Research Matters / 36
A Cambridge University Press & Assessment publication

ISSN: 1755-6031

Journal homepage: https://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

An analysis of the relationship between Secondary Checkpoint and IGCSE results

Tim Gill (Research Division)

To cite this article: Gill, T. (2023). An analysis of the relationship between Secondary Checkpoint and IGCSE results. *Research Matters: A Cambridge University Press & Assessment publication, 36*, 59–74. https://doi.org/10.17863/CAM.101745

To link this article: https://www.cambridgeassessment.org.uk/Images/research-matters-36-an-analysis-of-the-relationship-between-secondary-checkpoint-and-igcse-results.pdf

Abstract:
Secondary Checkpoint assessments are taken by students at the end of the Cambridge Lower Secondary programme (aged 14) in countries around the world. Many students continue with Cambridge after this and take IGCSE exams two years later.

Given that there is a high level of coherence between the curricula in the two stages, performance in Secondary Checkpoint should be a good indicator of performance at IGCSE.

In this article, I investigate whether there is evidence to support this contention, by calculating correlations between Checkpoint scores and IGCSE grades, across a range of subjects. I also look at whether students in schools offering the Cambridge Lower Secondary programme go on to perform better at IGCSE than schools not offering the programme.

An analysis of the relationship between Secondary Checkpoint and IGCSE results

Tim Gill (Research Division)

Introduction

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

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

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

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

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

Predictive validity studies

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

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

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

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

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

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

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

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

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

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

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

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

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

Impact of taking a particular qualification

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

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

Aim of the current research

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

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

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

Predictive validity of Secondary Checkpoint

Data and methods

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

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

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

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

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

Table 1: Progression from Checkpoint to IGCSE.

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

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

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

Matching process

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

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

Results

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

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

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

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

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

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

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

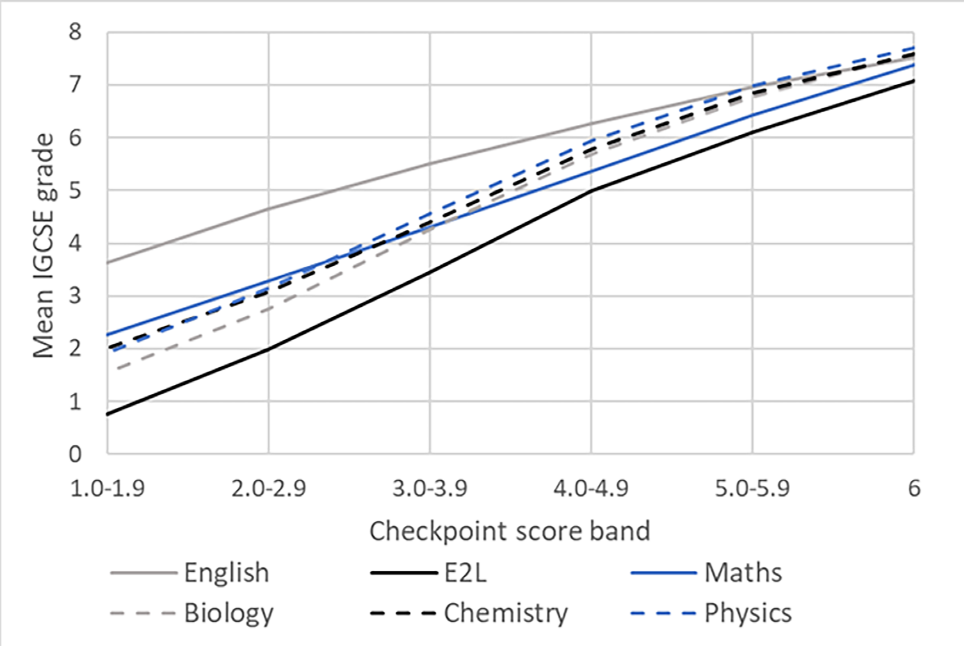

Figure 1: Mean IGCSE grade for each Checkpoint score band, by subject

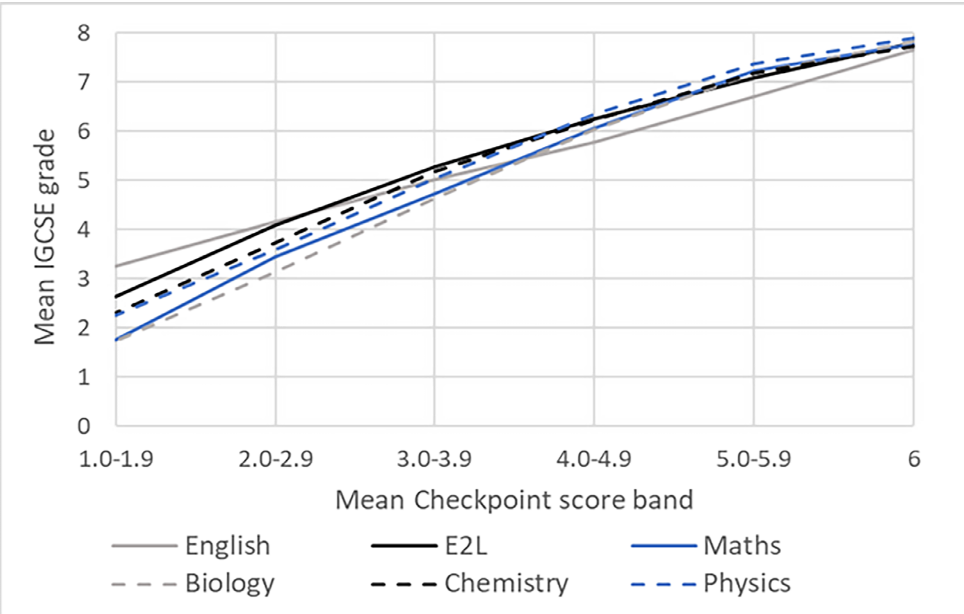

Figure 2: Mean IGCSE grade for each mean Checkpoint score band, by subject

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

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

Impact of adopting Checkpoint on IGCSE results

Data and methods

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

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

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

The general form of the model was:

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

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

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

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

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

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

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

Results

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

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

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

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

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

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

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

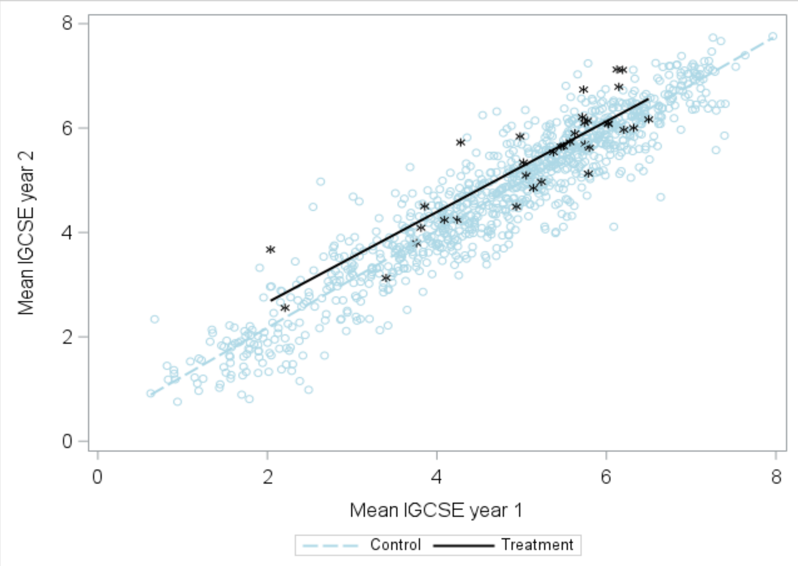

Figure 3: Mean IGCSE in Y1 and Y2, split by treatment group

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

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

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

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

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

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

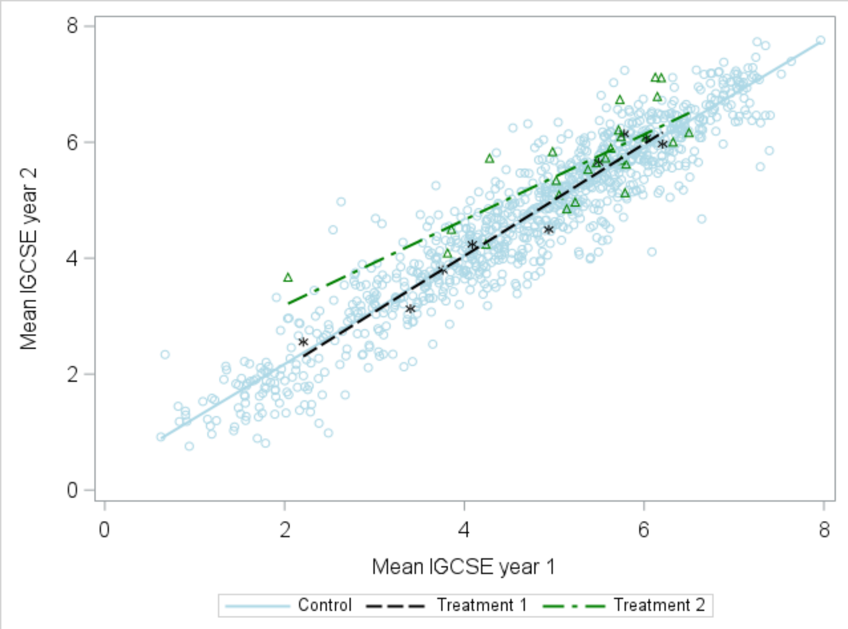

Figure 4: Mean IGCSE in Y1 and Y2, split by three-level treatment group

Regression analysis

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

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

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

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

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

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

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

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

Discussion

Predictive validity of Checkpoint

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

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

Impact of adopting Checkpoint on IGCSE results

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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