights examiners' concerns around malpractice from as early as 2014 up until reform, and the risk that this would lead to grade inflation in the controlled assessments. Despite this concern, typically the grade boundaries in the controlled assessment element were kept stable to reflect the fact that the task, and therefore the demand of the assessment, remained similar from year to year. It is perhaps surprising that despite grade boundaries typically being lowered in the examined element, and possible grade inflation in the controlled assessment, this did not lead to higher, rather than lower, outcomes over time. This implies a cohort which was weaker and less well prepared over subsequent years relative to their prior attainment.

In 2016 OCR made changes to one of their controlled assessments to make the task more open ended, in an attempt to avoid malpractice such as solutions being posted online. This change to the assessment could have resulted in a temporary change in performance due to the newness of this assessment, leading to a sawtooth-like pattern of performance. The grade boundary in the controlled assessment was lowered by one mark to compensate for the potential increase in difficulty (see Figure 8), however boundaries were raised again in 2017 as performance improved.

Reference year

One possible effect arising from the way standards were set during this period, is the potential cumulative effect of repeatedly awarding below prediction, followed by the updating of reference years for calculating future predictions. If outcomes are awarded below prediction in a particular year and that year becomes the reference year for a future year's prediction (to best reflect the most recently observed value-added relationship), then the expected value-added relationship is such that predicted results will be lower for students with the same prior attainment than would previously have been the case. If this happens repeatedly, as in GCSE computer science, this leads to a cumulative lowering of the expected value-added relationship for future cohorts, cumulatively lowering the expected outcomes for students with the same prior attainment over time, in order to reflect the observed performance of students.

Table 5 gives a rough estimate of the size of this effect, taking into account the reference year used in each year. Although this is only a simple calculation, that does not take into account possible changes in the prior attainment distribution over time, it does indicate a potential 'deflationary effect' on outcomes, albeit one based on judgements of the acceptability of students' work. Table 5 indicates that by 2019 this 'deflationary' effect could have led to predictions around 1.5pp lower at A/7, 3.5pp at C/4 and 1.8pp at grade G/1 relative to 2014.

Table 5. Estimated cumulative effect of awarding below prediction on future predictions. Figures indicate the cumulative difference between predictions and outcomes over time in percentage points.

Year	Reference year used for predictions	Grade A/7	Grade C/4	Grade G/1
2014	2013	-0.87	-0.37	0.05
2015	2014	-1.72	-0.42	0.04
2016	2014	-1.45	-3.33	-1.46
2017	2016	-1.33	-3.70	-1.78
2018	2017	-1.37	-3.63	-1.67
2019	2018	-1.46	-3.49	-1.78

In individual years, these outcomes reflect examiner judgements of the quality of performance that were made during that period. This lowering of predictions may therefore be justified if this represents a permanent change to the cohort’s expected value-added relationship, that is, if overall the cohort is performing worse relative to their prior attainment than in previous years and this is expected to continue indefinitely. However, if some of the weaker performance in previous years was due to temporary effects, such as sawtooth or sawtooth-like effects, this could result in an unjustified permanent shift in standards.

Summary

Overall, the awarding reports indicate that there were a number of challenges in maintaining standards over time, particularly around the period 2014 to 2016, when there were a large number of students from new centres. AOs set grade boundaries below those suggested by the prior attainment-based predictions on a number of occasions during this period, which may have been due to students at new centres demonstrating weaker performance. The position of the grade boundaries also suggests a growing mismatch in performance between the controlled assessment and exams, which could have been related to malpractice.

One potential risk highlighted from the review of awarding materials was related to the approach to calculating the predictions. The reference years were updated to ensure the predictions were a faithful representation of the awarded value-added relationship observed the previous year. This change, combined with successively awarding below predictions could have led to a small cumulative lowering of expected outcomes, the appropriateness of which may or may not have persisted in future years.

Strand 1. Analysis 3. Outcomes relative to other GCSE qualifications over time

Aim

One way to consider qualification standards is to look at how students taking a particular qualification performed in other qualifications they took alongside. The aim of this section is to analyse if the relationship between students' results in GCSE computer science compared with their results in the other subjects they took alongside changes over time. A change could indicate that standards have changed in computer science.

The intention for this analysis was not to focus on the direct statistical comparability between computer science and other subjects. *Absolute* differences between subjects in these analyses are not problematic, either in a particular year or persisting over time. Students' grades in different subjects may be higher or lower than in other subjects for a large number of reasons, which may include teaching time dedicated to the subject, student motivation, how long students have studied the subject, among other factors (for a more detailed discussion see Ofqual, 2015b). We therefore would not expect students' results to be perfectly aligned across subjects. Instead, the aim of this analyses was to use results in other subjects as a benchmark to identify if the *relative* difficulty of computer science had changed over time. The key assumption of this analysis is therefore that there is no reason to expect the relative difficulty of the subjects we are comparing to have changed over time.

Methods

We used 2 methods to provide a difficulty estimate for GCSE computer science compared with other subjects taken by the same students in each year, a Rasch difficulty model (see Coe, 2008) and Kelly's method (Kelly, 1976). These analyses give an indication of how well students performed on average in other GCSE subjects in each year relative to computer science and provide a relative 'difficulty' estimate for each subject.

It is worth noting, however, that although these methods effectively control for the 'general ability' of the cohort, as measured by students' performance in other GCSE subjects, they are not able to control for those other factors which may change over time and which may affect performance in specific subjects, such as teaching quality or student motivation.

The first method was to use a Rasch difficulty model to equate difficulty in different subjects in each year. For this model, each subject a student took was treated as an individual item on an assessment. However, only the key grades were used as a threshold for performance categories. To facilitate this, grades were converted to a score (see Table 6). Students not taking a subject were treated as missing responses. Only students who had taken at least 3 GCSEs were included in the analysis and only subjects with at least 1,000 entries in each year.

Table 6. Details of grade conversions to scores for the Rasch analysis.

Score	Legacy qualifications	Reformed qualifications
0	Ungraded	Ungraded
1	D, E, F, G	3, 2, 1
2	B, C	6, 5, 4
3	A, A*	9, 8, 7

The Rasch model was then fitted to simultaneously provide a ‘difficulty’ measure for each of the key grades for each subject and an ‘ability’ measure for each student. The difficulty measure is effectively the average ‘ability’ score of students achieving each grade in each subject in each year. A higher score on the Rasch difficulty scale therefore indicates that it is harder for the average student to achieve that grade. Outcomes for the Rasch model are inherently relative to other subjects and are on an arbitrary scale. Therefore, instead of presenting the Rasch difficulty scores in isolation, we provide the relative difference in difficulty estimates between computer science and 3 other subjects, maths, physics and English language in each year. We use these subjects because of their large and relatively stable entry and because maths and English language are taken by the vast majority of 16-year-old students in each year. Therefore, if we assume the ability distribution of students included in the analysis is similar in each year then we can compare these scores between years to see how they change. For a more detailed discussion of the methodology see He and Black (2020) and He and Cadwallader (2022).

The second approach, Kelly’s method, provides an alternative difficulty estimate. It involves calculating the grade ‘adjustment’ required in each subject for the average difference between each student’s grade in that subject and the average of their other subject grades to be 0 (for more details of the methodology see Coe et al, 2008). This estimate can be loosely interpreted as the mean difference in difficulty for each subject from the average subject. The adjustment is calculated on the A*-G

(8 to 1) grade scale. Therefore, for this analysis 9 to 1 grades were converted to an 8 to 1 scale, based on the estimated probability a student gaining each numbered grade would have received each lettered grade (see Table 7).

Table 7. Details of conversion of 9 to 1 grades to an 8 to 1 scale for analysis.

Grade on 9 to 1 scale	Grade converted to 8 to 1 scale
9	8
8	7.25
7	7
6	6
5	5.5
4	5
3	3.75
2	2.5
1	1.25
0	0

Again, instead of providing the absolute score we present the relative difference in scores between computer science and physics, English language and maths to identify if the gap between computer science and these subjects has changed over time.

Results

Rasch difficulty

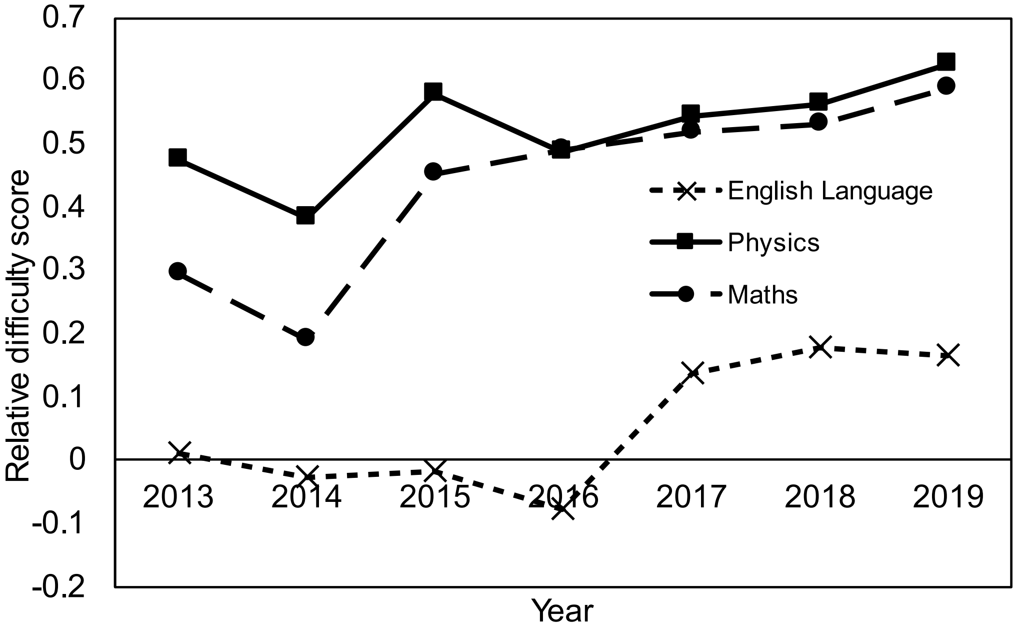

Figure 9. Relative difficulty of GCSE computer science compared with other subjects over time – A/7 grade.

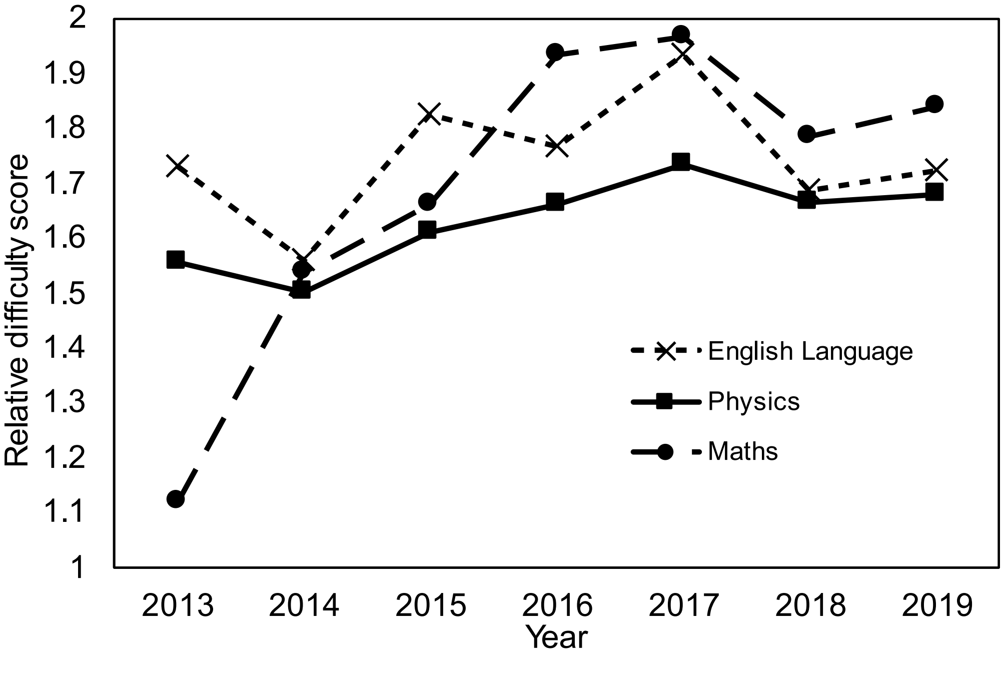

Figure 10. Relative difficulty of GCSE computer science compared with other subjects over time – C/4 grade.

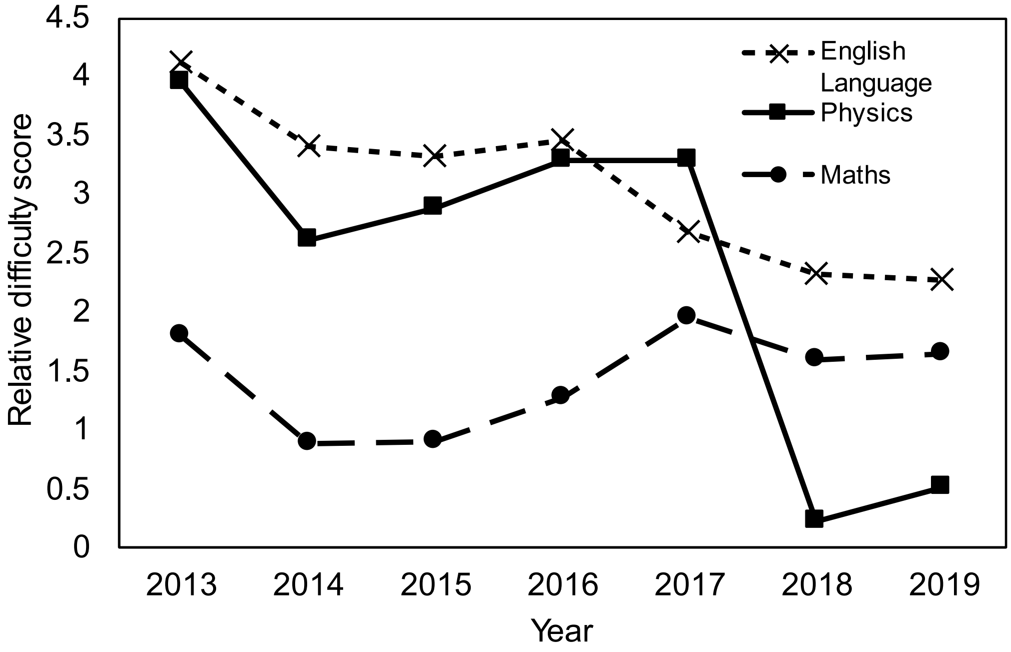

Figure 11. Relative difficulty of GCSE computer science compared with other subjects over time – G/1 grade.

Figure 9, Figure 10 and Figure 11 show the relatively difficulty of English language, physics, and maths compared with computer science based on the statistical definition described above. An increase in the relative difficulty score indicates that, based on these measures, computer science has become more difficult relative to the comparison subject.

As can be seen from Figure 9 and Figure 10, there has been a general upward trend in the difficulty of computer science over time relative to other subjects at both grade A/7 and C/4. At grade A/7 this is an increase of between 0.19 and 0.40 on the Rasch scale from 2014 to 2019, and at grade C/4 this is an increase of between 0.16 and 0.30 over the same time period. The absolute change in the score for computer science over this period is 0.1 at grade C/4 and 0.29 at grade A/7. Converting these scores to grades is challenging but, on average, a Rasch score value of 1.4 equates to approximately one grade on the 9 to 1 scale in each year across subjects, so the above represents an increase in difficulty of somewhere between 0.12 and 0.21 of a grade between 2014 and 2019 at grade C/4 and 0.14 and 0.28 at grade A/7.

At grade G/1 there are mixed results (Figure 11), with some evidence suggesting a reduction in the difficulty of GCSE computer science relative to other subjects. Using a similar procedure to convert the Rasch score to grades, this would suggest a reduction in the difficulty at grade G/1 of an average of 0.18 grades between 2014 and 2019, although this varies from between -1.5 grades (compared with physics) and +0.55 grades (compared with maths), depending on the comparison subject.

Kelly’s method

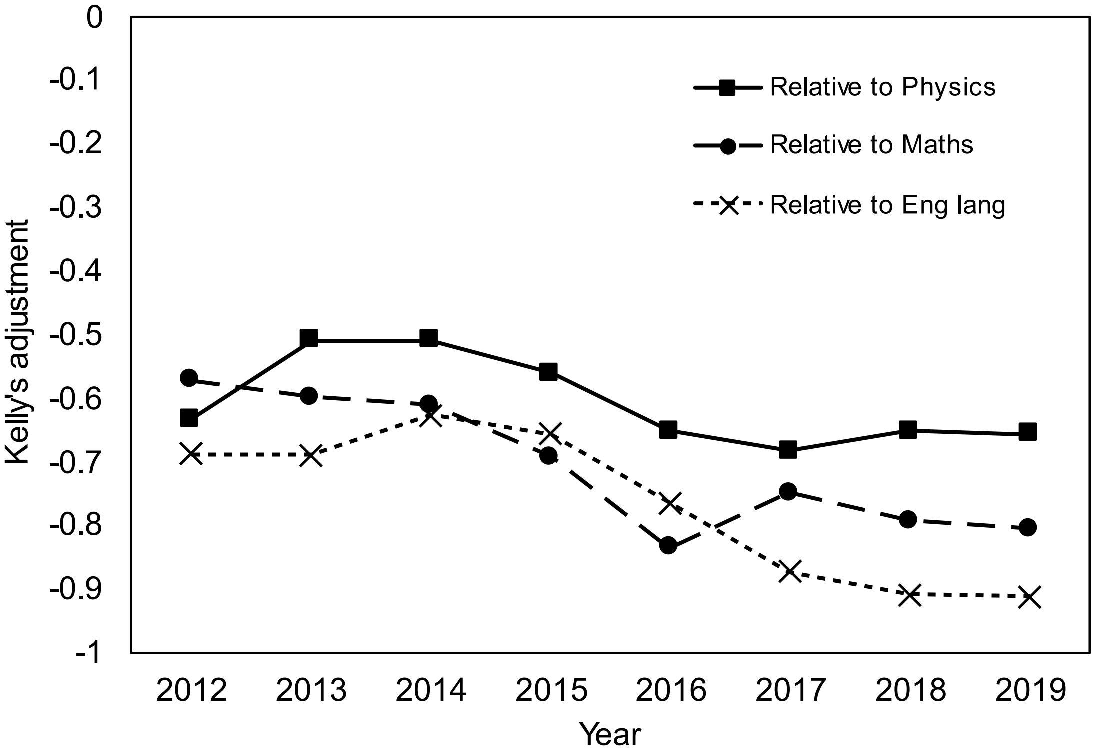

Figure 12. The relative difference between the average grade in computer science and their grade in other subjects.

The Kelly’s method analysis indicates that the difference in difficulty between computer science and the other 3 subjects included here has increased over time, particularly between 2015 and 2017 (Figure 12). The analysis estimates that between 2014 and 2019, students received a grade between 0.15 and 0.24 lower in computer science on an A* to G scale compared with the other subjects. Following a simple proportional scaling to the 9 to 1 scale, this equates to between 0.17 and 0.27 grades, with an average adjustment of 0.18 across all other GCSE subjects included in the analysis in each year.

Summary

Overall, both of the above methods indicate that students have generally gained increasingly lower outcomes in GCSE computer science compared with other GCSE subjects over time. It is also worth reiterating that here we are not focussed on the absolute difference in scores between the different subjects, which as discussed previously can arise for a number of different reasons, but the relative change over time. These relative changes could indicate a change in standards, representing an increase in the difficulty of GCSE computer science over time. However, this relative change in subject outcomes could also be due to other factors which could legitimately result in a change in outcomes in different subjects, such as students' preparedness for the assessments, which cannot be controlled for by this method.

Strand 1. Analysis 4. Progression analysis

Aim

One of the stated purposes of GCSEs is to prepare students for further study. The aim of this analysis is to identify if the relationship between GCSE and A level results in computer science has changed through time. If we assume that the standard of the A level has not changed, then the relationship between GCSE results and A level results should provide an indication of whether the value of a GCSE grade in indicating likely success at A level has changed through time. That is, do students with a particular grade in the GCSE show greater attainment in computer science in some years rather than others, leading to better (or worse) A level outcomes.

If the GCSE has become more difficult then we might expect to see students with the same GCSE grade performing better in the A level over time, as they have higher underlying attainment in the subject than students receiving the same grade in previous years. Conversely, we may expect that students receiving the same A level grade may have, on average, lower GCSE outcomes over time.

It is worth reiterating, however, that a key assumption of this analysis is that grading standards have not changed in the A level through time – an assumption that we do not test here. There may also be an interaction with centre entry policies for A level courses, which cannot be controlled for. However, unlike the GCSE, A level computer science is not a new subject, and there has been no systematic change in the qualification during the period of interest that would suggest that this assumption may be problematic.

Method

A level data was taken from the NPD for years 2014 to 2019 and filtered to 18-year-old students taking computer science. This was then matched to students’ GCSE computer science results from 2 years previous using their unique student ID.

The proportion of GCSE computer science students who went on to take the A level in the same subject was calculated in each year. The inverse was also calculated, that is, what proportion of A level students had previously taken the GCSE.

For the purposes of this analysis students’ A level grades were converted to numerical values with grades A* to E converted to a numeric 6 to 1 scale, respectively. For those that did take the A level, in each year the mean A level grade was calculated for students with different GCSE grades. We also calculated the proportion of students receiving at least a grade C at A level for students with each

GCSE grade. The mean grade students received in the GCSE was then calculated for students receiving different A level grades.

For these analyses, students who took their GCSE at a centre offering the qualification for the first time were removed. We only include data from students taking the GCSE until 2017, as after this, students would have received A level grades based on teacher judgements due to the cancellation of exams during the pandemic.

Finally, we created a linear model to examine the relationship between GCSE and A level computer science grade over time. The model took the form below:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 X_{ij} + u_j + \epsilon_{ij}$$

In this model, the dependent variable was A level grade *(y),* the key predictor was Year (x_1), and students' GCSE computer science grade was included as a covariate (x_2). The model also included a series of control variables (X), for KS2 prior attainment, ethnicity, gender, SEN status, FSM eligibility, language spoken and centre type. A random effect was included to take into account the clustering of students within centres (u). This is to control for the fact that student outcomes within the same centre are not independent of each other and therefore prevents the overestimation of model effects.

If we see the estimated A level grade from the model for each year increase over time (while keeping GCSE computer science attainment stable), this would indicate that students who attain a similar GCSE score are performing better at A level. For this analysis we only include 4 years for those sitting their GCSEs between 2014 and 2017 due to the small numbers of students available for analysis prior to 2014.

Results

Table 8. Percentage of students that took A level computer science who had previously completed GCSE computer science.

Year of sitting A level	N took A level	N previously took GCSE	Percentage previously took GCSE
2014	3781	234	6.2%
2015	4883	511	10.5%
2016	5473	1546	28.2%
2017	7289	3776	51.8%
2018	9259	6240	67.4%
2019	10076	7287	72.3%

Table 9. Percentage of students who took GCSE computer science who went on to do A level computer science.

Year of sitting GCSE	N took GCSE	N subsequently took A level	Percent subsequently took A level
2012	1745	234	13.4%
2013	4179	511	12.2%
2014	16011	1546	9.7%
2015	33773	3776	11.2%
2016	61751	6240	10.1%
2017	67374	7287	10.8%

As shown in Table 8, the proportion of students taking A level computer science who previously completed the GCSE has increased over time, from 6.2% in 2014 to 77.2% in 2020. This may reflect the increasing entry size to GCSE computer science over this period. The inverse is not true, however, and the proportion of students who took GCSE computer science going on to do A level has remained broadly stable (Table 9).

Table 10. Mean A level score for students receiving different GCSE grades over time. Values from cells with less than 100 students have been removed. Year indicates the year students took the GCSE.

GCSE Grade	2012	2013	2014	2015	2016	2017
A*			4.52	4.60	4.62	4.60
A		3.57	3.52	3.46	3.42	3.53
B			2.28	2.46	2.45	2.52
C				1.77	1.74	1.81
D						1.65
E						

Table 11. Proportion of students receiving at least a C at A level for students receiving different GCSE grades over time. Values from cells with less than 100 students have been removed. Year indicates the year students took the GCSE.

GCSE Grade	2012	2013	2014	2015	2016	2017
A*			0.942	0.963	0.965	0.960
A		0.795	0.792	0.788	0.780	0.811
B			0.429	0.476	0.474	0.509
C				0.268	0.252	0.261
D						0.284
E						

Table 10 and Table 11 present the mean A level grade achieved and proportion of students achieving A level grade C or above both differentiated by GCSE grade achieved. These analyses do not show any strong patterns for a change in the relationship between GCSE and A level outcomes over time. There is some slight indication that those who received grade A, B or C at GCSE in 2017 may have had higher attainment in computer science than those who received an A, B or C in 2016. This is because, as shown in Table 10, they gained a slightly higher mean A level grade and their probability of attaining at least a C at A level increased. However, between 2013 and 2016 students who gained an A at GCSE received lower mean A level grades each year, which may suggest higher performing students in the GCSE actually had lower attainment over time.

Table 12. Mean GCSE grade of students receiving different grades at A level. Values from cells with less than 100 students have been removed. Year indicates the year students took their GCSEs.

A level Grade	2012	2013	2014	2015	2016	2017
A*					7.80	7.74
A			7.43	7.43	7.42	7.40
B			7.09	6.97	6.96	6.91
C			6.64	6.51	6.58	6.43
D			6.37	6.20	6.18	6.06
E				5.76	5.97	5.80

Table 12 shows the mean grade of GCSE students achieving each grade at A level. Here, there is some indication that students receiving higher A level grades in 2017 had slightly lower mean GCSE scores than they did in previous years, across all grades. This could indicate that students receiving these grades had higher computer science attainment than in previous years. For example, between 2014 and 2017 the mean GCSE grade of students attaining a B at A level dropped from 7.09 (just over an A at GCSE) to 6.91 (a high B at GCSE). Which may suggest that on average students with lower GCSE grades are displaying the same level of ability in computer science as those who achieved slightly higher grades in previous years, as represented by their A level grade.

Figure 13 shows the high-level output from the linear model. This shows changes in the A level grades achieve by students controlling for differences in KS2 attainment, centre type and student background characteristics between years. Full model outputs can be seen in appendix A. The results of the linear model showed some indication that students in 2017 who achieved the same GCSE computer science grade as those in 2014, received a higher A level grade by approximately 0.1 grade (β=0.107, *p<0.05)*. Students in 2014 would need to have a grade 0.13 higher in the GCSE computer science (on an A*-G scale) to receive the same A level grade as similar students in 2017. Proportionally, this converts to around 0.15 grades on a 9 to 1 scale. However, Figure 13 indicates the effect is also not clearly linear, after controlling for other factors. Beyond the difference described between 2017 and 2014, taking into account the uncertainty in the model, there is not a clear trend over time.

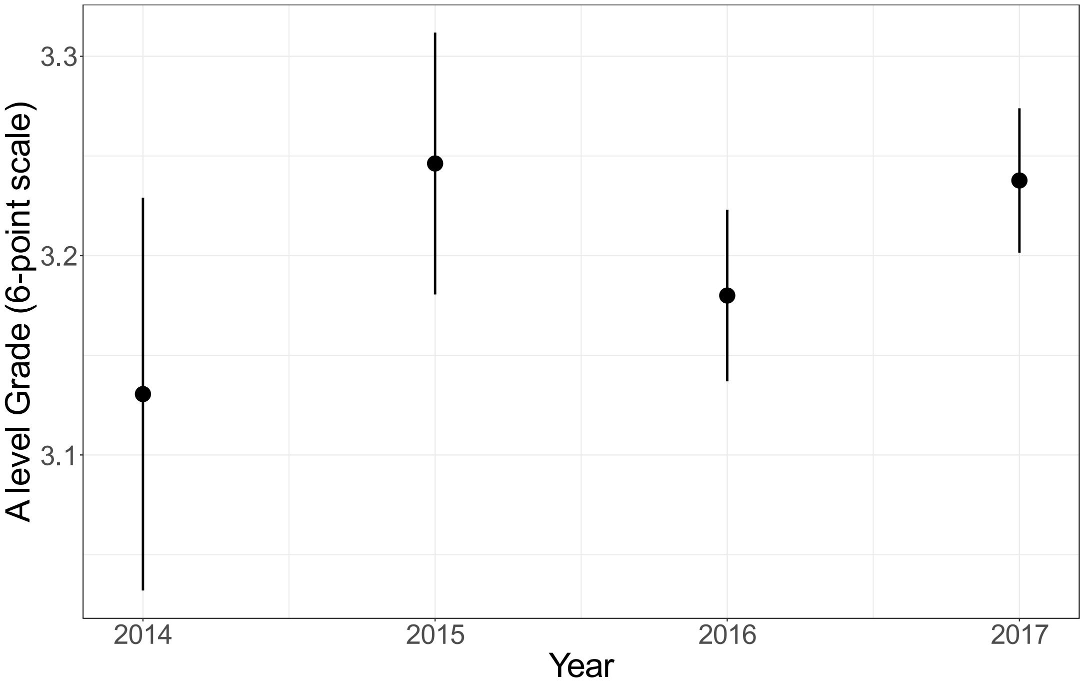

Figure 13. Marginal effects from linear regression model for reference group students by year.

Summary

The aims of the analysis presented in this section were to identify whether the relationship between students' performance in GCSE computer science and their success in A level computer science has changed over time. To summarise, the above analysis shows some evidence that students with a similar GCSE grade, and other characteristics, performed better at A level over time. This might indicate that these students are more able at computer science, suggesting the GCSE standard may have become more challenging, however, these effects are subtle. As discussed above, this interpretation relies on the assumption that the standard of the A level has not changed. These results could also indicate changes in centres' entry policies for their A level courses.

Strand 1. Analysis 5. Simulated Predictions

Aim

As discussed in the introduction, prior-attainment based statistical predictions are regularly used to support the setting of grade boundaries each exam series, alongside expert judgement and other technical evidence. Details of this approach are described in the section *Operationalising the setting and maintenance of standards*. A key assumption of this method is that the cohort of students in the current year is similar to the cohort of students that took the qualification in the reference year in all ways that would affect their outcomes, except their prior attainment distribution. Therefore, that we can reasonably expect the relationship between prior attainment and outcomes to be the same, on average.

The evidence discussed in Strand 1 Analysis 2 described the circumstances that led to the change in value-added relationship in GCSE computer science over time. The aim of this piece of analysis is to quantify the impact of those changes while considering the change in the cohorts prior (or concurrent) attainment distribution.

For this analysis we generate predictions based on different reference years. We generate 2 sets of predictions, accounting for students' prior attainment (KS2 score) and concurrent attainment (mean GCSE score) respectively. If there are large differences in the predictions generated depending on the reference year this may indicate that standards have changed between years. However, it could also indicate that other factors have changed that would affect outcomes, such as the makeup of the cohort or teaching time dedicated to the subject.

One additional factor that we can attempt to control for here is how familiar teachers are with the qualification. As discussed previously, outcomes from centres entering students for the first time may be lower if students at those are less well prepared for the assessments. We therefore look at the impact of excluding these 'new' centres from the predictions generated, as these centres may have a different value-added relationship.

Method

We calculated predictions using a range of reference years (2012-2018) to predict outcomes in 2019, but otherwise following the same methodology as would be typically used by AOs.

For 2015 around 20% of the cohort were missing KS2 prior attainment data (due to boycotts of the KS2 assessments in 2010), meaning using this group as a reference for prior-attainment based predictions may be less reliable. Therefore, 2 sets of predictions were produced. The first set of predictions included all 16-year-old students with prior attainment data, excluding students at selective and independent centres. This is typically the approach taken when GCSE predictions are produced in practice since students at selective and independent centres have a different relationship between prior attainment and GCSE outcomes to other centres. The second set of predictions was produced using concurrent attainment (that is, mean GCSE), rather than prior attainment, and included all 16-year-old students that had taken at least 3 GCSEs. The second set of predictions are therefore based on the relationship between a student's mean GCSE grade in the other subjects that they took concurrently, and their grade in computer science. For this analysis, students at all centre types were included.

A normalised KS2 prior attainment score was calculated for each student replicating the process for calculating prior-attainment-based predictions used in awarding. A similar process was followed to produce a 'concurrent attainment' score based on the students mean GCSE score (converted to an 8 to 1 scale) across all of the other subjects each student studied at GCSE.

For each year normalised prior or concurrent attainment scores were divided into 10 equal deciles based on results for the whole GCSE cohort. For each reference year, the proportion of students in each decile attaining each grade in GCSE computer science was calculated in an outcome matrix. For 2019, we then calculated how many students fell into each attainment decile. The outcome matrix was then used to predict how many of the students in each decile in 2019 would receive each grade, based on the proportions in the reference year. The number of students predicted to receive each grade was then summed over all deciles and used to calculate a cumulative percentage predicted outcome at the grades A/7, C/4 and G/1.

Finally, based on the results of the other analyses and the differences observed in the patterns between new and existing centres, a set of predictions was produced excluding 'new' centres in both the reference year and the current year (2019) for each prediction. New centres were defined as those with entries to the qualification for the first time in the year being analysed.

Results

Prior attainment based predictions

Table 13. Simulated predictions for 2019 based on different reference years – matched candidates only, excluding selective and independent centres. 'Difference'

indicates the percentage point difference between each prediction and actual outcomes in 2019.

Reference Year	Matched Entry	Cum. % Predicted A/7	Cum. % Predicted C/4	Cum. % Predicted G/1	Difference A/7	Difference C/4	Difference G/1
2013	3210	17.3	59.8	96.3	-1.0	0.0	-0.5
2014	13100	18.4	61.8	97.0	0.2	2.0	0.3
2015	20869	17.6	62.0	97.4	-0.7	2.1	0.6
2016	53297	18.1	59.0	96.1	-0.1	-0.8	-0.6
2017	58042	17.9	58.5	96.0	-0.4	-1.4	-0.7
2018	59718	18.1	59.2	96.7	-0.1	-0.7	-0.1
2019	62287	18.3	59.8	96.7	0.0	0.0	0.0

Table 14. Simulated predictions for 2019 based on different reference years – matched students excluding students at new centres and selective and independent centres. ‘Difference’ indicates the percentage point difference between each prediction and actual outcomes in 2019.

Reference Year	Matched Entry	Cum. % Predicted A/7	Cum. % Predicted C/4	Cum. % Predicted G/1	Difference A/7	Difference C/4	Difference G/1
2013	1116	21.5	65.7	97.8	3.0	5.5	1.0
2014	3601	22.7	66.7	97.3	4.2	6.6	0.5
2015	10450	19.3	64.8	97.7	0.8	4.6	0.9
2016	33555	20.1	62.0	96.9	1.7	1.9	0.1
2017	50693	18.4	59.1	96.1	-0.1	-1.1	-0.7
2018	56031	18.2	59.4	96.7	-0.2	-0.8	-0.1
2019	59047	18.5	60.2	96.8	0.0	0.0	0.0

The above analyses indicate that when including all students, prior-attainment-based predictions based on 2014 outcomes would suggest outcomes around 2pp higher at grade C/4 than actual outcomes in 2019 (Table 13). The difference at grades A/7 and G/1 were much smaller and less consistent between years. When students at centres which had never offered GCSE computer science before were excluded, the size of the difference increased in most years (Table 14). When 2014 was used as the reference year, predictions were almost 7pp higher at grades C/4 and 4pp higher at A/7 than actual outcomes in 2019. This suggests that students at new centres tend to receive, on average, lower GCSE results relative to their prior attainment, and if excluded would have led to higher predictions for non-new centres.

Concurrent attainment based predictions

Table 15. Simulated predictions for 2019 based on different reference years – all students. 'Difference' indicates the percentage point difference between each prediction and actual outcomes in 2019.

Reference Year	*Matched Entry*	*Cum. % Predicted A/7*	*Cum. % Predicted C/4*	*Cum. % Predicted G/1*	*Difference A/7*	*Difference C/4*	*Difference G/1*
2013	3756	23.3	65.9	97.1	1.9	3.3	0.3
2014	15092	23.5	65.3	97.1	2.1	2.7	0.3
2015	31928	22.2	65.3	97.3	0.8	2.7	0.5
2016	59334	22.5	62.5	96.4	1.0	-0.1	-0.4
2017	65897	21.9	62.0	96.3	0.4	-0.6	-0.5
2018	68966	21.7	62.9	96.9	0.2	0.3	0.1
2019	74530	21.5	62.6	96.8	0.0	0.0	0.0

Table 16. Simulated predictions for 2019 based on different reference years – excluding students at new centres. 'Difference' indicates the percentage point difference between each prediction and actual outcomes in 2019.

Reference Year	*Matched Entry*	*Cum. % Predicted A/7*	*Cum. % Predicted C/4*	*Cum. % Predicted G/1*	*Difference A/7*	*Difference C/4*	*Difference G/1*
2013	1530	26.3	70.0	97.5	4.8	7.2	0.6
2014	4612	27.6	70.1	97.4	6.1	7.3	0.5
2015	16309	24.3	68.0	97.8	2.8	5.2	0.9
2016	37995	24.2	65.0	97.1	2.7	2.2	0.3
2017	57537	22.1	62.2	96.4	0.6	-0.6	-0.4
2018	64507	21.5	62.9	96.9	0.0	0.1	0.0
2019	69949	21.5	62.8	96.8	0.0	0.0	0.0

Concurrent-attainment-based predictions show a similar pattern to prior-attainment-based predictions, but with slightly higher predictions than those generated using prior attainment. When including all centres, predictions for 2019 based on 2014 outcomes were around 3pp higher at grade C/4 and 2pp at grade A/7, than actual outcomes (Table 15). After new centres had been removed this prediction was around 7pp higher than actual outcomes at C/4 and around 6pp higher than actual outcomes at grade A/7 (Table 16).

Summary

These analyses suggest that predictions based on 2014 outcomes would have been higher than actual outcomes in 2019, regardless of whether the predictions are based on prior attainment or concurrent attainment. This indicates that the value-added relationship has changed such that there is lower value-added relationship for students taking GCSE computer science over time, that is, the same prior or concurrent attainment is associated with lower grades in 2019 compared with 2014. Further, the size of this effect increased when new centres were removed. This suggests that students at new centres tended to perform less well than students at other centres who had a similar prior or concurrent attainment.

In practice predictions are only used to guide awards and it therefore cannot be assumed that a different prediction would have led to different outcomes, particularly in years where examiners recommended grade boundaries below predictions anyway. However, it is not possible to know how different statistical evidence may have influenced the final judgements of examiners in a particular year.

It is also worth considering whether the cohort in each reference year was similar enough to that in the ‘current’ year (2019) to expect a similar value-added relationship. The descriptive analysis presented previously indicated that there have been a large number of changes to the cohort since 2014. These changes could have led to legitimate differences in the value-added relationship over time. The reference year for predictions needs to be carefully considered to ensure the cohort is representative of the current year. A larger number of years between the reference year and the current year results in a higher likelihood that the cohort, and therefore outcomes, may have changed for legitimate reasons.

Disentangling these legitimate changes in outcomes from illegitimate ones is challenging. Therefore, in the next section we carry out some more sophisticated modelling aiming to control for some of these potentially confounding effects.

Strand 1. Analysis 6. Modelling of outcomes over time

Aim

In this section we present a series of models of outcomes in GCSE computer science in each year, which as in the previous analysis, control for concurrent or prior attainment, but also for a variety of other student and centre characteristics which may be related to outcomes. The aim of this modelling is to disentangle some of the factors which may be related to changes in outcomes over time, but that are not appropriate to be factored into the statistical predictions, to identify if the changes in outcomes can be reasonably accounted for by these factors.

Primarily we controlled for students prior or concurrent attainment, however we also controlled for other student characteristics which may be related to outcomes. We calculated both a model of GCSE grade on a linear scale, and models of the probability students would receive at least a grade A/7, grade C/4 or grade G/1. If the analysis indicates that outcomes differed between years, after controlling for other variables which might be related to outcomes, this may suggest that standards have changed between years.

However, as discussed previously there may be other factors influencing outcomes over time which do not directly relate to observable student characteristics. We therefore aim here to control for 2 additional factors which could be related to outcomes. Firstly, the experience of centres of delivering the qualification. We control for this by removing centres entering students for only the first or second year from the analysis. Secondly, outcomes could differ if there are qualitative differences between centres entering in different years. We therefore carry out some further models only including the same set of centres in each year. If in these models, we still see a change in outcomes over time, this suggests that there has been a change in standards which cannot easily be explained by other factors.

Method

A numeric grade variable converting both A* to G grades and 9 to 1 grades to an 8-point scale was created (see Table 7 in Strand 1 Analysis 3) along with binary variables indicating if each student received at least a grade G/1, C/4 or A/7. A variable was also created indicating how long each centre had been delivering GCSE computer science, by calculating the number of years since a student at that centre had first received a grade. This was then converted into a binary variable (new/not new centres). For this analysis a slightly more conservative approach was used to the previous analyses and new centres were classed as those entering students for the first or second year.

The primary models used a linear relationship, with students' numeric GCSE grade as the target variable, and a series of logistic regression models evaluating the probability of a student receiving at least each grade – G/1, C/4 and A/7. All models included *Year* as the key predictor. Models were developed using both prior attainment (standardised KS2 score) and concurrent attainment (standardised mean GCSE). These variables were trialled as both continuous variables and as categorical variables (that is, attainment deciles), all producing similar results, however the continuous models resulted in better model fit. Prior attainment data was missing for around 20% of students in 2015 due to boycotts of KS2 assessments 5 years earlier. Therefore, we focus on the results of the concurrent attainment models in the main text and figures (see appendix B for all full model results).

All of the models controlled for other student characteristics, namely; gender (male/female), SEN status (SEN, no SEN, missing), FSM eligibility (yes, no, missing), primary language spoken (English, other, missing), ethnic group (Asian, Black, Chinese, Mixed, White, other, missing) and centre type (college, selective, independent, mainstream, missing). A random effect of centre number was included in all models to control for centre level clustering. Models only included 16-year-old students with a valid grade, from England only, with prior or concurrent attainment data available (depending on the model). See Table 17 for a summary of the sample included in the analysis.

Table 17. Summary of sample used for modelling of outcomes over time.

Year	***N students - Prior attainment models (all centres)***	***N students - Concurrent attainment models (all centres)***
2012	1,583	1,614
2013	3,876	3,756
2014	14,768	15,092
2015	23,322	31,928
2016	57,163	59,334
2017	62,321	65,897

2018	65,167	68,966
2019	68,814	74,530

Results

The majority of models indicated that *Year* had a statistically significant effect on the probability of students attaining key grades, except for the models at grade G/1 (see Table 18). Adding *Year* to models also improved model fit, however, the additional explanatory power was relatively small (increase in R^2/pseudo-R^2 between 0.1pp and 0.9pp). This likely represents that the main predictor of a students' outcome in an exam is inherently their own ability and other variables only have a weak relationship with outcomes in comparison.

For each of the models we estimate what the difference in mean grade predicted by the model, or probability of receiving key grades would be, for the full cohort of students included in the model in 2019, using the estimated model coefficients for 2014. This estimate takes into account the effect of changes in the distributions of different subgroups of students, prior attainment and centre types and so gives an estimate of the size of the *Year* effect on actual outcomes (see Table 18).

Table 18. Summary of Year model effects from various different models using concurrent attainment.

Model	**Restriction**	**Year-2019 coefficient [Ref 2014] (SE)**	**Estimated difference in outcomes from 2014 predicted for 2019 cohort**
Linear	All centres	-0.12 (0.01)***	-0.11
Linear	Excluding new centres	-0.41 (0.03)***	-0.41
Linear	2014 centres only	-0.31 (0.03)***	-0.31
Linear	2015 centres only	-0.33 (0.02)***	-0.33
A/7 Grade	All centres	-0.02 (0.03)	-0.17pp
A/7 Grade	Excluding new centres	-0.48 (0.08)***	-3.47pp
A/7 Grade	2014 centres only	-0.40 (0.10)***	-4.52pp
A/7 Grade	2015 centres only	-0.30 (0.06)***	-3.40pp

C/4 Grade	All centres	-0.06 (0.03)*	-0.76pp
C/4 Grade	Excluding new centres	-0.77 (0.10)***	-8.72pp
C/4 Grade	2014 centres only	-0.39 (0.11)***	-4.43pp
C/4 Grade	2015 centres only	-0.57 (0.07)***	-5.60pp
G/1 grade	All centres	+0.09 (0.07)	-0.15pp
G/1 grade	Excluding new centres	-0.92 (0.33)**	-1.17pp
G/1 grade	2014 centres only	-0.27 (0.39)	-0.26pp
G/1 grade	2015 centres only	-0.22 (0.20)	-0.16pp

Note. Statistical significance is indicated by p<0.001 ‘***’ p<0.01 ‘**’p<0.05 ‘*’

It is notable that in all cases excluding new centres increases the estimated size of the *Year* effect. This suggests that including these centres may have masked a potential change in standards. The sections below discuss the different models in detail. Figures show the predicted mean grade or predicted probability of receiving the key grade or above in each year for students in the reference group (that is students with an average attainment score, white, male, not FSM eligible, English speaking, not registered as SEN and attending a mainstream school).

All centre models

We start by looking at the linear models. These models are based on an 8-point grade scale equivalent to A* to G.

Figure 14 shows the output of the model including all 16-year-old students from all centres and gives an indication of the estimated mean grade of similar students in the reference group in each year after controlling for other factors.

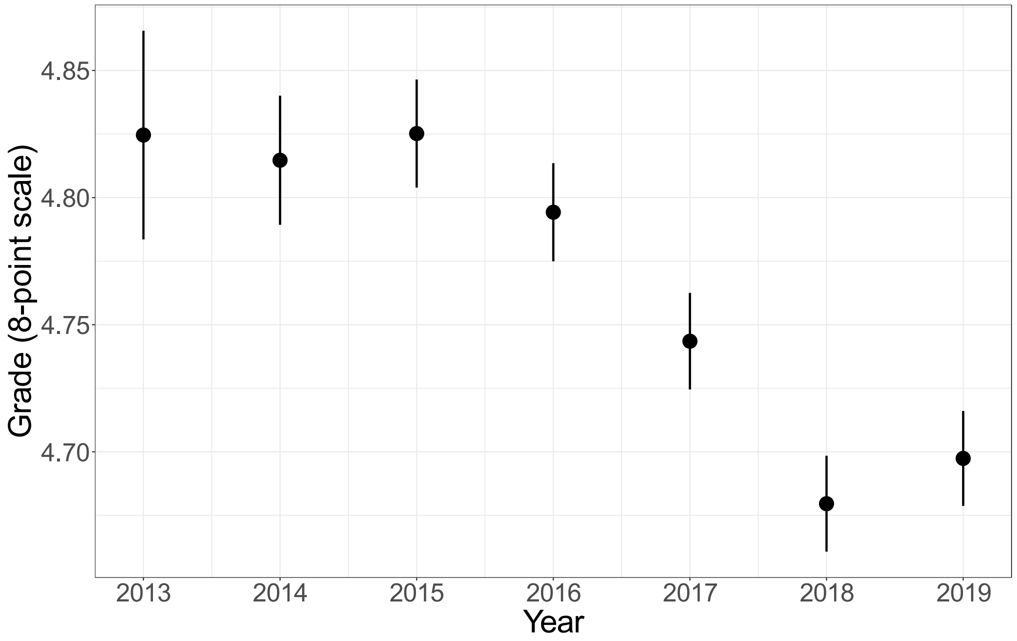

Figure 14. Estimated mean grade for students in the reference group for students with average concurrent GCSE attainment in each year. Includes all centres.

From Figure 14 a clear pattern of declining mean grade can be seen between 2015 and 2018. This is after controlling for student characteristics, centre type and students' attainment in other GCSEs. Although this effect is relatively small, with an average estimated difference in outcomes of 0.12 grades between 2014 and 2019 for an average attaining student, this is still a notable change, representing over one in 10 students gaining a grade lower in 2019 when compared with 2014. However, this model does not account for the effects discussed previously which could impact on outcomes; whether centres are new to delivering the qualification or unmeasured differences between centres in different years such as changes to teaching quality. For the next set of models, we therefore first exclude students at centres that have entered students for fewer than 2 years previously.

Model excluding new centres

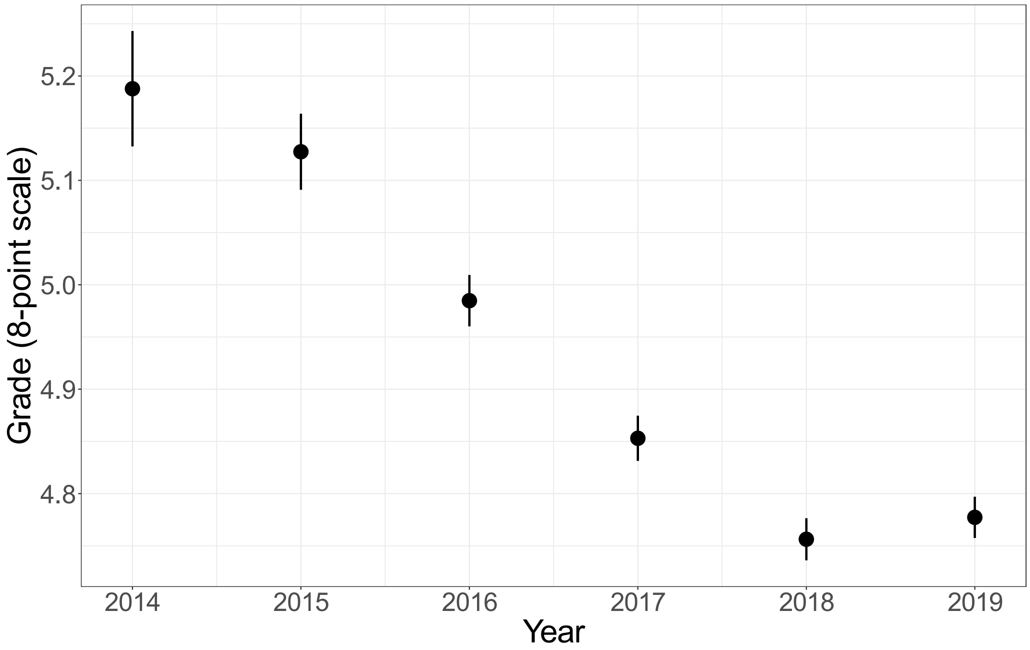

Figure 15. Estimated mean grade for students in the reference group for students with average concurrent GCSE attainment in each year. Only includes students at centres offering GCSE computer science for the third year or more.

Figure 15 shows that after excluding ‘new’ centres, the effect of declining outcomes becomes more pronounced. This suggests that the different value-added relationship in these centres, combined with differing numbers of new centres in each year, may have masked a larger shift in standards. This model estimates that this shift results in a difference of mean grade of 0.41 grades (once converted to a 9 to 1 scale) between 2014 and 2019.

However, this model is still not accounting for potential qualitative differences between centres taking up the qualification in different years, for example related to teaching quality or resources. For the final set of models, we therefore only include centres entering students to the qualification in every year that is included in the model. For the period between 2014 and 2019 this results in only 85 centres being included in the analysis, so we therefore repeat the analysis for centres entering students every year between 2015 and 2019, which increases the sample to 205 centres.

Models restricted to same set of centres in each year

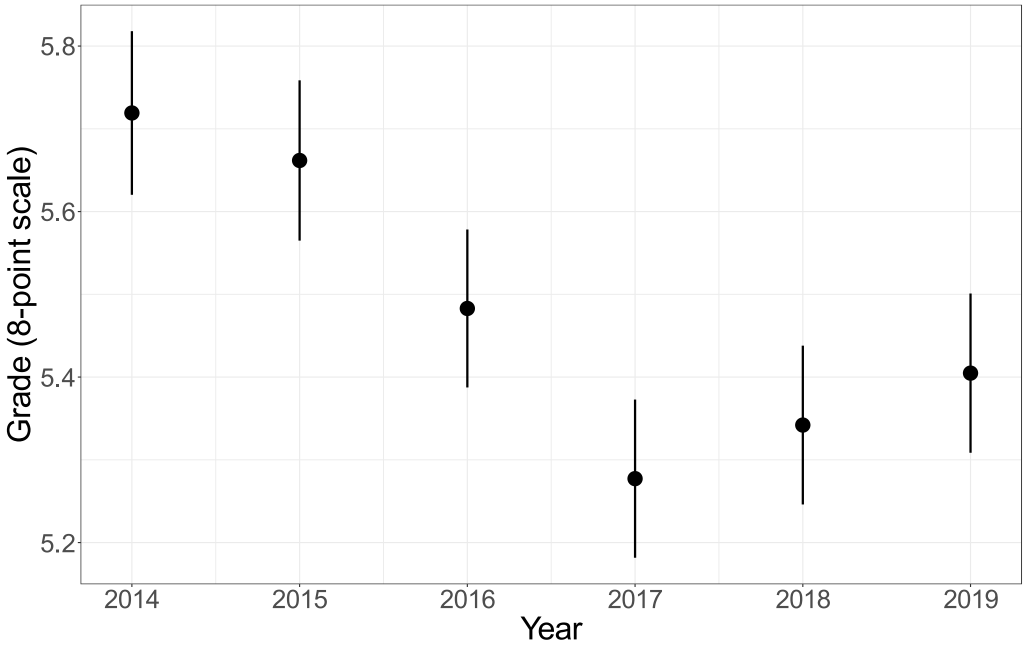

Figure 16. Estimated mean grade for students in the reference group for students with average concurrent GCSE attainment in each year. Only includes students at centres offering GCSE computer science for the third year or more who entered students in every year 2014 to 2019.

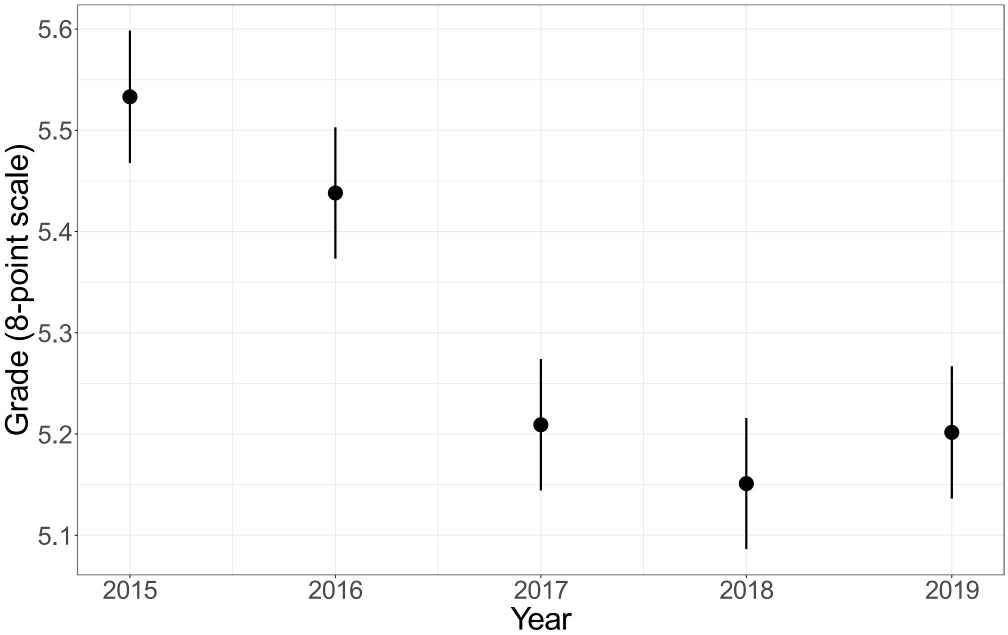

Figure 17. Estimated mean grade for students in the reference group for students with average concurrent GCSE attainment in each year. Only includes students at

centres offering GCSE computer science for the third year or more who entered students in every year 2015 to 2019.

Figure 16 and Figure 17 show that, even when the analysis only includes centres who entered students in every year, the estimated mean grade still declines between 2015 and 2017 by around 0.3 grades on average (see Table 18 above). This decline cannot be explained by effects due to centre unfamiliarity, as new centres were excluded from the model, and it also seems unlikely that teaching quality would have consistently declined in this same set of centres over time. There are other factors that might have changed though such as entry policies for the subject or factors relating to student preparation or motivation during the period. It seems unlikely, however, that these effects would be consistent across centres.

Figure 18 and Figure 19 show outputs from logistic regression models estimating the probability of students attaining the key grades A/7 and C/4 or above. Like the previous model, these models only include centres with entries in all years 2015 to 2019 who first entered students in 2012 or 2013. We have not included the figures for grade G/1 or the models for centres entering students every year between 2014 and 2019 as the sample sizes for these models were small and therefore the models were unreliable.

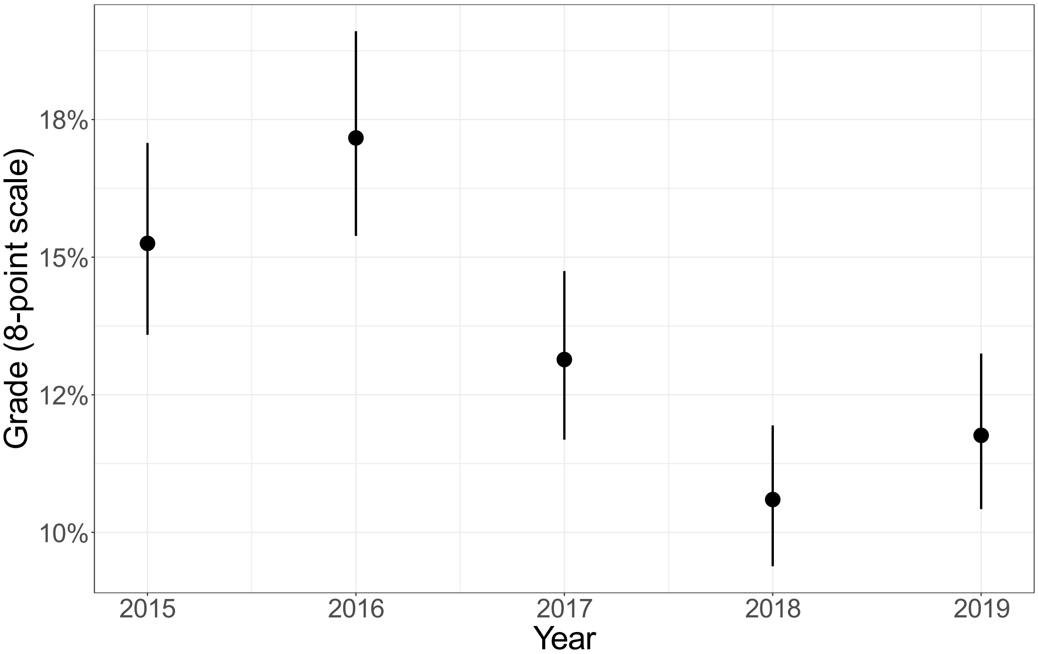

Figure 18. Estimated probability of attaining an A/7 or above for students in the reference group for students with average concurrent GCSE attainment in each year. Only includes students at centres offering GCSE computer science for the third year or more who entered students in every year 2015 to 2019.

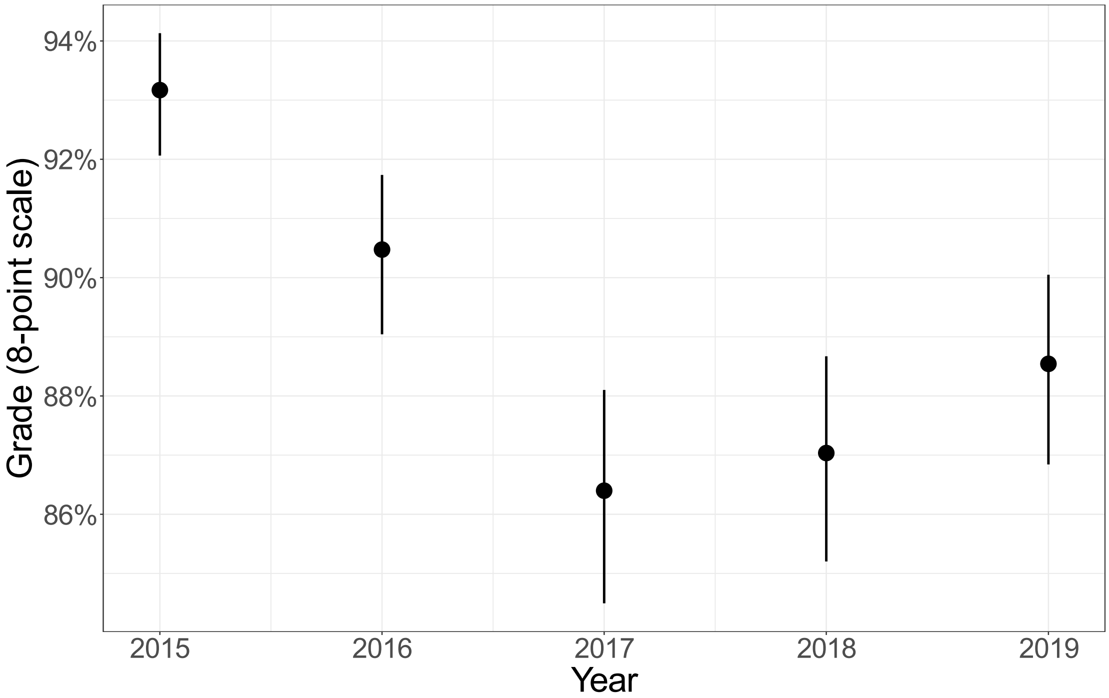

Figure 19. Estimated probability of attaining an C/4 or above for students in the reference group for students with average concurrent GCSE attainment in each year. Only includes students at centres offering GCSE computer science for the third year or more who entered students in every year 2015 to 2019.

The same pattern of declining outcomes can be seen at both grades A/7 and C/4. The models estimate the size of the difference as 3.4pp fewer students attaining an A/7 in 2019 compared with 2015 and 5.6pp fewer students attaining a C/4. Interestingly, the main decrease is slightly later for grade A/7 (occurring between 2016 and 2018), whereas for grade C/4 it occurs between 2015 and 2017.

For grade G/1, the modelling did not consistently indicate a statistically significant difference in the probability of students achieving a grade G/1 between 2019 and previous years (Table 18). A very small number of students receive a grade U, which means consistently estimating model effects is challenging. If an effect exists at this grade it is likely very small, the estimate for the model including centres with entries between 2015 and 2019 model suggested outcomes -0.16pp lower in 2019 compared with 2015.

Summary

In summary, these analyses indicate that after controlling as far as possible for changes in cohort characteristics, possible teacher unfamiliarity at new centres and changes between groups of centres entering the qualification in different years, there is a trend of lower results over time. This change is focussed around the grade C/4 boundary, with a slightly smaller effect at the grade A/7 boundary. In the following section we aim to carry out a similar analysis focusing on centre level outcomes over time for centres with entries in adjacent pairs of years.

Strand 1. Analysis 7. Common centres analysis

Aim

Although evidence shows that outcomes do vary for individual centres from year to year, it is expected that, on average, across a large number of centres, outcomes remain fairly stable when standards are maintained, assuming that the cohort of students entering from each centre remains fairly stable. Schools or colleges which offer the same qualification across 2 or more years are referred to as 'common centres' as they are centres 'in common' across those years. The aim of this section is to consider evidence relating to the maintenance of standards over time based on changes (or an absence of changes) in outcomes for these common centres.

As outlined previously, we are focusing here on whether outcomes have changed, and are not considering other factors such as the quality of student work. This analysis relies on the assumption that centres typically have similar outcomes between years, reflecting a similar level of student performance over time. Where outcomes do change it is expected to be statistically random, that is, some centres outcomes go up, but balanced by those where outcomes go down. This is built on the premise that students entering a qualification at the same centre will be similar from one year to the next, in terms of things like socioeconomic background, motivation and so on. It also assumes that centre-level factors will remain stable from one year to the next (at least on average), things such as entry policies, resourcing, and teaching quality. Therefore, if these assumptions hold, on average across the population of common centres, a large and consistent difference between the common centres predicted outcomes and the percentage of students who actually received each grade in each year may indicate a change in standards.

Method

The simplest approach to common centres analysis is to consider all centres that offer the qualification in a pair of adjacent years and to directly compare the outcomes across the 2 years. For this simple common centres approach we are assuming that the distribution of grades (that is, the proportion of students attaining each grade) remains the same on average across all the centres included.

This approach does not take into account any changes in the entry size from individual centres. For example, if higher performing centres increased their entries, whereas lower performing centres decreased their entries, we might expect overall outcomes to improve. Therefore, we can calculate a weighted common centres analysis by weighting the outcomes from individual centres to take such changes into account. In this case the assumption is made that the distribution of grades remains the same within each centre regardless of changes in entry size.

A more complex version of common centres analysis takes account of the change in the prior attainment distribution between pairs of years for the centres included in the analysis. This is achieved by applying the prediction matrix methodology, similar to that used to aid in setting standards in GCSEs and A levels, but only applied to the centres in the sample. This is referred to as a 'prior attainment adjusted' common centres analysis.

Given that we are looking historically we can also use a fourth alternative. This approach is similar to the prior attainment adjusted analysis but using concurrent attainment. This 'concurrent attainment adjusted' analysis utilises a prediction matrix based on the centres in the sample, but uses mean GCSE score to group students by ability in the place of KS2 prior attainment scores.

A common restriction applied to common centres analysis is to only include 'stable' common centres. Typically, these are classed as centres with a minimum number of students in each year, and/or those where the number of students has not changed by over a certain percentage. The rationale is that we might expect outcomes in these centres to be more consistent than in other centres. In practice, the effectiveness of these restrictions on improving prediction accuracy requires careful consideration. Previous analysis has shown the potential increased accuracy gained by restricting the sample to more stable centres, is often outweighed by the loss of sample size (Benton, 2013). However, we include them here for comparison and to potentially control for centres with large changes in entries in the early years of the qualification, where it may have a larger impact.

We applied all of the above methods to identify a range of potential predicted outcomes for each year based on each method; simple common centres, weighted common centres, prior attainment adjusted and concurrent attainment adjusted analyses. We also carried out each method with different levels of restriction of the sample of centres included. For the initial analysis, we include all centres with entries in each pair of consecutive years. For the 'stable' common centres analysis we carried out 2 versions, the first restricted the sample to only centres with a minimum of 10 students in each of the pair of years being analysed and whose entry did not fluctuate by more than 40% between the first and second year, for the second 'very stable' analysis we restricted to centres with at least 20 students and whose entry fluctuated by less than 15%.

For the prior attainment analyses we excluded selective and independent centres, as students at these centres tend to have a different relationship between their KS2 results and GCSE outcomes. Students without prior attainment or concurrent attainment data were also excluded from the respective analyses. For these attainment-adjusted analyses in each pair of years the first year was treated as the reference year. The standard methodology to produce prior-attainment-based predictions was applied here, but only for the subset of centres identified as common across years.

For each of the analyses we compared the common centres predicted outcomes against the actual outcomes at grades A/7, C/4 and G/1. However, pairs where there were less than 500 students retained in the sample in either year have been removed as the predictions are unlikely to be reliable. Therefore, predictions in most cases cover changes in outcomes during the period from 2014 to 2019, except for analyses using very stable centres, which include the period 2015 to 2019 and the prior attainment adjusted analysis with very stable centres, which only covers the period 2016 to 2019.

Similar to the previous analysis, we removed centres that have entered students for the assessments for less than 2 years prior to the ‘reference year’ in all methods as this is the period when their outcomes are most likely to change due to sawtooth-like effects.

Results

Figure 20, Figure 21 and Figure 22 show, for each year, the difference between the common centres predicted outcomes and the actual outcomes for the sample of centres included in each analysis. The figures are cumulative over time, to give an indication of the possible cumulative change in standards over time from 2014. A separate line is included for the combination of each method (simple, weighted, prior attainment adjusted and concurrent attainment adjusted) and each sampling approach (all common centres, stable centres and very stable centres). Table 19 then shows a summary across different methods of the difference between the predicted outcomes and actual outcomes in each year.

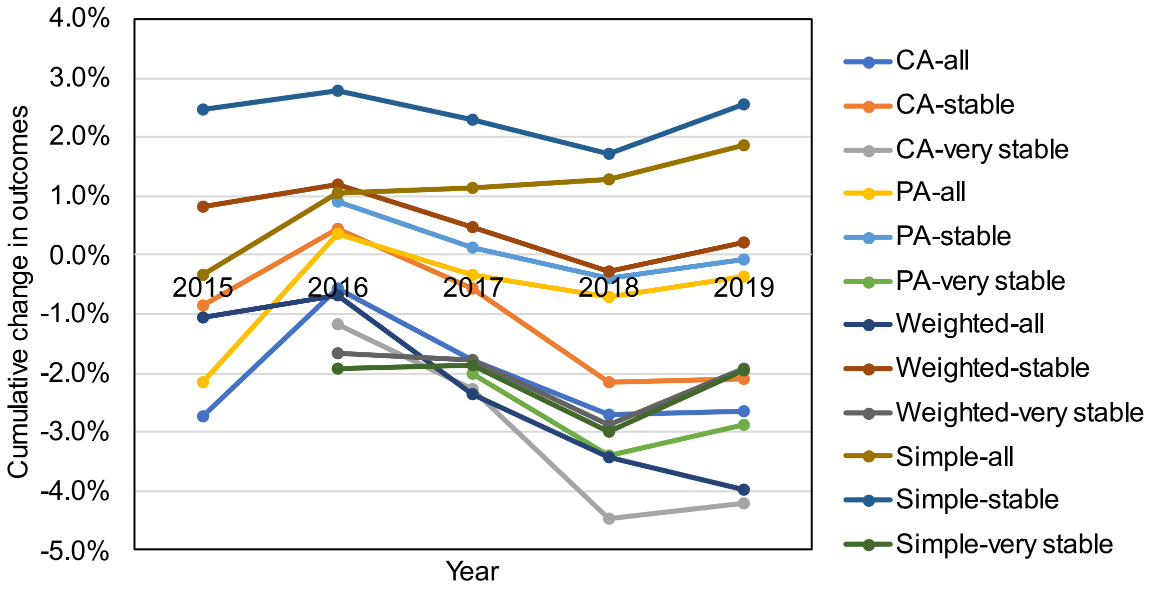

Figure 20. Cumulative difference between common centres predictions and actual outcomes over time by common centres method. Grade A/7.

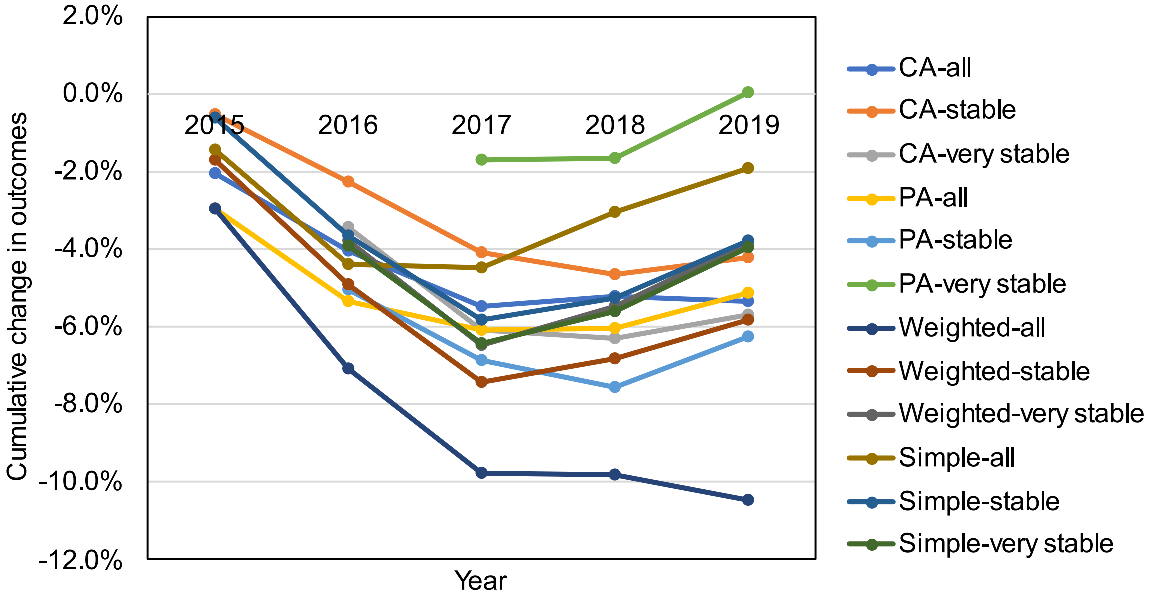

Figure 21. Cumulative difference between common centres predictions and actual outcomes over time by common centres method. Grade C/4.

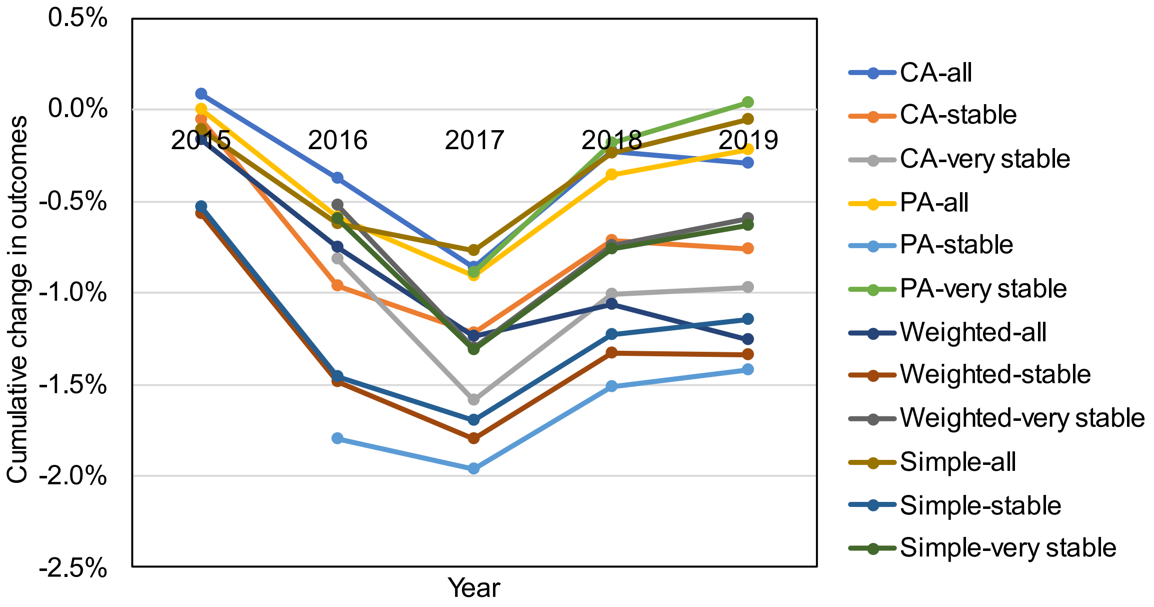

Figure 22. Cumulative difference between common centres predictions and actual outcomes over time by common centres method. Grade G/1.

Table 19. Summary of common centres analyses across methods, showing the mean and median percentage point difference between predictions and outcomes in each year and the cumulative effect, with 95% confidence intervals.

Grade	Method	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019	Cumulative (2014-2019)
A/7	**mean**	-0.5	0.4	-0.8	-1.0	0.4	-1.3
A/7	**CI**						**-2.5 to -0.1**
A/7	**median**	-0.8	0.4	-0.7	-1.0	0.4	-1.9
C/4	**mean**	-1.7	-3.2	-1.9	0.3	0.9	-4.7
C/4	**CI**						**-6.1 to -3.3**
C/4	**median**	-1.7	-3.2	-2.0	0.2	1.1	-4.7
G/1	**mean**	-0.2	-0.8	-0.5	0.5	0.1	-0.7
G/1	**CI**						**-1.0 to -0.4**
G/1	**median**	-0.1	-0.6	-0.4	0.5	0.1	-0.7

Although there is some variation across the different common centres methods, they present a similar picture. At grade A/7 outcomes were slightly lower than predicted by the common centres analyses in 2015, 2017 and 2018, however this was somewhat compensated for by outcomes being above those predicted in 2016 and 2019. If we total estimates across all years, then outcomes are lower in 2019 by somewhere between 0.1pp and 2.5pp than what we may have expected if centres outcomes had remained stable over the period studied.

At grade C/4 the average effect across our different methods suggests that outcomes were below predictions in 2015, 2016 and 2017, by around 1.7pp, 3.2pp and 2pp respectively. In 2018 and 2019 outcomes may have been slightly above predictions, although with some variance across methods. This results in a total difference of outcome being around 3.3pp to 6.1pp lower than what would be expected if centres outcomes had remained stable between 2014 and 2019.

At grade G/1 the effects are much smaller. Analysis suggests outcomes were again below prediction in 2016 and 2017, although this was mostly counterbalanced by outcomes being above prediction in 2018 and 2019. Overall, this suggests a slight negative effect of outcomes being around -0.7pp below predictions by 2019.

It is worth noting that the size and even the direction of these effects varied somewhat depending on the common centres method employed. The figures in Table 19 represent an average across methods, whereas individual methods suggest a larger or smaller effect. Estimates ranged from suggesting the cumulative difference in outcomes in 2019 were almost 11pp below predictions to only 1.9pp below prediction at C/4. For grade A/7 there was some variance in the direction of the effect, with estimates ranging from 4.2pp below prediction to 2.6pp above. Methods including all centres typically gave a more negative estimate than only including stable centres.

However, estimates across methods for the cumulative change between 2014 and 2019 were almost uniformly negative. Only the simple common centres method, for both all and stable centres at grade A/7, suggested a positive change in outcomes relative to common centres predictions, estimating outcomes 1.9pp and 2.6pp above predictions respectively.

It is also worth noting that even though we may expect outcomes to remain stable on average for centres over time, there will always be some fluctuation in outcomes. This is because multiple students will always receive the same mark in each year, so it may be impossible to exactly reproduce cumulative percentage outcomes, even if this were desirable.

Summary

Overall, these findings suggest there may have been a change in standards over time, particularly at grade C/4. Analysis showed a similar pattern to the previous modelling of a fall in outcomes at grade C/4 between 2015 and 2017, and at grade A/7 between 2016 and 2018. It seems unlikely that the same centres would have outcomes consistently worse in subsequent years of offering the qualification.

One possible reason would be if the centres included in the analysis entered on average lower performing students in subsequent years. However, the prior and concurrent attainment adjusted analyses should have compensated for changes in the general ability of the cohort, yet still generally showed a decline in outcomes. The number of students entering at the centres also did not consistently increase between pairs of years, which does not suggest centres were changing their entry policies and expanding their intake, which may have resulted in less able students taking computer science (see Table 20).

Table 20. Number of centres and change in number of students between each pair of years for common centres analysis.

Group	Value	2014-2015	2015-2016	2016-2017	2017-2018	2018-2019
All centres	Change in total entry	323	636	-911	-726	1269
All centres	N Centres	85	196	652	1278	1994
Stable centres	Change in total entry	-12	51	-113	-262	179
Stable centres	N Centres	40	108	353	666	1020
Very stable Centres	Change in total entry	-7	30	-35	1	-18
Very stable Centres	N Centres	10	41	147	225	374

Strand 1. Analysis 8. Comparative judgement of script quality

Aim

The previous strands of work all took a statistical approach to comparing standards, focussing on measures of outcomes over time. The aim of this strand of work was to take a different approach, instead focussing on the performance standard, that is the quality of work demonstrated by students to attain the key grades in each year. If the quality of work at the grade boundaries is different between years, this indicates that the performance standard of the qualification has changed.

Method

Overview

This strand of research utilised subject experts to judge the quality of students' work holistically and to compare the quality of students' work across different assessments over time. To facilitate this, judgements were collected from experts using a paired comparative judgement (CJ) task. Comparative judgement allows us to collect the consensus view of a group of expert judges, while minimising the potential bias introduced by individual judges' views. The method requires experts to make relative judgements about students' work, which is arguably psychologically easier, and more intuitive, than making absolute judgements of quality.

Within this study, judges were presented with pairs of examples of students' work, in the form of exam scripts from different years, and asked which script was higher quality. Multiple comparisons between different pairs of exam scripts based on experts' holistic view of the quality of students' work make it possible to construct a scale of 'perceived quality.' The location of each script on the scale of perceived quality depends on both the proportion of times it 'won' and 'lost' each paired comparison, but also the location of the scripts it was compared with (Bramley, 2007). If the distance on this scale between 2 scripts is greater, this means there is a larger probability that the higher scored script is judged as having greater perceived quality than the lower scored script (Bramley & Oates, 2011).

In this CJ exercise experts judged the quality of work in students' exam scripts at the grade A/7 or C/4 boundary for one exam paper which was broadly comparable pre and post reform. Given that the specifications changed, it was not possible to compare exactly the same exam over time. Therefore, exam papers which were the most similar in terms of content and structure were selected to facilitate comparisons. However, this means caution is needed when interpreting the findings pre and post reform as there were some changes to exam content and to the overall structure of the qualification. Non-exam assessments were not included as there was no comparator post-reform and due to the size of the assessment materials, they were deemed not suitable for inclusion in the CJ exercise. Details of the exams included are shown in the materials section below. The aim was to identify if the performance standard at the grade boundaries in the exam had changed over time.

Materials

For this exercise, assessments from the AOs with the 2 largest entries for GCSE computer science were considered (AQA and OCR). Pre-reform (2011-2017) each AOs' specification comprised one exam and either one or 2 controlled assessments. Post-reform, following the removal of the non-examination assessment, each AOs' specification comprised 2 exams. To allow comparison pre-and post-reform, only one of these 2 post-reform exams was considered. For both AOs one of the post-reform exams was similar in content and structure to the pre-reform exams, this exam was therefore used to make the most valid comparison. Details of the assessments are included in Table 21.

Table 21. Details of GCSE computer science assessments pre-reform (2012 to 2017) and post-reform (2018 and 2019). Assessments included in CJ exercise shown in grey.

OCR

Pre Reform		
Exam	Computer Systems and Programming	40% of the total GCSE, 1 hour 30 mins, 80 marks
Controlled Assessment	Practical Investigation	30% of the total GCSE, ~20 hours, 45 marks
Controlled Assessment	Programming Project	30% of the total GCSE, ~20 hours, 45 marks

Post Reform		
Exam	Computer Systems	50% of the total GCSE, 1 hour 30 mins, 80 marks
Exam	Computational thinking, algorithms and programming	50% of the total GCSE, 1 hour 30 mins, 80 marks

AQA

Pre Reform

Post Reform

Exam	Computing Fundamentals	40% of the total GCSE, 1 hour 30 mins, 84 marks
Controlled Assessment	Practical programming	60% of the total GCSE, ~50 hours, 126 marks

Exam	Written Assessment	50% of the total GCSE, 1 hour 30 mins, 80 marks
Exam	Computational thinking and problem solving	50% of the total GCSE, 1 hour 30 mins, 80 marks

The CJ exercise included student scripts on the grade boundaries from both AOs for each year that the assessments were available between 2011 and 2019. The OCR specification was first available in 2011 and the AQA specification was first available in 2014, resulting in 15 sets of student scripts. For each AO in each year, students' scripts were requested from AOs. Up to 5 students' scripts were requested at each of the A/7 and C/4 grade boundaries for each exam paper. For OCR only 3 scripts were available at each boundary in each year, and for AQA 5 scripts were available in most cases (4 in one case). Student scripts were requested that, as far as possible, showed a relatively even or typical performance across the paper.

Students' scripts were anonymised to remove information identifying the student, year and AO. All mark information was also removed from the scripts, and they were each given a unique ID. This ID could be matched to the blank question papers and mark schemes which were also provided to the judges (any information identifying the AO and year were also removed from these).

Judges

Sixteen judges were employed to complete the exercise, all of whom had experience of teaching GCSE computer science. Judges were initially recruited from Ofqual's list of subject matter specialists and additional judges were then recruited by contacting teachers directly. Judges were paid for their time.

Judging procedure

Judges initially attended an orientation meeting where they were informed of the aims of the study, given an introduction to comparative judgement and the software they would be using. Following the meeting they were sent detailed instructions and access to the judging platform and all additional materials which were stored in a secure online environment.

Following the meeting, judges were asked to familiarise themselves with the exam papers and mark schemes for all of the assessments included in the judging. They were then asked to provide a rating for how demanding they felt each of the individual exam papers was on a 7-point scale, from significantly less demanding than the average paper to significantly more demanding than the average paper. Their reference for this was how demanding they felt the papers were on average. Experts were told that a paper would be considered more demanding if a typical student would likely score proportionally fewer marks, or overall perform less well, than if they had taken another paper. The judges were asked to revisit these scores after they had completed the CJ exercise in case reviewing actual student responses to the papers had changed their opinion.

We know that exam papers differ in demand from year to year, as it is highly challenging to write exam papers which are of the exact same demand. This is usually compensated for by the setting of grade boundaries, as discussed in the introduction. Therefore, exams varying in demand was not a direct concern. Instead, the aims of the rating exercise were 3-fold. First, to initially orientate the judges to the exam papers and to ensure they had thoroughly familiarised themselves with the papers and mark schemes. Second, to attempt to avoid judges' views on the quality of students' responses being influenced by the demand of the assessments. Previous research has shown that judgements of the quality of students' work can be influenced by the demand of the assessment being judged (Good and Cresswell, 1988). Third, so we could evaluate the relationship between judges' perceptions of paper demand and student performance over time.

The judges were then asked to complete the CJ exercise. For this exercise they were given a unique login for an online judging platform where each judge was given a unique set of judgements to complete. For each judgement, judges were presented with 2 random scripts side by side and asked to consider "Which of these 2 students is the better computer scientist, based on a holistic judgement of script quality?". Judges were able to scroll up and down on each script individually before making their decision. Judges were asked to make their judgements based on the overall quality of the students' responses and not to attempt to re-mark the scripts to come to their decision. Judges were asked to make relatively rapid decisions and were informed that it should take around 5-6 minutes to judge each pair.

Initially each judge was given an allocation of 70 or 71 judgements, aiming for a total of 20 judgements per script across judges. Due to one judge not being able to complete the full task, their additional allocation was given to one of the other judges, resulting in one judge only completing 52 judgements and another completing 90 judgements.

Following completion of judging, a Bradley-Terry model (Bradley and Terry, 1952) was applied to the judgements to give each script a score indicating its likelihood to 'win' individual pairings. For this study, the script scores can be interpreted as indicating the quality of students' responses, relative to other scripts. For a detailed discussion of CJ methodology and analysis in this context see Curcin et al (2019).

Finally, after judges had completed all other tasks, they were sent a short survey asking how they had found the judging process, how confident they were in their judgements, and their general views about the quality of the students' work they had seen.

Results

Paper demand

Ratings of paper demand were first standardised within each judge (to a mean rating of 0 and standard deviation of 1) before being averaged across judges. The demand ratings are shown in Figure 23. On average, AQA papers were deemed to be more demanding than OCR papers. Comments from the surveys suggested that judges felt they were less accessible than OCR's papers. OCR papers were considered to be more demanding in 2015 and 2016, whereas AQA papers were considered to be most demanding in 2017. Post-reform (2018 and 2019) the demand of the papers between the 2 AOs was judged to be more similar. It is difficult to directly interpret the size of these perceived differences in demand as they were all on a relative scale. Discussions with experts indicated that they did think some assessments were more challenging than others (and this is not just an artefact of us asking the question), however it is unclear of how much impact this may have had on student performance.

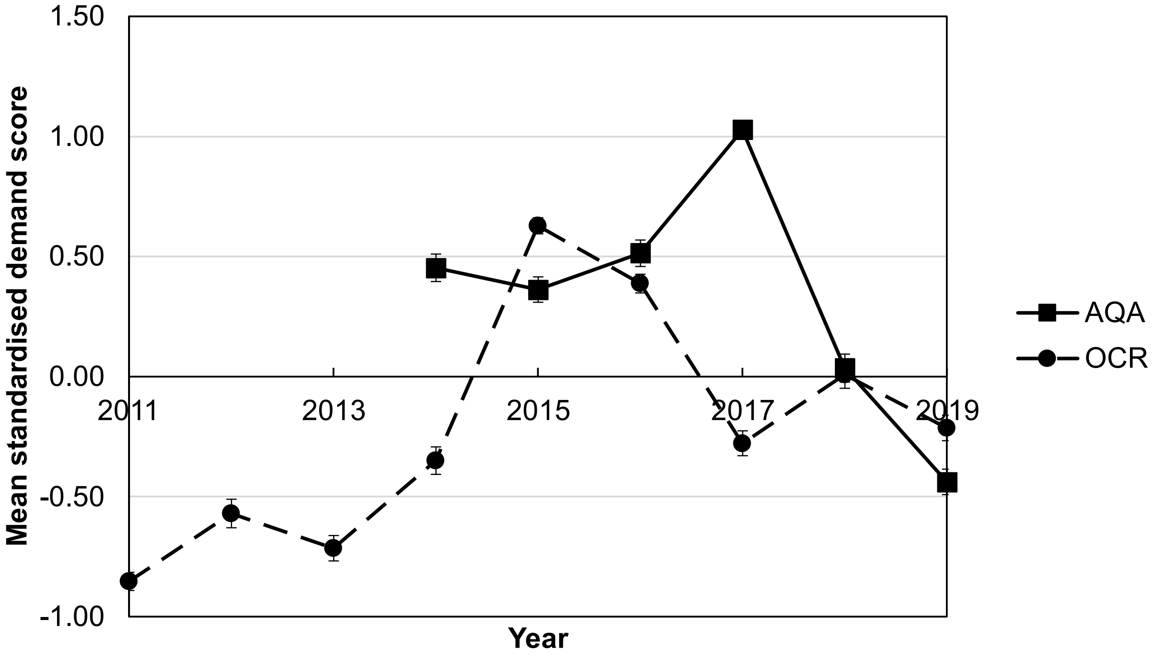

Figure 23. Mean standardised relative paper demand ratings by judges, with standard errors.

Figure 24 below shows the C/4 grade boundary positions for the same 2 assessments between 2012 and 2019. If all else remains stable, we would typically expect grade boundaries to change to compensate for a change in demand of the assessment. We might therefore expect the inverse pattern from Figure 23, that is, in years when the demand increases the grade boundaries should decrease to maintain the same standard within the assessment.

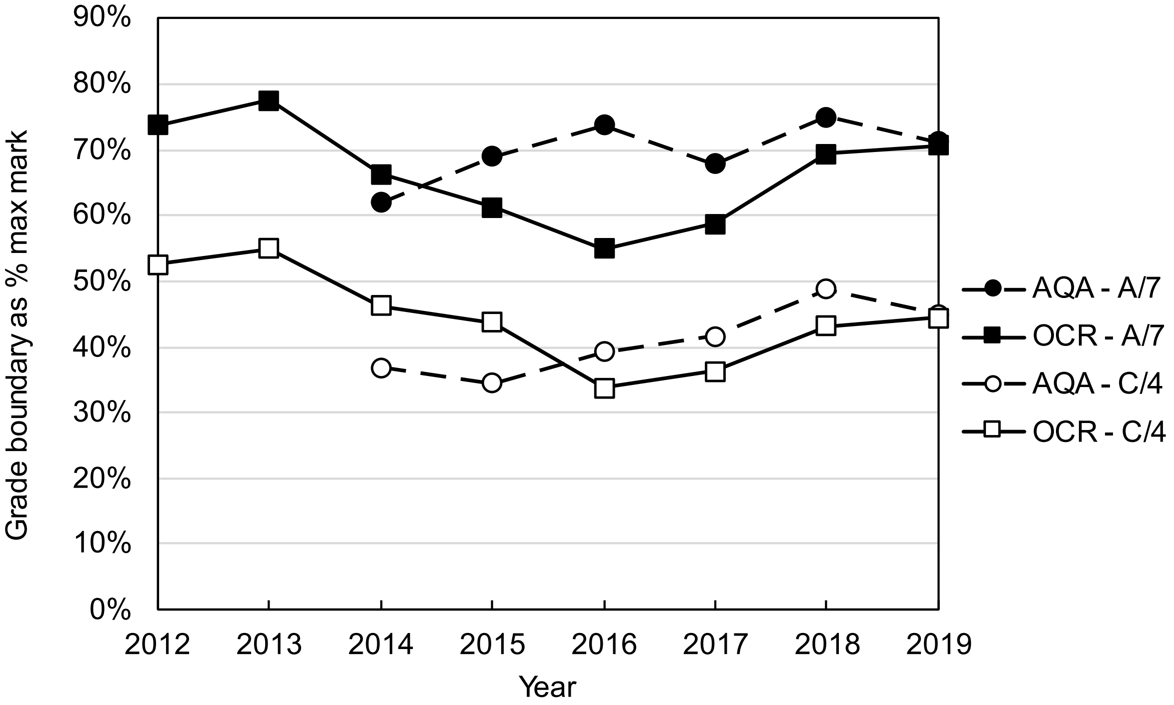

Figure 24. Grade boundaries for the assessments used in the CJ study over time.

Although there is some relationship between the patterns of changing assessment demand and grade boundaries, it is evident from Figure 23 and Figure 24 that the grade boundaries do not move solely as a response to a change in the assessment demand ratings. However, there may be other factors that affect grade boundary position beyond assessment demand. In particular, as standards are maintained at qualification rather than assessment level, we need to take into account the relationship between different assessments which make up the qualification when interpreting changes in grade boundaries.

The main purpose of this exercise was to familiarise the experts with the assessments and to take into account their demands when making their judgements for the CJ exercise. These ratings of paper demand also provide useful context to interpret the main CJ findings presented in the next section.

CJ analysis of script quality

Outputs from the CJ model allow us to evaluate the reliability of the ratings provided by the judges. In particular, infit is a measure of the consistency of the judgements made by judges, compared with the overall model fit. A high infit indicates that a judge was either inconsistent within their own judgements, or when compared with the judgements made by the other experts. Similarly, a script with a high infit may indicate that script was unreliably judged.

Judges took on average just over 7 minutes per judgement. One judge was removed from further analysis as their infit score was notably higher than other judges (1.44) suggesting that their judgements were not consistent with those of the other judges. Their median judging time was only 47 seconds which suggests that they may have not taken sufficient time to make accurate judgements. After removing this judge, the separation reliability was 0.85, which provides reassurance that judgements were consistent between and within judges. Scripts were judged on average 18.65 times (range 15-20). Four scripts were removed from the final presentation of results as they had a notably higher infit score than other scripts (over 1.5), suggesting they may have been particularly hard to judge and therefore their script quality scores may have been somewhat unreliable.

Figure 25 summarises the ratings across the different scripts for each year and each AO. Scripts with a higher score, and therefore further up the chart, are those rated as higher quality. Where all scripts in a year are rated as higher quality than other years this may suggest the performance standard needed to attain that grade was higher, which could be described as it being more difficult for students to receive that grade in that year.

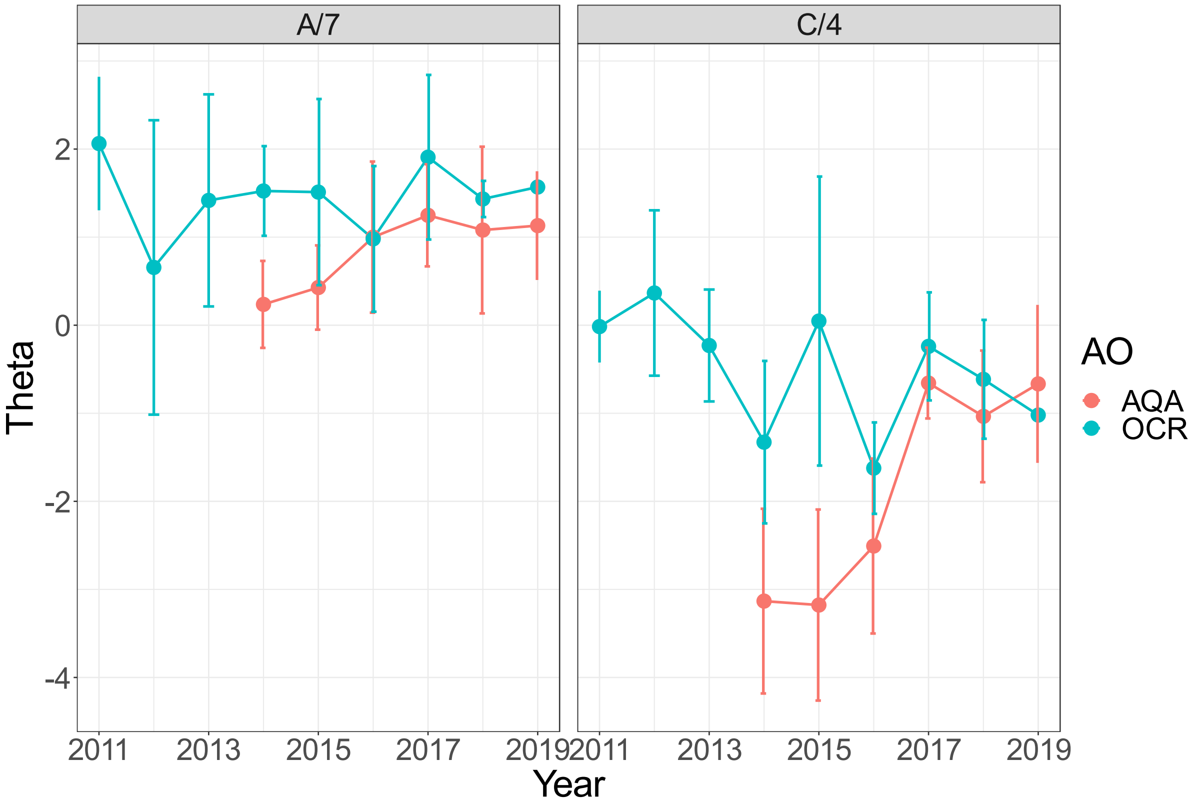

Figure 25. Line chart showing mean script scores at both A/7 and C/4 boundaries, with 95% confidence intervals.

A few patterns are notable from the CJ results, shown in Figure 25. At both the A/7 and C/4 boundaries during the period 2014 to 2015 the quality of work was judged as lower for AQA than for OCR. The quality of students' work for AQA also gradually improved between 2014 and 2017. It should be noted, however, that these temporary differences and changes over time do not appear to be problematic, as this is consistent with patterns expected from the 'sawtooth effect'. As centres offering the AQA specification were likely to be initially unfamiliar with the teaching material and the structure of the assessment when they were first available, we would expect gradual improvement as they become more familiar with the qualification and assessments. The grade boundaries indicated by the predictions during this period, which were initially based on the OCR outcomes, will have automatically compensated for this effect, suggesting a slightly lower quality of work at the boundary. This can also be seen in the increasing grade boundaries over this period in Figure 24. It is also important to note that the comparisons made between AOs here are based on a subset of assessments for each specification. The relationships between the assessments analysed should not be assumed to mirror those for the qualifications overall.

For OCR, the script quality ratings at the C/4 grade boundary from 2015 are noticeably higher and more variable than most other years. This paper was judged as one of the most demanding papers by experts (Figure 23) and the scripts for this paper were judged the most variable in quality. This may suggest caution in interpreting the results from this paper as experts may not have effectively compensated for the demand of the paper when comparing scripts from this paper with other papers.

The quality of work for OCR papers was lowest at the C/4 boundary for the years 2014 and 2016. Comparing these 2 years with 2017, this would represent a shift to a higher quality of script being required in 2017, which could indicate that it became more difficult to achieve a C/4 in this year. This pattern is similar across both OCR and AQA papers at grade C/4 and is somewhat evident at grade A/7 for OCR. However, this finding should be treated with caution for 2 reasons. First, for AQA, the grade boundaries may still have reflected sawtooth effects being present across this period indicating a change in required quality for reasons other than an unintentional change in standard. Second, this finding ignores the 2015 data point for OCR at grade C/4, which suggests a higher quality of work was required in that year.

Summary

Overall, the results of the comparative judgement exercise are somewhat inconclusive. The quality of work required to achieve a C/4 may have dropped in 2014 (for OCR) before rising again in 2016 and 2017 (for both AQA and OCR). The drop in quality in 2014 may have come from predictions being used to aid setting boundaries, when the students may have been less well prepared as new and less specialist centres started offering the qualification. The suggestion of a rise in quality following this would agree with previous analysis which indicate an increase in difficulty in 2016 and 2017, although the findings are not clear cut.

Strand 1. Discussion

To aid discussion of findings, Table 22 summarises all of the above analyses and what they indicate the size of the potential change in standards between 2014 and 2019 might be. These changes are expressed in terms of the estimated percentage point change in students receiving at least an A/7, C/4 or G/1 grade and in terms of mean grade, where appropriate. As discussed previously, each of these analyses has different assumptions and limitations and, in some cases, include slightly different samples of students. Therefore, there needs to be caution in directly comparing the results from the different analyses.

Table 22. Summary of findings from analyses in strand 1.

Method	A/7 change from 2014	C/4 change from 2014	G/1 change from 2014	Mean grade change from 2014 (9 to 1 scale)
Raw change in outcomes	-1.3pp	-4.3pp	+0.02pp	-0.26
Cumulative effect of awarding under prediction	-1.5pp	-3.5pp	-1.8pp	
Rasch analysis				-0.12 to -0.28
Kelly's method				-0.17 to -0.27
Progression analysis				-0.15
Simulated predictions (excl new) - prior	-4.2pp	-6.6pp	-0.5pp	
Simulated predictions (excl new) – concurrent	-6.1pp	-7.3pp	-0.5pp	
Model of outcomes over time (2014/2015 centres only) - concurrent attainment	-4.5pp / -3.4pp	-4.4pp/-5.6pp	No change	-0.31/-0.33
Model of outcomes over time (2014/2015 centres only) - prior attainment	-4.3pp/-2.0pp	-3.9pp/-4.3pp	No change	-0.27/-0.28
Common Centres (95% CI of mean of different models)	-2.5pp to -0.1pp	-6.1pp to -3.3pp	-1.0pp to -0.4pp	
Comparative Judgement study	No difference	More difficult?		

Overall, based on the range of analyses conducted in strand 1, between 2014/2015 and 2017 there appears to have been a subtle shift in standards in GCSE computer science. This is particularly noticeable around the C/4 grade boundary, but there also appears to be a slightly smaller change at the A/7 grade boundary. The various modelling carried out estimates that by 2019 similar students may have been around 3 to 6pp less likely to attain at least a grade C/4 when compared with students in 2014. At the A/7 boundary the evidence is somewhat less consistent, but there is some evidence that a similar shift may have happened resulting in students being around 2 to 3pp less likely to attain an A/7 in 2019 compared with 2014.

This change in standards was identified across the analysis methods employed, although the exact size of this effect varies between the different analyses. In the introduction we discussed 3 possible factors that could cause a drop in outcomes; 1. Changes in the ability of the cohort, 2. Sawtooth-like effects where centres were offering the subject for the first time and may have been unfamiliar with the qualification or assessments, 3. Changes in other factors over time which may affect students' preparation for the assessments, such as teaching quality. However, even once we had controlled for the possible impact of all 3 factors, particularly in the common centres analysis and modelling approaches, the change in outcomes was still evident. This suggests that there has been an unintended change in the grading standards in the qualification over time. These changes are likely to have been small within each year and may not have been observable to awarders within each year but resulted in a larger cumulative effect over time.

The reasons for this shift in standards we can only surmise from the data available and from awarding documents acquired from AOs. During the period 2015 to 2017, the number of students taking the qualification more than doubled, many of whom were being taught at centres that had never offered GCSE computer science before. The prior attainment of these students from these centres was, on average, lower than those in previous years. Prior-attainment-based predictions subsequently fell during this period, however, despite this fall in predictions, across AOs grade boundaries were set such that outcomes were below predictions. All of this combined suggests that students may, on average, have performed less well in the assessments over time, potentially leading to a valid decrease in outcomes.

Prior to the first year of reformed assessments in 2018, the qualification included a controlled assessment element, consisting of a project carried out in class. Due to the fact the controlled assessment tasks typically stayed similar from year to year, the grade boundaries were typically kept consistent from year to year. This meant that AOs were most likely to take into account the evidence provided by the statistical predictions through the setting of the grade boundary on the examined component. On average, grade boundaries for the examined components were lowered between 2014 and 2017, but it was judged not appropriate to lower them such that student outcomes met predictions. Awarding reports suggest that awarders did not feel comfortable lowing the grade boundaries any further to meet predictions as they believed the quality of the work was not of a sufficient standard. This may have led to a disparity between performance on the different assessments, which would have made maintaining appropriate standards at qualification level highly challenging.

Alongside these changes, AOs were also dealing with issues of malpractice. Malpractice in the controlled assessment elements was an issue since the inception of the qualification, ultimately leading to the removal of NEA post reform in 2018. In an attempt to counteract this, OCR (the largest provider of the qualification) made one of their controlled assessments more open ended, and therefore potentially more challenging in 2016. Changing assessments can lead to temporary sawtooth-like effects as teachers become familiar with the new assessment structure, which may result in lower student performance.

The above suggests that AOs not meeting predictions in 2016 and the subsequent dropping outcomes in 2016 and 2017 may have been somewhat justified by the weaker performance of the candidature, in part due to the changing composition of the cohort. However, some of the effects that led to this weaker performance may well have been temporary.

In computer science the reference year for predictions was updated almost every year between 2012 and 2019. That is, the year used to benchmark the relationship between prior attainment and outcomes. The intention of this was to reflect the changing cohort and any resulting changes to the value-added relationship. While this is likely to have aided the management of changes in the cohort during this period to a considerable degree, the unintended effect may have been to also carry forward any changes to the value-added relationship from years when performance may have been temporarily weaker. 2018 and 2019 were also the first 2 years of the reformed qualification, a period where there are additional challenges to ensure that standards are maintained. This change may have made it impossible to identify any positive changes in performance standard, which may have followed if some of the effects leading to lower performance were temporary.

Ultimately this means that between 2014/2015 and 2017 there was a relatively rapid (albeit small) reduction in the value-added relationship between students' prior attainment at KS2 and their performance in GCSE computer science. Then between 2017 and 2019 this relationship stayed relatively stable. This could be a valid reflection of the changing cohort and therefore a true representation of their ability in computer science. However, the number of transitory effects seen during this period, and the fact that from our analysis we also saw a fall in outcomes in centres whose outcomes we would expect to have been stable during this period, causes us to question the validity of this reduction in reflecting a genuine, permanent, change in student attainment.

In 2023, GCSE computer science had been widely available to teach for over 10 years. Therefore, it is reasonable to believe that any of the temporarily transitional effects that may have impacted on students' performance in the early years of the qualification should have passed.

Caveats and limitations

One key limitation to almost all of the analyses presented here is the assumption that there have not been legitimate changes in student outcomes in GCSE computer science assessments, due to factors that have not been controlled for. These legitimate changes could come from a variety of sources. For example, for the analyses that compare outcomes in GCSE computer science to prior or concurrent attainment, a key assumption is that the relationship with prior attainment should have remained stable. Value-added relationships can vary for a wide variety of reasons which may be legitimate and which may be challenging to take account of during standard setting. For example, changes in teaching time, teaching quality, student motivation or changes to content taught could all cause legitimate changes in outcomes. This also applied to methods comparing outcomes in computer science to other subjects.

Another potential legitimate change in outcomes may have come from the reduction in malpractice following the removal of NEA. Prior to reform there were substantial issues with malpractice which may have led to inflated outcomes on this assessment. In 2016, OCR changed its controlled assessment structure to attempt to limit this, which may have led to a small drop in outcomes. Removal of the NEA component following reform could also cause challenges in maintaining a performance standard, although this was aimed to be compensated for through the approach to maintaining standards during reform. Hypothetically a drop in outcomes could represent a more valid reflection of students' attainment, than that represented by the pre-reform controlled assessments.

Along similar lines there is anecdotal evidence of students being better prepared for controlled assessment tasks than they were for exams when both contributed to the qualification grade. Once this option was removed this could have led to a fall in outcomes if teachers were not well prepared to teach exam content. However, this effect could have been temporary along with other effects during reform.

One additional assumption we have made throughout these analyses is that the standard set in the initial years of the qualification was an appropriate one, and subsequently that 2014 was an appropriate year to benchmark standards to. Setting standards in the first years of a qualification is a challenging task and ultimately the appropriateness of the standard in a subject can only be determined by experts and stakeholders within that field. This is something we return to in strand 2 of this work.

Strand 2 - Performance standard in summer 2023

Aims

The previous analyses have considered whether there is evidence of a potential change in standards over time. The aim of this strand was to examine performance standards in GCSE computer science assessments in the most recent series they were available (summer 2023) to understand the impact any change in standards would have on the quality of work needed to be demonstrated by students to receive the key grades considered during awarding (grade 7 and grade 4). This study was therefore focussed on the minimum level of performance required for these grades.

There were 2 elements to this. The first was to understand at which point on the mark scale a difference in standards from the grade boundary was consistently identified by experts and, where it was noticeable, whether experts believed the quality of work was acceptable to receive the relevant grade. The aim of the second element was to understand where in the range of student performance experts felt the quality of work indicated students would succeed in further study in computer science. Here we rely on one of the key aims of GCSEs "*to provide a strong foundation for further academic and vocational study and for employment*" (Ofqual, 2023) as a benchmark for the qualification standard.

This work required the subject experts to make holistic judgements about the quality of students' work at various points in the mark distribution, and to identify where they could reliably perceive differences in the quality of work. This is a highly challenging task and is particularly difficult where students' responses are uneven across the assessment, or when judges need to keep in mind a large amount of evidence to make their judgements (Leech and Vitello, 2023). However, given that expert judgement is a key component of setting standards for GCSEs, identifying where experts can identify differences in the quality of work and the magnitude of those differences is important to understanding the impact of any changes to that standard.

Finally, we also wanted to receive any other qualitative insights that the expert group of computer science specialists might have about the standard of the current GCSE computer science qualifications.

Methodology

Recruiting subject experts

Computer science subject experts were recruited to carry out the review exercise. We recruited experts from a range of backgrounds, all of whom had some familiarity with the current A level or GCSE qualifications. Experts were recruited via a number of sources including Ofqual's register of subject matter experts, recommendations by BCS - the chartered institute for IT, and from contacting an AO to recommend senior examiners to take part in the work. The intention was to recruit a panel of computer science experts with varied backgrounds, that represent a range of stakeholders in the qualification, to provide detailed insights into the standard of GCSE computer science. We successfully recruited 8 experts with a wide range of experience including current and previous A level and GCSE computer science teachers, representatives of BCS and Computing At Schools (CAS), those with marking experience for different AOs, those with experience of being a senior examiner and awarding (grade boundary setting), those with experience of training other computer science teachers, and those with experience of writing textbooks and other materials to aid in teaching computer science (see Table 23 for summary).

Table 23. Summary of subject experts' background and computer science (CS) experience

Experience	**Number of experts (8 total)**
Number of years teaching CS or related qualifications	Median 17.5, min 14, max 36
Experience of teaching GCSE CS	8
Experience of teaching A level CS	7
Worked as an examiner for CS (any AO)	5
Experience writing/developing CS assessments	3
Experience writing CS training materials or textbooks	7
Experience training other CS teachers	6
Degree or higher in CS (or closely related subject)	6
Worked in CS outside of teaching	2

Exam materials

Exam scripts were requested from AQA, the AO with the second largest entry into GCSE computer science. The experts were generally less familiar with the AQA specification and so were likely to have less preconceived ideas about script quality or assessment demand. Our previous analyses suggested that the standard between AOs is highly similar, and no substantial concerns have been raised about inter-AO comparability, or the lack thereof. Therefore, we believed it was an appropriate assumption that conclusions from one AO about the performance standard could be applicable to all AOs offering GCSE computer science.

Student work was requested across a range of marks, based on the total qualification mark achieved (more details about this are included below). A number of examples of student work were requested at each mark point – 5 on the grade boundaries and 3 on other marks. AQA’s specification includes 2 exam papers, Paper 1: Computational thinking and programming skills and Paper 2: Computing concepts. Paper 1 is available in 3 versions (1A, 1B, 1C) depending on the programming language used (C#, Python or VB.Net respectively). To aid the experts in making comparisons between scripts we only included students who had taken Paper 1B (Python) as it has by far the largest entry. Both exam papers from the same student were requested, and scripts were anonymised to remove any student identifiers and all mark information. Scripts were requested that had a relatively even profile across both exam papers. Both exam scripts from the same student were combined into a single PDF. ‘Packs’ of scripts of students with the same mark total (at qualification level) were then created.

Method

Subject experts attended an orientation session where the researchers introduced the aims of the project, explained the tasks and allowed the experts to ask questions and seek any clarification. Experts were then asked to complete 2 tasks at home, in their own time. Finally, there was a review meeting to discuss the results of the tasks and for the experts to provide any additional insights.

Task 1

For task 1, experts were provided with packs of scripts at the grade 7 and grade 4 qualification level grade boundary, along with mark schemes and specification documents. These were borderline students who received just enough marks for each grade. Experts were asked to review the scripts in these packs and provide a summary of the strengths and weaknesses demonstrated by students, and what skills or knowledge they displayed (or did not display). The experts were then asked to indicate if they thought the quality of work was at the level they expected for a GCSE grade 7 or 4.

Following this, the experts were presented with a series of packs of scripts at various mark points above and below the grade 4 and grade 7 boundary. Each pack was given a randomly assigned ID and the mark totals on the scripts were removed, so the experts were not aware which pack was which. There were 3 packs above each grade boundary at every other mark from +2 marks above the boundary to +6 marks, and there were 7 packs below each boundary from -2 marks to -14 marks below the boundary.

For each pack the experts were asked if they thought that the overall quality of work (across all students in the pack) was typically much better, slightly better, slightly worse, much worse or not noticeably different to the work at the grade boundary. Where the expert thought there was a difference, they were asked to provide a short summary of what these differences were, that is, were students typically better or worse at demonstrating particular skills or knowledge. Experts were asked as far as possible to form a holistic judgement across the students included in each pack, to get a sense of what was 'typical' at each mark point. We were aware that this could be challenging in some cases. Despite the scripts having a relatively even mark profile, different students may have had very different performance profiles across the exams.

Task 2

After completing task 1 experts were asked to complete task 2. For task 2, experts were asked to think about a GCSE level student who they believe showed enough aptitude to go on to further study in computer science and be successful. They were then asked to describe what skills or knowledge they would expect this student to display. Experts were not initially told what 'success' should consist of, but this was discussed with the experts following the task (see results – task 2).

Experts were then presented with a different series of packs of scripts, with the mark information removed. However, this time they were numbered and presented in order of descending marks starting with the grade 7 boundary scripts. Packs were provided in 5-mark intervals working down the mark range until the grade 3 boundary. For this task experts were aware that the packs were ordered by mark total, although they were not aware of the exact mark or grade of each pack or the difference in marks between packs.

For each pack, experts were asked to indicate if they believed the students within that pack were highly likely to succeed in further study, somewhat likely, somewhat unlikely or highly unlikely to succeed. Finally, experts were asked to provide a short rationale for their decision and describe the skills or knowledge they had seen at different parts of the mark range that had informed their decision.

Review meeting

Following completion of the tasks, 2 review meetings were conducted each with 4 of the subject experts. At the meetings, experts were presented with a summary of the findings from the first 2 tasks and were asked for additional insights and reflections. A key part of this was discussing which of the packs presented in task 1 they felt demonstrated enough knowledge and skills to receive a grade 4 or grade 7 where the quality of work was noticeably different from the grade boundary. Experts were also invited to share their views on the overall standard of the qualification and any reflections on the perceived difficulty of GCSE computer science.

Results

Task 1

After reviewing the grade boundary scripts for each grade but before reviewing the packs for task one, the experts were asked whether the quality of work at the qualification level grade boundaries was as they expected for that grade, in a free text response. A summary of experts' responses is shown in Table 24. Experts were not given any guidance about what they should refer to when considering their expectations, as we were interested in their diverse views depending on their background and experience. When questioned, experts stated that they variously drew on their experience of teaching A level, teaching GCSE, awarding the subject and professional experience.

Table 24. Summary of experts' responses when asked if the quality of work at the grade boundary was as they expected.

Comparison of quality to expectations	**Number of responses – Grade 4**	**Number of responses – Grade 7**
Better than expected	1	3
Slightly better than expected	1	1
As expected	4	3
Slightly worse than expected	2	1
Worse than expected	0	0

At grade 4 there was a fairly even split between experts believing that the quality of work was higher or lower than expected, with only 1 expert expressing a strong view that the work was higher quality than they would expect from a borderline grade 4 student. However, at grade 7, 4 of the experts thought the work was better than they expected, whereas only 1 thought it was worse than expected for a borderline grade 7 student.

The main findings of task 1 are shown in Figure 26 and Figure 27 below. These figures show the percentage of the experts who rated each pack of scripts as being much better, slightly better, slightly worse, much worse or not noticeably different from the quality of work at the grade boundary.

There was a fair amount of variation in subject expert responses to each pack, both in terms of there being differences between experts but also in relation to the mark totals. The discussions with the subject experts indicated that this may have been due to different experts prioritising different skills or parts of the exam papers when making their judgements. This may, in part, reflect the diverse nature of content in the subject. Experts also said that they found this task challenging as there was often a large amount of variation in the skills and knowledge displayed by students within each pack who received the same overall mark. This made it challenging to make a holistic judgement about the quality of student work at each mark point. These results are not dissimilar to previous work which showed that it can be challenging for examiners to consistently identify differences in the quality of students’ work when the difference in total marks is small (Baird and Dhillon, 2005).

This may also have led to differences between expert ratings and what was credited in the mark schemes, as subject experts may have put more weight on some areas of skills and knowledge than others. For example, the experts noted that they typically prioritised performance on the programming and/or extended response questions when making their judgements about the quality of student work. For some scripts, experts highlighted that students could be inconsistent, for example showing high quality responses in some areas but not responding to all questions, which led to lower mark totals.

Grade 7

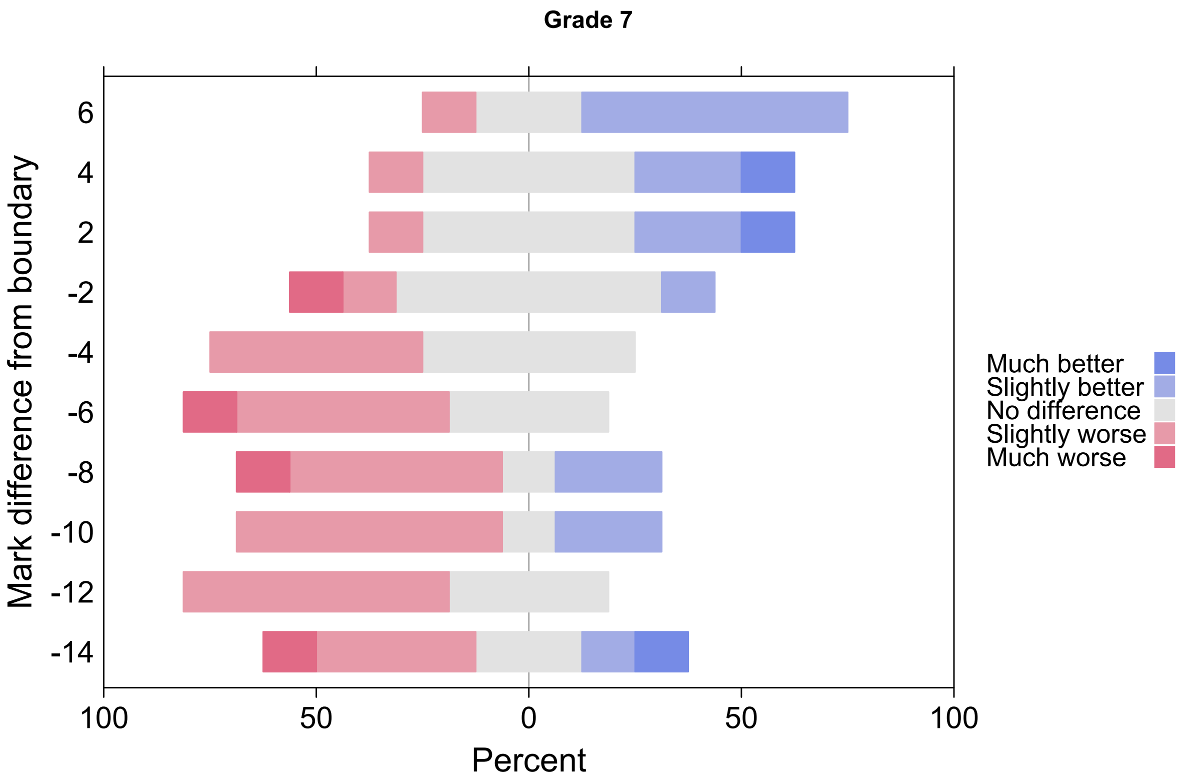

Figure 26. Subject expert ratings of packs of scripts at different mark points around the grade 7 boundary.

The results from the review of students' scripts around the grade 7 boundary indicate that within the range of approximately +4 to -2 marks from the grade boundary, the experts did not consistently identify any difference in the quality of students' work. Within this range, less than 50% of experts indicated that the packs were noticeably different from the grade boundary scripts in the appropriate direction.

At the review meeting experts were asked to review the scripts below this range, and to provide further views on the quality of work that students demonstrated at these mark points. At -4 marks from the boundary, although experts thought the work was weaker than the work at the grade boundary, the majority of the experts still believed the scripts showed enough knowledge and skills to receive a grade 7 without it having a strong impact on the performance standard indicated by that grade. At -6 marks from the boundary there was some disagreement. While a few of the experts believed that students' work at this mark showed high enough quality to receive a grade 7, others disagreed. Below this point, however, the majority of experts believed that there were too many weaknesses in students' work, with students showing a lack of understanding and having too many gaps in their knowledge for a grade 7.

Grade 4

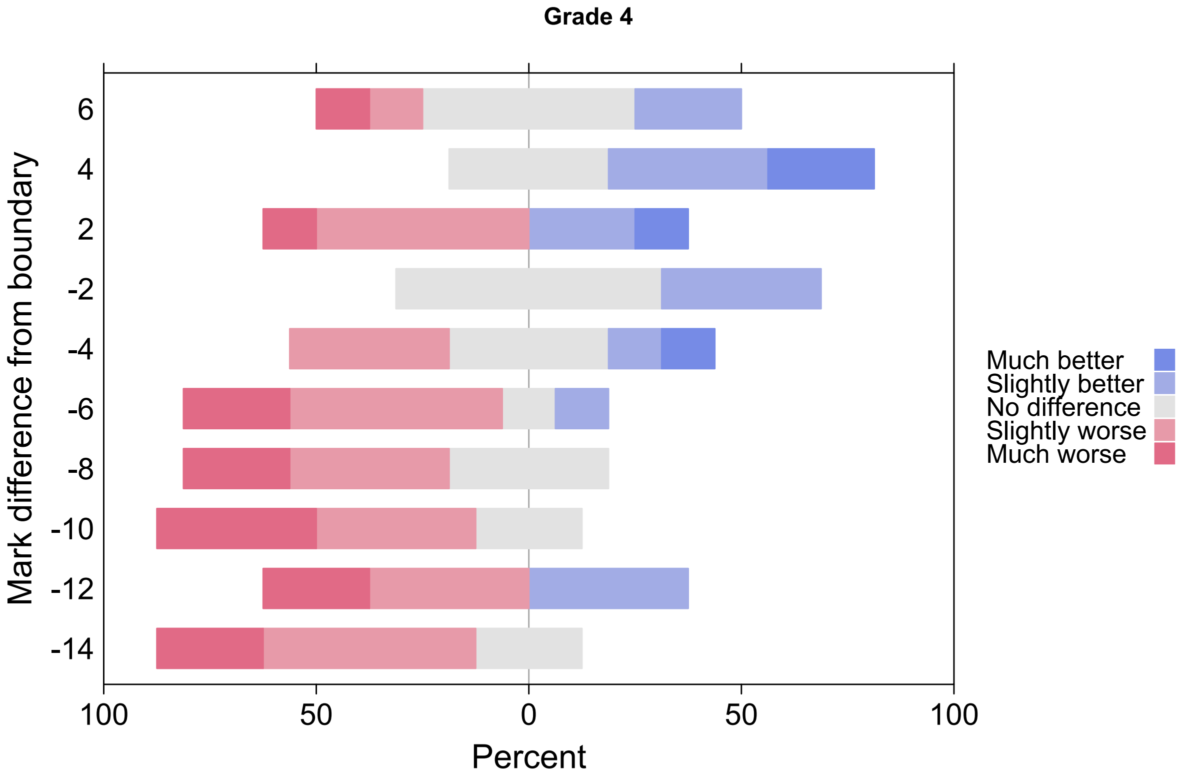

Figure 27. Subject expert ratings of packs of scripts at different mark points around the grade 4 boundary.

At grade 4 there was also a zone around the grade boundary where the experts did not consistently identify differences in the quality of scripts from the grade boundary. This ranged from +2 marks above the boundary to -4 marks below.

Discussions with the experts after reviewing scripts below this zone again indicated that students at -4 marks may show enough skills and knowledge for a grade 4, with experts agreeing that overall, the quality of work met their expectations. At -6 marks there were mixed views from the subject experts, with some believing the scripts indicated students showed enough aptitude for a grade 4, however others believed that there were too many missed answers and gaps in knowledge. Below this point, all of the experts believed that students showed significant weaknesses, with misunderstandings in key concepts and notably weaker programming skills.

Based on the outcomes from students who sat these assessments in summer 2023, we can convert the mark differences from the grade boundaries into a percentage point change in students who would receive the grade if the grade boundary were moved to different mark points. For this calculation we only included 16-year-old students. This information is summarised in Table 25 for the different mark points discussed above. We present this alongside a simplified summary of the experts' comments on the difference in the quality of work.

Table 25. Summary of differences in performance standard at different mark points and the difference in percentage points (pp) of students at each mark point

Grade	Mark Difference	Difference from boundary performance standard	Change in % of students attaining grade
7	-2	Not noticeable	+1.9pp
7	-4	Minor	+3.6pp
7	-6	Moderate	+5.4pp
7	-8	Significant	+7.1pp
4	-2	Not noticeable	+1.2pp
4	-4	Not noticeable	+2.5pp
4	-6	Minor/Moderate	+3.7pp
4	-8	Significant	+4.9pp

Task 2

The purpose of task 2 was to understand the subject experts' view of what a successful student in further study in computer science would know and could do. During discussions with the experts at the review meeting, the majority of the experts confirmed they had considered a student continuing down a traditional academic route in computer science when completing this task. Most experts focussed on students likely to achieve at least a grade C in A level computer science. However, the experts emphasised that for some students a grade E could still be considered a success. A minority of our experts also considered other routes such as a T Level in digital skills.

The experts were asked to provide a summary of what skills and knowledge they would expect a student who would go on to be successful at computer science to show at GCSE level. Experts provided a broad list of skills (summarised in Table 26), which may reflect their varied experience, expertise and perceived priorities within the subject. However, in further discussion during the review meetings, experts agreed that they would not expect a single student to demonstrate all of these skills. They suggested that many of these skills could be summarised as good programming and problem-solving skills. Experts also made it clear that it is often not possible to define a list of skills or knowledge that indicate success and, beyond the content of the qualifications, it is often less tangible factors such as motivation, maturity or willingness to learn which are predictors of a successful student.

Table 26. Summary of skills and knowledge identified by subject experts that might indicate a student who would be likely to succeed in further study in computer science.

Ability to read and debug code
Good communication skills and use of technical terms
Good understanding of theoretical concepts, although not necessarily their application
Clear understanding of data types and data representation
Basic understanding of computer systems and hardware
Ability to discuss legal and ethical issues relating to technology
Basic/strong mathematical skills
Understand the basics of networking and communication between devices
Passionate about the subject and able to see beyond the curriculum
Ability to think logically
Able to interpret and apply algorithms to solve problems (with reasonable efficiency)
Able to apply the principles of computational thinking
Reasonable/strong programming skills
Confidence with simple data structures (for example, arrays)
Proficiency in one programming language
Ability to think creatively
Strong ability in various number systems (base 8 and 16)
Able to write programs to solve non-trivial problems
Good ability at problem solving
Understand the underlying abstraction in computer systems

During the discussions, the experts noted that they also found task 2 challenging, particularly because of the varied profile of students within each pack. However, they found task 2 easier than task 1, as the packs were presented in mark order and therefore, they had higher confidence in their judgements.

Figure 28 below shows the ratings given by experts to packs of scripts at different points in the mark scale. As in the previous figures the coloured bars indicate the percentage of experts who gave each rating; from students highly likely to succeed in further study to those highly unlikely to succeed in further study.

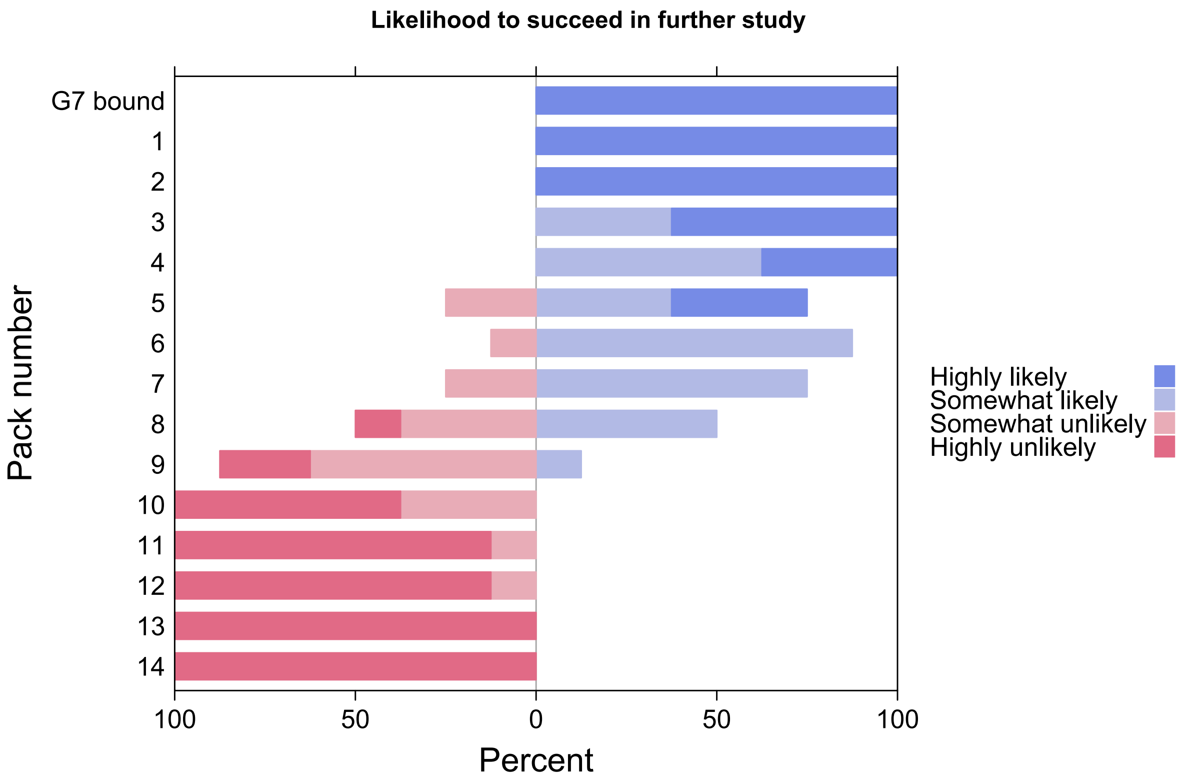

Figure 28. Subject expert ratings of the likelihood of students in each pack to succeed in further study in computer science.

The key points of Figure 28 are pack 7, the lowest pack where the majority of experts believe that students would be likely to succeed in further study, pack 4 where all experts believe students would be likely to succeed, and pack 3 where the majority of experts believed students would be highly likely to succeed in further study. In terms of grades, pack 7 represents students achieving a low grade 5, pack 4 a high grade 5 and pack 3 a low grade 6.

It is worth reiterating that for this task the packs were presented in order from highest mark total to lowest and the experts were aware of this ordering. This is likely the cause of the greater consistency of ratings across the mark range than in task 1.

The results of this task were shared with the experts at the review meeting. Some of our experts were surprised at how low in the grade distribution students had been rated as likely to succeed in further study, stating that they expected these students to have received a higher grade and that students wouldn't typically be admitted onto

A level courses with a grade lower than a 6. Some experts attributed this to giving the benefit of the doubt to students while completing the tasks, with experts trying to find evidence in the scripts which might indicate quality, particularly where students may have shown aptitude but lost marks by articulating themselves poorly in the exam. Other experts were less surprised, particularly in light of the previous discussion that it is often not the subject specific skills and knowledge displayed in the exam which indicate a promising student, but other factors. These other factors which experts may have seen in students' answers may often not have been 'credit worthy' in the mark scheme resulting in students receiving lower grades.

Broader views on the standard from subject experts

As part of the discussions the experts were encouraged to provide broader views about the qualification and particularly any views they had about the current qualification standard. Although not a unanimous view, some of the experts believed that the GCSE was too challenging; a view which was explored further through discussion. It was expressed that this view was based on a variety of reasons and not simply due to grading standards. In this section we discuss the main points made by the subject experts about the current qualification and factors which may impact on the actual or perceived difficulty for students, these comments are summarised into themes below.

Teaching Quality

The experts noted on a number of occasions that they thought that the majority of centres offering computer science do not have a specialist computer science teacher and that was reflected in the quality of students' work. It was believed by a number of experts that this was the primary driver of students receiving relatively lower grades in computer science to other subjects and the perception of the subject being difficult. However, it was also suggested that this is a difficult problem to overcome as there are likely to be opportunities for employment in other sectors for good computer scientists that might be more financially rewarding. This concern has been raised elsewhere (Royal Society, 2019), with reports indicating that computer science teachers can earn more money in careers outside of teaching (Sibieta, 2018). Previous statistics have also suggested that historically only around 15% of computer science teachers were subject specialists (Dallaway, 2016). In 2017 46% of computer science teachers in secondary schools held a relevant computing qualification (36% computer science, 10% ICT or business with ICT) (Royal Society, 2017). More recent data from the academic year 2022/2023 indicate that just over half of the hours taught in computing in secondary schools were taught by teachers with a relevant post-A level qualification (54.1%), which is in contrast to other Ebacc science subjects where the majority of hours were taught by a subject specialist (73-95%) (DfE, 2023).

In some cases, the experts felt like they could see evidence of good students answering questions badly, which could be due to students being poorly prepared for the exam resulting in poor exam technique. They also felt that there was some evidence of certain areas of the content being prioritised over others. It was also noted that computer science is a very practical subject, that can be difficult to teach in a classroom setting – especially if schools do not have the right equipment available.

Content and teaching time

There was discussion throughout both of the review meetings about the variety of content included in the GCSE computer science curriculum. Experts variously showed disagreement about what skills or knowledge should be prioritised, as expressed through their ratings in task 1. It was noted that this was a broader concern within the subject, as there were varying views from those in the field about which skills were important. Experts speculated that this may have resulted in the GCSE content being too broad, leading to difficulty.

Some experts noted that when the content had been originally designed it was expected to sit alongside the GCSE in ICT which has since been discontinued, and perhaps were it to be redesigned it might be advantageous to include some ICT content alongside the computer science content. There was a belief that this may make the subject more accessible.

Experts also noted that due to the broad content they did not think that there was sufficient time to adequately teach it all. This particularly related to the programming elements, which experts believed take much longer to teach than reflected by the weighting they are given in the curriculum and assessments. These concerns have been raised elsewhere (Royal Society, 2017; Ofsted, 2022). There also appears to be a trend of less, rather than more time dedicated to teaching computer science over time (Kemp & Berry, 2019; Royal Society, 2019). The experts suggested that students need to be sufficiently engaged to practice programming skills outside of the classroom if they are to succeed.

Exam structure and mark schemes

One comment that recurred in our discussions with the subject experts was that in some cases there was disparity between the marks students received and their judgements of ‘quality’. Experts in a number of cases identified students who they believed showed some skill or understanding but missed out on marks. There was some supposition that this could be due to the material being taught poorly, or that students having poor exam technique and so missed out on ‘easy’ marks by articulating themselves badly. On the other hand, some experts believed that due to the nature of the exams, students who did not have a good grasp of the subject, could still gain a reasonable number of marks across the paper.

It was commented on a number of occasions that testing coding skills in a written exam may not give a valid indication of a student’s ability. Experts expressed a preference for computer science assessments taken on-screen where students can edit or even trial their code, although experts acknowledged that there were good reasons why the NEA had been removed.

Progression

There was some suggestion from the experts that attainment at GCSE did not necessarily indicate how well students would do at A level. This may be due to many students taking the A level in computer science not having taken the GCSE, therefore A level teachers expect students to have gaps in their knowledge. Experts believed success in further study was more to do with effort, attitude and not being afraid to have a go and make mistakes, than subject knowledge. Related to this, one expert believed that students who received as high as a grade 7 in the GCSE were not necessarily well prepared for A level, as they could do well in the GCSE but still not have had the skills to progress further in computer science. However, experts also believed that programming skills would be beneficial, along with creativity and problem-solving skills.

Grading standards

It was noted that reducing the expectations for each grade may benefit some students but might risk the integrity of the subject. This was born out by our discussions with the experts following task 1, where there was consensus that work only a few marks below the grade boundaries did not show enough knowledge and skills necessary for each grade. A small number of experts commented that they believed the quality of work at the boundaries was actually below what they expected, particularly at grade 7.

Those experts that expressed concern that the current exam standard was too challenging were not consistent about where in the grade range their concerns were focussed, with different experts suggesting that they thought the standard was too challenging at the higher grades (grade 7/8/9) or at the middle grades (4 and 5). Other experts instead thought that the assessment was inaccessible to weaker students at lower grades (4 and below). There was also some suggestion from experts that it was relatively easy to gain a grade 1.

Strand 2. Discussion

This strand aimed to review the performance standard in GCSE computer science in summer 2023, by this we mean the quality of work demonstrated by students to receive the key grades (grade 7 and grade 4). We did this by seeking the views of a group of 8 experts with a diverse range of experience in computer science. These experts represented a variety of views from teachers, industry experts and subject bodies. Overall, there were mixed views from subject experts about whether the current performance standard is appropriate. Some subject experts believed the quality of work at the boundaries was lower and others higher than expected, although slightly more experts believed that the standard of work at grade 7 was higher than expected.

The results of the first task and discussions during the review meeting indicate that there is a region around each grade boundary where the experts did not consistently identify a noticeable difference in the quality of work produced by students (extending to -2 marks below the boundary at grade 7, -4 marks at grade C). Therefore, for these assessments, if the grade boundary had been anywhere within this range, it would have a negligible impact on the performance standard of the qualification. Evidence from experts suggested that moving slightly lower down the mark distribution (-4/6 marks at grade 7 and -6 marks at grade 4), the difference in the quality of work became more noticeable. However, in discussion, experts believed that moving the standard within this range would not undermine the purpose of the qualification. Experts believed that moving any further than this would result in students showing noticeably less skills and knowledge and would have a significant negative impact on the qualification standard.

The second task aimed to understand how well students receiving different grades in GCSE computer science were prepared for further study in the subject, representing one of the main purposes of GCSEs. The results of this task were mixed. The findings suggested that the experts believed that students with a high grade 5 could be successful in further study. To some experts, however, this came a surprise, as it would be unusual to accept a student with a grade 5 to study A level computer science. From discussions with the experts though it was apparent that that it can be difficult to judge which students will do well at A level. Experts believed success typically has less to do with students' subject content knowledge on entering the A level than their attitude and approach to the subject. This may be because a number of students take the A level without having taken the GCSE and so teachers presume very little, or patchy knowledge from students entering the A level. Experts may therefore have been generous in their rating of students' work for this exercise, looking to find evidence of potential even when students performed poorly in the assessment.

Finally, discussions with the experts indicated that although there was some belief that the assessments in GCSE computer science were challenging, the experts thought that there are a large number of other potential reasons that the qualification is seen as too difficult, beyond exam grading standards. Principally among these may be issues with recruiting subject specialist teachers, challenges of validly assessing programming skills and the breadth of content in the qualification.

To conclude, overall, the experts considered that there was some justification for adjusting the standard in GCSE computer science, although views on this were mixed. The findings suggests that the standard in the assessment could be lowered by a small degree, without undermining the qualification, but any larger changes would potentially be considered undesirable by subject experts. Broader issues with the perception of difficulty within GCSE computer science were highlighted which cannot be addressed through changes to the assessment standard or through grade boundary setting.

Overall Conclusion

The aim of strand 1 of this study was to understand if there was any evidence of a change in standards in GCSE computer science over time, which may have led to the subject being more difficult than intended. Across a variety of methods there was an indication of a small change in standards over time, particularly during the period 2014 to 2017. During this period there were a large number of changes to the qualification in terms of the number and make-up of students taking the qualification, the number of new centres entering students to the qualifications for the first time, and some changes to the assessment design and structure. These changes produce challenges in maintaining standards, which in this case may have led to some small incremental changes to the qualification standard. Given that such changes were likely to be small, they are unlikely to have been detectable by senior examiners when setting grade boundaries each year. Cumulatively though, this appears to have led to a more substantive change in standards. Across the methods used in strand one, evidence suggested that there had been a small change in standards at grade A/7 between the period 2014 to 2019, and a slightly larger change at grade C/4. Evidence for any change in standards at grade G/1 was weak.

In strand 2 we aimed to explore what the impact of any change in the standard of the qualification would be on the skills and knowledge demonstrated by students in the assessments and to understand what impact this might have on student progression. The findings indicated that a small change in standards at grade 7 and grade 4 would have a minor impact on the performance standard for each grade, and that this would be unlikely to impact on the progression of students to further study in computer science. However, any larger changes would start to have undesirable consequences for the skills and knowledge that our subject experts would expect students to demonstrate and may risk undermining the value of the qualification. The other feedback from experts in strand 2 did not indicate that a larger change in grading standards was felt to be necessary. Subject experts highlighted a number of factors which may influence the perceived and actual difficulty in the subject beyond grading standards in the assessments. These include teacher expertise, curriculum time, subject content and resourcing.

In summary, the evidence in this report suggests that consideration should be given to making an adjustment to grading standards in GCSE computer science. Evidence from strand 1 indicates that there is likely to have been a small change in standards over time in the qualification, and the findings from strand 2 suggest that a small adjustment to grading standards is unlikely to undermine the value of the qualification or the progression of students to further study in computer science.

References

Benton, T. (2013). *Formalising and evaluating the benchmark centres methodology for setting GCSE standards.* Cambridge Assessment Research Report.

Benton, T. S. T., & Sutch, T. (2014). *Analysis of use of Key Stage 2 data in GCSE predictions.* ARD Research Division.

Bradley, R.A., & Terry, M. (1952). The rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39, 324–345.*

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

Bramley, T., & Oates, T. (2011). Rank ordering and paired comparisons - the way Cambridge Assessment is using them in operational and experimental work. *Research Matters: A Cambridge Assessment Publication, 11, 32-35*

Brown, N. C., Sentance, S., Crick, T., & Humphreys, S. (2014). Restart: The resurgence of computer science in UK schools. *ACM Transactions on Computing Education (TOCE), 14(2), 1-22.*

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

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

Cresswell, M.J. (2003). *Heaps, prototypes and ethics: The consequences of using judgements of student performance to set exam standards in a time of change.* University of London Institute of Education.

Cuff, B. M., 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.

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots.* Ofqual.

Dallaway, E. (2016). *GCSE Reform: A New Dawn of computer science.* CREST.

DfE (2015). *computer science: GCSE Subject content.* Department for Education.

DfE (2023). *Reporting year 2022: School workforce in England*. National Statistics.

Good, F. J., & Cresswell, M. J. (1988). Grade awarding judgements in differentiated examinations. *British Educational Research Journal*, *14(3), 263-281.*

He, Q. and Black, B. (2020). *Impact of calculated grades, centre assessment grades and final grades on inter-subject comparability in GCSEs and A levels in 2020.* Ofqual.

He, Q. and Cadwallader, S. (2022). *An investigation of inter-subject comparability in GCSEs and A levels in summer 2021.* Ofqual.

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

Kemp, P.E.J. & Berry, M.G. (2019). *The Roehampton Annual Computing Education Report: pre-release snapshot from 2018.* University of Roehampton

Newton, P. (2020). *What is the Sawtooth effect?* Ofqual.

Ofqual (2015a). *Further Decisions for Completing GCSE, AS and A Level Reform in 2017*. Ofqual.

Ofqual (2015b). *Comparability of Different GCSE and A Level Subjects in England: An Introduction.* Ofqual.

Ofqual (2016). *Decisions on setting the grade standards of new GCSEs in England - part 2*. Ofqual.

Ofqual (2017). *Consultation on revised assessment arrangements for GCSE computer science.* Ofqual.

Ofqual (2019). *Decisions on future assessment arrangements for GCSE (9 to 1) computer science*. Ofqual.

Ofqual (2023). *GCSE (9 to 1) Qualification Level Conditions and Requirements.* Ofqual.

Ofsted (2022). *Research review series: computing.* Ofsted.

Royal Society (2017). *After the reboot: Computing education in UK schools*. The Royal Society.

Royal Society (2019). *Policy briefing on teachers of computing: recruitment, retention and development.* The Royal Society.

Sibieta, L. (2018). *The teacher labour market in England: shortages, subject expertise and incentives.* Education Policy Institute.

Appendix A – Progression analysis modelling output

Table A1. Linear model output for the relationship between mean GCSE score and A level outcomes between years. Model includes control variables for Ethnic group, FSM eligibility, Language group, Gender, SEN status and Centre type, coefficients not shown.

Variable	**Coefficient**	**SE**	**p value**
Standardised KS2 score	0.353	0.014	<0.001
Standardised mean GCSE	0.806	0.012	<0.001
Year 2015 [2014]	0.116	0.056	<0.05
Year 2016 [2014]	0.049	0.051	0.336
Year 2017 [2014]	0.107	0.051	<0.05
Marginal r-squared	0.391		
Conditional R-squared	0.473		
N Students	12103		
N Centres	1778		

Appendix B – Outcomes over time modelling output

Table B1. Summary of Year model effects from various different models using prior attainment. See text for details.

Model	Restriction	Year 2019 coefficient [Ref 2014] (SE)	Estimated difference in outcomes from 2014 predicted for 2019 cohort
Linear	All centres	-0.12 (0.01)***	-0.15
Linear	Excluding new centres	-0.35 (0.04)***	-0.39
Linear	2014 centres only	-0.24 (0.05)***	-0.27
Linear	2015 centres only	-0.25 (0.03)***	-0.28
A/7 Grade	All centres	-0.02 (0.03)	0.19pp
A/7 Grade	Excluding new centres	-0.28 (0.07)***	-2.75pp
A/7 Grade	2014 centres only	-0.23 (0.08)**	-4.27pp
A/7 Grade	2015 centres only	-0.11 (0.06)	-2.04pp
C/4 Grade	All centres	-0.08 (0.03)**	-1.81pp
C/4 Grade	Excluding new centres	-0.53 (0.08)***	-10.47pp
C/4 Grade	2014 centres only	-0.21 (0.10)*	-3.85pp
C/4 Grade	2015 centres only	-0.24 (0.06)***	-4.30pp
G/1 grade	All centres	-0.05 (0.07)	-0.06pp
G/1 grade	Excluding new centres	-0.91 (0.30)**	-0.53pp
G/1 grade	2014 centres only	-0.24 (0.38)	0.00pp
G/1 grade	2015 centres only	-0.02 (0.20)	-0.06pp

*Note. Statistical significance is indicated by p<0.001 '***' p<0.01 '**'p<0.05 '*'*

Table B2. Detailed model output for prior attainment linear models. Student and centre characteristic control variables not shown, see text for details.

Variable	M1 (all centres)	M2 (no new centres)	M3 (same centres 2014)	M4 (same centres 2015)
Year 2012 [2019]	-0.054 (0.039)	NA	NA	NA
Year 2013 [2019]	0.071 (0.026)**	NA	NA	NA
Year 2014 [2019]	0.128 (0.014)***	0.35 (0.038)***	0.241 (0.045)***	NA
Year 2015 [2019]	0.129 (0.012)***	0.296 (0.025)***	0.223 (0.046)***	0.248 (0.03)***
Year 2016 [2019]	0.105 (0.009)***	0.176 (0.013)***	0.031 (0.042)	0.167 (0.027)***
Year 2017 [2019]	0.059 (0.008)***	0.08 (0.01)***	-0.171 (0.042)***	-0.012 (0.027)
Year 2018 [2019]	-0.025 (0.008)**	-0.035 (0.009)***	-0.059 (0.042)	-0.097 (0.027)***
Standardised KS2 score	1.069 (0.003)***	1.118 (0.004)***	0.995 (0.015)***	1.031 (0.01)***
R-Squared (Marginal/conditional)	0.36/0.46	0.39/0.49	0.38/0.45	0.38/0.47
N (students/centres)	297014/3432	173787/2662	12198/85	26238/203

*Note. Statistical significance is indicated by $p<0.001$ ‘***’ $p<0.01$ ‘**’$p<0.05$ ‘*’*

Table B3. Detailed model output for prior attainment A/7 binomial models. Student and centre characteristic control variables not shown, see text for details.

Variable	M1 (all centres)	M2 (no new centres)	M3 (same centres 2014)	M4 (same centres 2015)
Year 2012 [2019]	-0.317 (0.071)***	NA	NA	NA
Year 2013 [2019]	-0.101 (0.049)*	NA	NA	NA
Year 2014 [2019]	-0.019 (0.028)	0.283 (0.069)***	0.233 (0.084)**	NA
Year 2015 [2019]	-0.091 (0.023)***	0.139 (0.048)**	0.186 (0.086)*	0.111 (0.057)
Year 2016 [2019]	0.062 (0.018)***	0.184 (0.027)***	0.157 (0.079)*	0.241 (0.052)***
Year 2017 [2019]	0.018 (0.017)	0.065 (0.021)**	0.032 (0.08)	0.092 (0.052)
Year 2018 [2019]	-0.009 (0.017)	-0.011 (0.018)	-0.048 (0.079)	-0.102 (0.052)*
Standardised KS2 score	1.45 (0.008)***	1.511 (0.01)***	1.338 (0.034)***	1.401 (0.024)***
R-Squared (Marginal/conditional)	0.39/0.48	0.41/0.5	0.38/0.44	0.4/0.47
N (students/centres)	297014/3432	173787/2662	12198/85	26238/203

*Note. Statistical significance is indicated by $p<0.001$ ‘***’ $p<0.01$ ‘**’$p<0.05$ ‘*’*

Table B4. Detailed model output for prior attainment C/4 binomial models. Student and centre characteristic control variables not shown, see text for details.

Variable	**M1 (all centres)**	**M2 (no new centres)**	**M3 (same centres 2014)**	**M4 (same centres 2015)**
Year 2012 [2019]	-0.125 (0.077)	NA	NA	NA
Year 2013 [2019]	-0.025 (0.048)	NA	NA	NA
Year 2014 [2019]	0.081 (0.026)**	0.525 (0.082)***	0.211 (0.101)*	NA
Year 2015 [2019]	0.084 (0.021)***	0.363 (0.051)***	0.067 (0.1)	0.244 (0.062)***
Year 2016 [2019]	0.002 (0.015)	0.085 (0.025)***	-0.243 (0.088)**	0.031 (0.055)
Year 2017 [2019]	-0.052 (0.015)***	-0.024 (0.019)	-0.578 (0.087)***	-0.196 (0.055)***
Year 2018 [2019]	-0.04 (0.015)**	-0.055 (0.016)***	-0.206 (0.09)*	-0.21 (0.055)***
Standardised KS2 score	1.419 (0.007)***	1.512 (0.009)***	1.408 (0.037)***	1.412 (0.024)***
R-Squared (Marginal/conditional)	0.39/0.5	0.43/0.52	0.44/0.53	0.43/0.52
N (students/centres)	297014/3432	173787/2662	12198/85	26238/203

*Note. Statistical significance is indicated by p<0.001 '***' p<0.01 '**'p<0.05 '*'*

Table B5. Detailed model output for prior attainment G/1 binomial models. Student and centre characteristic control variables not shown, see text for details.

Variable	**M1 (all centres)**	**M2 (no new centres)**	**M3 (same centres 2014)**	**M4 (same centres 2015)**
Year 2012 [2019]	0.17 (0.251)	NA	NA	NA
Year 2013 [2019]	-0.136 (0.125)	NA	NA	NA
Year 2014 [2019]	0.051 (0.068)	0.907 (0.305)**	0.238 (0.385)	NA
Year 2015 [2019]	0.159 (0.057)**	0.499 (0.153)**	0.218 (0.385)	0.024 (0.2)
Year 2016 [2019]	-0.107 (0.037)**	-0.036 (0.061)	-0.775 (0.288)**	-0.374 (0.175)*
Year 2017 [2019]	-0.174 (0.035)***	-0.172 (0.046)***	-1.176 (0.273)***	-0.935 (0.16)***
Year 2018 [2019]	0.012 (0.036)	-0.027 (0.04)	-0.477 (0.303)	-0.418 (0.172)*
Standardised KS2 score	1.177 (0.014)***	1.279 (0.019)***	1.159 (0.086)***	1.216 (0.057)***
R-Squared (Marginal/conditional)	0.34/0.52	0.38/0.52	0.92/0.95	0.65/0.76
N (students/centres)	297014/3432	173787/2662	12198/85	26238/203

*Note. Statistical significance is indicated by p<0.001 '***' p<0.01 '**'p<0.05 '*'*

Table B6. Detailed model output for concurrent attainment linear models. Student and centre characteristic control variables not shown, see text for details.

Variable	**M1 (all centres)**	**M2 (no new centres)**	**M3 (same centres 2014)**	**M4 (same centres 2015)**
Year 2012 [2019]	-0.005 (0.029)	NA	NA	NA
Year 2013 [2019]	0.127 (0.02)***	NA	NA	NA
Year 2014 [2019]	0.117 (0.011)***	0.411 (0.027)***	0.314 (0.034)***	NA
Year 2015 [2019]	0.128 (0.008)***	0.35 (0.017)***	0.257 (0.032)***	0.331 (0.021)***
Year 2016 [2019]	0.097 (0.006)***	0.208 (0.01)***	0.078 (0.031)*	0.237 (0.02)***
Year 2017 [2019]	0.046 (0.006)***	0.076 (0.007)***	-0.128 (0.031)***	0.008 (0.02)
Year 2018 [2019]	-0.018 (0.006)**	-0.021 (0.006)***	-0.063 (0.031)*	-0.051 (0.02)*
Standardised mean GCSE score	1.592 (0.002)***	1.625 (0.003)***	1.485 (0.011)***	1.527 (0.007)***
R-Squared (Marginal/conditional)	0.65/0.7	0.68/0.73	0.64/0.68	0.65/0.7
N (students/centres)	321117/3442	185439/2654	13663/84	28408/203

*Note. Statistical significance is indicated by p<0.001 ‘***’ p<0.01 ‘**’p<0.05 ‘*’*

Table B7. Detailed model output for concurrent attainment A/7 binomial models. Student and centre characteristic control variables not shown, see text for details.

Variable	**M1 (all centres)**	**M2 (no new centres)**	**M3 (same centres 2014)**	**M4 (same centres 2015)**
Year 2012 [2019]	-0.356 (0.084)***	NA	NA	NA
Year 2013 [2019]	-0.058 (0.059)	NA	NA	NA
Year 2014 [2019]	0.024 (0.033)	0.476 (0.08)***	0.4 (0.097)***	NA
Year 2015 [2019]	-0.064 (0.025)*	0.308 (0.052)***	0.257 (0.093)**	0.3 (0.063)***
Year 2016 [2019]	0.114 (0.021)***	0.328 (0.031)***	0.255 (0.091)**	0.441 (0.061)***
Year 2017 [2019]	0.031 (0.02)	0.1 (0.025)***	0.046 (0.092)	0.127 (0.062)*
Year 2018 [2019]	-0.021 (0.019)	-0.018 (0.022)	-0.078 (0.091)	-0.118 (0.061)
Standardised mean GCSE score	3.064 (0.013)***	3.2 (0.018)***	2.911 (0.055)***	3.036 (0.04)***
R-Squared (Marginal/conditional)	0.71/0.75	0.73/0.77	0.7/0.72	0.71/0.75
N (students/centres)	321117/3442	185439/2654	13663/84	28408/203

*Note. Statistical significance is indicated by p<0.001 ‘***’ p<0.01 ‘**’p<0.05 ‘*’*

Table B8. Detailed model output for concurrent attainment C/4 binomial models. Student and centre characteristic control variables not shown, see text for details.

Variable	M1 (all centres)	M2 (no new centres)	M3 (same centres 2014)	M4 (same centres 2015)
Year 2012 [2019]	-0.05 (0.091)	NA	NA	NA
Year 2013 [2019]	0.079 (0.059)	NA	NA	NA
Year 2014 [2019]	0.061 (0.03)*	0.769 (0.095)***	0.389 (0.113)***	NA
Year 2015 [2019]	0.11 (0.022)***	0.657 (0.056)***	0.269 (0.107)*	0.568 (0.068)***
Year 2016 [2019]	-0.01 (0.018)	0.181 (0.03)***	-0.195 (0.099)*	0.206 (0.065)**
Year 2017 [2019]	-0.087 (0.017)***	-0.046 (0.023)*	-0.659 (0.098)***	-0.196 (0.063)**
Year 2018 [2019]	-0.021 (0.017)	-0.032 (0.019)	-0.24 (0.1)*	-0.141 (0.063)*
Standardised mean GCSE score	2.994 (0.012)***	3.231 (0.017)***	2.849 (0.059)***	2.941 (0.041)***
R-Squared (Marginal/conditional)	0.69/0.75	0.73/0.78	0.69/0.74	0.7/0.75
N (students/centres)	321117/3442	185439/2654	13663/84	28408/203

Note. Statistical significance is indicated by p<0.001 ‘***’ p<0.01 ‘**’p<0.05 ‘*’

Table B9. Detailed model output for concurrent attainment G/1 binomial models. Student and centre characteristic control variables not shown, see text for details.

Variable	M1 (all centres)	M2 (no new centres)	M3 (same centres 2014)	M4 (same centres 2015)
Year 2012 [2019]	0.235 (0.275)	NA	NA	NA
Year 2013 [2019]	0.016 (0.15)	NA	NA	NA
Year 2014 [2019]	-0.088 (0.074)	0.921 (0.325)**	0.279 (0.393)	NA
Year 2015 [2019]	0.057 (0.056)	0.546 (0.154)***	0.196 (0.355)	0.216 (0.198)
Year 2016 [2019]	-0.205 (0.041)***	-0.005 (0.068)	-0.744 (0.292)*	-0.211 (0.181)
Year 2017 [2019]	-0.248 (0.038)***	-0.226 (0.051)***	-1.264 (0.278)***	-0.886 (0.165)***
Year 2018 [2019]	0.014 (0.039)	-0.002 (0.044)	-0.524 (0.305)	-0.227 (0.177)
Standardised mean GCSE score	2.533 (0.02)***	2.761 (0.029)***	2.319 (0.12)***	2.525 (0.082)***
R-Squared (Marginal/conditional)	0.61/0.71	0.66/0.73	0.93/0.95	0.72/0.79
N (students/centres)	321117/3442	185439/2654	13663/84	28408/203

Note. Statistical significance is indicated by p<0.001 ‘***’ p<0.01 ‘**’p<0.05 ‘*’

Published by:

ofqual

2nd floor
1 Friargate
Station Square
Coventry
CV1 2GN

0300 303 3344

public.enquiries@ofqual.gov.uk

www.gov.uk/ofqual

July 2024 Ofqual/24/7143/1