Students' effort and educational achievement: Using the timing of the World Cup to vary the value of leisure

Robert Metcalfe
Boston University

Simon Burgess
University of Bristol

Steven Proud
University of Bristol

September 2018

Abstract

We study the effect of a sharp, exogenous, and repeated change in the value of leisure on educational achievement, arising from the overlap of major international football tournaments with high-stakes tests. Using administrative data covering almost all students in England, we find a significant negative average effect of the tournament on exam performance. The odds of reaching the achievement benchmark fall by 12% on average, considerably more for students likely to be interested in football. Analysis of within-student variation shows a 0.02 SD fall in grades, 0.06SD for the interested. We interpret our results as reflecting changes in student effort.

Keywords: educational achievement, student effort, schools, value of leisure
JEL Codes: I20, J24

Acknowledgments

We are very grateful to the Editor and anonymous referees for very helpful and insightful comments. We are grateful to the DfE for access to the NPD data, and to the ESRC for funding through CMPO. We are grateful for comments to Rebecca Allen, Patty Anderson, Dan Hamermesh, Lindsey Macmillan, Andrew Oswald, Carol Propper, Helen Simpson, Sarah Smith, Liz Washbrook, Nicolas Ziebarth and seminar participants at ISER, the Geary Institute and CMPO.

1. Introduction

We study the effect of a sharp, exogenous and repeated change in the value of leisure on educational achievement. In many countries around the world, achievement is assessed at least in part by examinations. In England, performance is measured using universal high-stakes tests that students in England take at the end of compulsory schooling, which are scheduled at the end of the academic year in May and June. Also held at this time are the world's two most-watched football tournaments (the FIFA World Cup and the UEFA European Championship), and these overlap with the exam period. These tournaments are both attention-grabbing and highly salient for many students in England, and substantially raise the value of leisure time for many students. They happen every other summer, so each year is sequentially either a treatment year (even-numbered years) or a control year. We estimate the overall effect of the tournament and compare within-student variation in performance during the exam period between tournament and non-tournament years using seven years of student-subject data on practically all the students in England. This data allows us to bring out the heterogeneity of impact as well as quantifying the average effect. The maximum potential treatment is very strong: the tournaments always completely dominate TV, radio, and other media during the weeks it takes place; for example, the 2018 World Cup Quarter Final match involving England was watched by 87.7% of all TV viewers in England.[1]

The treatment is ideal for a causal study. Identification comes straightforwardly from the fact that the students and schools are faced with two timetables outside of their individual control: the schedule of the international tournaments and the schedule of the exams season. Every other year there is no tournament, providing a set of alternating treatment and control periods. Furthermore, since there is essentially no grade repetition in England, exposure to the treatment is random: simply whether a student is born to take her exams in an even year or an odd year. The only potential room for manoeuvre is the choice of optional subjects, but these are chosen over two years ahead of the relevant exams, under guidance from schools, and for reasons we explain below, are very hard to strategize. Furthermore, because only the second half of the exams period overlaps with the tournaments, this generates within-student variation in the value of leisure in tournament years.

The important high-stakes examinations in England (called the General Certificate in Secondary Education or GCSE) are achievement tests, testing both knowledge and skills. They are always scheduled for May and June at the end of compulsory schooling (at age 16). We use administrative data on subject-level grades for almost all students in England over seven years. We obtained data on exam timetables for each subject, and compare these with the tournament dates. We observe neither time spent thinking about the tournament nor hours of study time, so this is an intention to treat study, but

[1] See https://www.theguardian.com/media/2018/jul/08/england-world-cup-win-sweden-watched-by-20m-bbc-tv-viewers

we have enough data to estimate heterogeneity in effects across groups with varied likely interest in football.

We find a significant negative average effect of the tournament on exam performance. We first estimate the overall effect of the tournament on the standard benchmark in England of student achievement, that is, whether the student achieved at least good passes (C grade or better) in at least 5 subjects. We find that in tournament years the odds of achieving this fall by approximately 12% (or by 6 percentage points). However, the impact is much more substantial for some demographic groups, students who are likely to be very interested in football (defined below). For them, the odds of reaching this benchmark decline by 28% in tournament years. This is particularly noteworthy as this group is already a low-achieving group in England. The lower performance is likely to have a persistent impact on the life-chances of students as this metric is important for a student's chance of continuing in education and for job prospects. The rate of return to achieving at least 5 good passes has been estimated at around 25% - 30% for women and 28% - 31% for men (McIntosh, 2006, and Dearden et al., 2002). Also, while it might be thought that a period of just a few final weeks of intense studying would not have long run effects on human capital, Falch et al. (2014) have shown that even a few days of intense preparation for high-stakes tests (as in this paper) makes a difference to longer run outcomes such as enrolment in higher education. Looking at within-student variation, we find that average effect is a 0.017 SD fall in grades in tournament years, and for the "keen on football" sub-group achievement falls by 0.065 SD. For the sub-group, these are very sizeable effects.

Our study is a reduced form analysis of the relationship between achievement and the timing of major international football tournaments. We interpret the substantial overall effects as an overall reduction in effort and engagement, on top of any reallocation of effort between different subjects at exam time. We see the effects we find as evidence of the importance of student 'effort', broadly understood to include both the quantity of study time and its quality in terms of concentration on the task. The within-student analysis rules out many alternative explanations of our results. Also, student effort is not confounded with teacher effort: there is no teaching at this time of year, leaving the student time to study alone. An effort interpretation matters because it has implications for understanding test score performance and for designing policies to improve that. There is a small but growing literature on incentivizing students to raise effort, some showing substantial positive effects of financial incentives on primary/elementary and secondary/high school students.[2] Recently, Oswald and Backes-Gellner (2014) have shown that incentives for learning have a much greater effect on impatient students than patient ones. This helps in the interpretation of our results. Patient students understand the value of qualifications and do not need incentives to promote study, whereas they are an effective additional

[2] For instance, Angrist et al. (2002), Henry & Rubinstein (2002), Jackson (2010), Kremer et al. (2009), Angrist & Lavy (2009), Dearden et al. (2009), Dee (2011), Levitt et al. (2011) and Pallais (2009), although others demonstrate a lack of positive effects of financial incentives on educational attainment, e.g., Bettinger (2010), Sharma (2010), Fryer (2010) and Rodriguez-Panas (2010).

stimulus for the impatient.

There are a few other studies examining how much effort matters for test scores. Announcing financial incentives prior to a test has a positive and significant effect on test scores (see Braun et al., 2011; Levitt et al., 2012; O'Neil, 1997; 2004). One other study has attempted to examine the link between changes in the value of leisure and achievement, and interpreting this as effort. Lindo et al. (2012) show how college students in the University of Oregon are negatively affected in their exams when the college football team is doing well.[3] While there are a number of similarities, there are also some important differences between the studies. First, the results of Lindo et al. relate to a single university, whereas this research exploits data on almost all students in England; as a result, our results generalize to the whole student population. Second, the nature of our data and the coincident timing of the high-stakes tests and the football tournaments provide us with both within-student and between-student variation in exposure to the treatment, raising internal validity. Stinebrickner & Stinebrickner (2008) found that randomly being assigned to a college roommate who has a video game console significantly reduces time allocated to studying (using self-completed surveys), which then negatively impacts on educational achievement. The estimated effect may include peer effects as well as the changing marginal value of leisure.

The next section describes the nature of student assessment and the timing of the exams and the tournaments; section 3 describes our data. We present our results in section 4 and section 5 discusses the implications of our results.

2. Student assessment and tournament timing

We first describe the nature of assessment in England: the subjects taken, the structure and grading of exams, and the assessment of a student's overall performance. Second, we explain the timing of the tests and how these overlap with the timing of the tournaments.

a. Assessment, subjects and tests

Education in England is organised into Key Stages, and our focus is Key Stage 4 (KS4), covering the final two years of compulsory schooling, with students aged 15 to 16. There is essentially no grade repetition in England, each student moves up to the next school year (grade level) each year. Each Key Stage finishes with assessments, and at the end of KS4 this takes the form of a series of assessments,

[3] Clotfelter (2011) examines the impact on journal article downloads in universities that have their basketball teams in the NCAA tournaments for a longer period of time, but does not examine the impact on actual attainment.

separately for each subject; these are known as GCSEs. This process is high stakes, crucial for continuing in school or looking for jobs.

Subjects and exams

Students take on average around eight subjects at GCSE, and most students will attempt at least five. Among these, English, maths and science are compulsory[4]; others are chosen from a list of possibilities. All these subjects are studied for two years up to the summer exams.

Typically, a subject will be assessed by several different exams, held on separate days, plus also some coursework. For example, in Maths in 2002 there were 4 different exams held on separate days over a period of 13 days. All the exams are nationally set and are marked remotely and independently from the school. For each subject, marks from all of the exams and the coursework are combined into a single grade for that subject: from A*, A, B, …G and U (indicating a Fail). The distribution of subject grades is not normed each year, and the average grade increased year on year over our period. Our data include each student's grade in each subject, but the results for each individual exam are not readily available.[5] On average, coursework contributes about half of the final grade, but this varies somewhat between subjects, a fact that we exploit in our analysis below. The variation is not great, but we distinguish subjects with a high weight on exams (English Language, English Literature, Geography, Mathematics, History) and those with a medium weight (Science, Design and Technology, French, German, IT studies, Media Studies).[6]

Choice of subjects

Optional subjects are chosen over two years in advance of the exam period. While obviously the occurrence of the football tournaments is fully predictable, potential differential overlap of this with the exams is probably not a major reason for subject choice. Even if some students considered this, some subjects' exam dates switch between late and early year-to-year, making those students' attempts to strategize subject choice more or less impossible. Students are also strongly advised by their schools, so it is not plausible that they would choose subjects on such grounds.

[4] While Maths and English are straightforward in the date, science is more complicated to code as there are different ways of taking science. Some schools offer it as three separate subjects: Biology, Chemistry and Physics; others do one course called Science, but devote twice as much time to it and its double-weighted; there are also much more applied courses that are science-equivalents.

[5] This rules out an even more fine-grained analysis, for example looking at the exam score the day after an important match.

[6] We omitted the following minor subjects from this analysis: Business Studies, Drama, Latin, Music, PE, Religious Education, Sociology, and Spanish. In these, the weight attached to coursework varied across exam boards. In England, there are different providers of exams and schools differ in which they use. We cannot use that variation as our data do not tell us which exam board was used by each school. For the main subjects, exam boards used the same weighting between exams and course-work.

Overall student assessment

The widely used summary measure of performance is whether they achieved at least a C grade in at least 5 subjects. This is the typical benchmark[7] used for a student's chance of continuing in education and for job prospects, and is what we use below to measure the overall impact of the tournaments. During the period our data cover, some 50-60% pupils did achieve at least 5 C grades or better, rising over the period. While this criterion of "getting at least 5 C grades or better" matters for the pupil and the school, it is not all that matters. Getting more and higher-graded GCSEs is better. It matters not least for access to university as the GCSE grades are the latest actual marks that a student has on her CV: she is much more likely to be made a university offer with (say) 8 C grades than 5 Cs and 3 fails, or with 8 A* grades than 5 A*s and 3 Cs. The public record, and what goes on CVs, is not just a binary achieved/did not achieve this benchmark; a typical CV might record "3A*s, 4As, 2Bs" for example.

b. Timing of the tournaments and of the exams

The football tournaments

Every four years (on even years) the FIFA World Cup takes place in June and July, and every other four years (on the different even years, so always two years apart) the UEFA European Championships also take place in June and July.[8] The FIFA World Cup attracts a massive worldwide audience. For instance, the 2006 World Cup in Germany had television coverage in 214 countries around the world, with 73,000 hours of dedicated programming, which generated a total cumulative television audience of 26.29 billion people (FIFA, 2007). The UEFA European Championships are not as large as the World Cup, although in the 2008 Euro tournament, the final rounds were shown in over 200 countries (UEFA, 2008). Appendix Table 1 reports the time frame for the World Cups and European Championships[9] from 2002 to 2008, the years in our data, alongside a summary of the exam timetable (discussed further below).

Demonstrating the salience of the tournaments for students in England

Football is the national sport in England and generates huge interest. For example, latest figures[10] show that attendance at one weekend's ten matches in the top flight (Premier League) was around 375,000 people, totalling around 14.3m attendances over a whole season. TV viewing data provide useful support for our assumption that watching the major international football tournaments is a very

[7] The school system in England does not have a "graduation" status, there is no graduation requirement. Pupils reaching the final year of compulsory schooling either continue on to more years in school and possibly university, or they simply stop going to school. Students who have passed no GCSE subjects simply leave school with that on your CV.

[8] The history and background to the FIFA World Cup and European Championships can be found at http://www.fifa.com/worldcup/ and http://www.uefa.com/ respectively.

[9] As some readers may know, England did not progress very far through the knock-out stages of any of these tournaments. We considered whether we could differentiate between exams sat before and after England were eliminated, but in fact the team did manage to remain in the tournament for almost all the exam period.

[10] https://www.worldfootball.net/attendance/eng-premier-league-2018-2019/1/

widespread phenomenon. Appendix Table 2 shows that football programmes dominate the list of most-watched programmes in the UK in this window. To scale these numbers, 26.8 million households owned a TV in 2010.[11]

Contemporary data is very powerful in demonstrating the huge popularity of football in England in relative terms too. First, England's quarter-final match in the 2018 World Cup was watched by 87.7% of everyone watching BBC TV in the UK at that time.[12] Nor is it just England games: the 2018 Portugal-Spain game was watched by over 40% of all UK TV viewers. While this is some ten years later than our analysis sample, we believe that the fragmentation of viewing habits over the last decade would suggest that if anything the fraction of viewers watching the tournament might well have been higher in the mid-2000s. Of course, the interest in the tournaments extends beyond watching TV: 80% of all regular UK Twitter users tweeted within 30 minutes of the end of England's semi-final game in 2018.[13]

Overlap of tournament and exam schedules

In each of these years, the tournament overlapped with national UK examinations. We report the start dates and end dates of the GCSE exam season in Appendix Table 1. There is little difference in exam dates between those years in which there is a football tournament and those in which there is not. The proportion of exams during the football tournament ranges between 46% and 61%. The data on examination dates for each subject were obtained from Cambridge Examinations[14]. The list of subjects used in our analysis is provided in Appendix Table 3. Some subjects have no exams during the tournament, others have a proportion of exams during the football period, and others have all the exams during the football period.

Using the dates of each individual exam, we calculate the proportion of exams for a subject that are within the time period of the football tournament, for each subject in each year in our data. We measure this degree of overlap with a continuous variable P_{tz}, the proportion of exams in subject z that are scheduled during the tournament period in year t. In non-tournament years, we take the calendar dates from the previous tournament year to define late subjects.

We also looked at the timing of exams for each student, specifically whether for some students exams were closely bunched and for other students they were spread out over the exam period. In fact, there is very little variation in this across students. We computed the range of exam dates by student

[11] https://www.barb.co.uk/resources/tv-ownership/.
[12] The 87.8% represents 19.6m viewers at peak, 15.8m on average, relative to an all-ages population of the UK of 66m. Interest in this game is likely to be close to a maximum level of interest, to illustrate the reach of the football tournaments into the attention of the population.. See for details: https://www.theguardian.com/media/2018/jul/08/england-world-cup-win-sweden-watched-by-20m-bbc-tv-viewers.
[13] https://www.sporttechie.com/twitter-world-cup-brazil-england/
[14] Although different examination boards set their own exams, the exams of different boards for the same subject across the country are on the same day.

(measured as the standard deviation in days of exam dates), and the mean range was 11.3 days with an SD of 1.6 days; there are very few students below 8 or above 13.

An alternative to this proportion overlap measure is to use the TV viewing data in a very fine-grained way to distinguish affected exams and unaffected exams. In fact, there is very little variation over time, the tournament as a whole is very popular. The historical TV data we could access was the top 30 programs per TV channel (BBC1, BBC2, and ITV) per week (hence the top 90 programs per week). These show that there is at least one game every day that is in the top 90 programs for that week. So once the tournament has started (i.e. in the 'late' period), the TV data actually provide little additional variation in attention across 'late' subjects. By the time interested students have taken in the pre-match speculation and post-match analysis, and viewed potential opponents' games, there is something to grab attention all the time. While we present some robustness analysis on the impact of games with the very highest viewing figures of all, we chose to use the early-late difference as our main measure of potential disruption to study.

Note that the 'early' versus 'late' categorization of subjects does not match the compulsory versus optional categorization of subjects. Among compulsory subjects, English and maths are early and science is late (though it switches to early in 2008); and there are early optional subjects as well as late ones – for example, from Appendix Table 3, in 2002, History (taken by 198k students) overlapped 100% with the tournament, as did IT (149k); French (321k) overlapped 25%, and Geography (211k) overlapped 0%.

To emphasise, we know the dates of all of the individual exams within each subject, but we only know the overall subject grade, not the marks on individual exams. Hence we cannot connect the date of a specific exam to the mark on a specific exam; what we do is to relate the overall grade in a subject to the degree of overlap of the exams in that subject with the tournament.

Dealing with 2008

The English national football team qualified for the first three of the four international tournaments in our sample, but not for the 2008 European Championships. The TV data show considerably less UK interest in the tournament in 2008: first, only half as many matches were shown on TV (23/40) relative to 2006; second audiences were 7% down on 2006 and there were no "media frenzy" games in 2008 with a viewership of over 8 million. We therefore classify 2008 as a "non-tournament" year, and test the robustness of our results to this decision. We test whether this assumption on 2008 is pivotal in our results.

3. Student Data

The data on students are taken from the National Pupil Database (NPD). This is an administrative

dataset of all students in the state-maintained system, some 93% of all students, made available to researchers by the Department for Education. It includes a census of students, taken each year in January, from 2002 onwards. In each cohort there is approximately half a million students, and so over the seven year period we use we have some 3.5 million students. We have data on each student's gender, within-year age, ethnicity, an indicator of Special Educational Needs (SEN, which measures learning or behavioural difficulties), whether English is an additional language, and eligibility for Free School Meals (FSM), which is dependent on eligibility for welfare benefits and is widely used as a measure of family poverty.

The key idea in this paper is that the value of leisure increases for some individuals when a major football tournament takes place. This taste for watching football will depend on cultural factors and an idiosyncratic component, which we expect to be substantial. The cultural factors may be associated with observable student characteristics, for example social class, ethnicity and gender. An interest in football is by no means confined to the working class or to men, but in England it remains a bigger part of male culture than female culture.[15] Football has also been more associated with lower income and working class families (see Baker, 1979; Goldblatt, 2006). Asians are traditionally perceived as being much less interested in football than Whites (see for example, McGuire et al., 2001, and Kilvington and Saeed, 2011). This may in part be because there are not many role models for them: according to the English Football Association[16], of the 3,000 professional footballers in the English leagues in 2017, only 10 of them are Asian; it does not seem likely that there would have been more in 2002 – 2008.

We use this information to produce two composite measures proxying a student's likely level of interest in football. We create a dummy for "likely to be keen on football", equal to one if the student is male, eligible for Free School Meals, and White (this group accounts for 4.1% of our sample); and a dummy for "unlikely to be keen on football", equal to one if the student is female, not eligible for Free School Meals, and of Asian ethnicity (2.2% of our sample). These small and narrow groups are an attempt to define students with relatively homogenous tastes for the tournaments, to more precisely pin down the impact of the increased value of leisure on grades.

We also have data on student test scores. Our analysis uses the subset of students that are identifiable within the state system throughout this period, which amounts to 90% of the cohort.[17] Our outcome variable is the grade each student receives for each subject. Very few students re-take exams, so there is almost always just the one grade per student per subject; in cases where there is more than one grade

[15] To the extent that male and female students share leisure time, there will be an indirect effect on females too. This of course could go either way: if the boys are watching the game, the girls might as well watch too, or if the boys are watching the game, the girls might as well study.

[16] https://www.bbc.co.uk/news/av/uk-england-birmingham-41428071/son-challenges-dad-why-couldn-t-i-go-to-man-utd-trial

[17] Those that are excluded may have attended a private school for a period, may have spent time abroad (including Wales or Scotland), or may have been entirely educated in the English state system but their Unique Pupil Number was lost during a school transfer.

we take the results in the Summer of the academic year when they turn 16. We assign numerical values to the letter grades (A*, A, B etc) using the National Curriculum points system. We normalise the scores separately for each subject to remove any differences in subject difficulty; obviously the normalisation is done over all the years together as our focus is on across-year within-subject variation. As a measure of prior achievement, we use test scores in English, Maths and Science taken at age 11 (these are the tests taken at the end of Key stage 2, KS2). We average the three subjects and then take three quantiles of that combined score. These tests are compulsory for all students, and are also set and marked at a national level, remote from the school. One important and useful feature is that these tests are always taken in early May and so never overlap with the tournament.

Table 1 provides an overview of our data. Over this period, around 12% of students are eligible for FSM, and around 85% of the students are white. The subset of students who have both "late" and "early" subjects account for 81.4% of the total data. This subset differs a little: those students taking only "early" subjects are slightly more likely to be poor, and have slightly lower prior achievement. Since the compulsory subjects have "early" exams, there are no students only taking "late" exams.

4. Empirical Results

We first discuss identification and the empirical framework. Second, we present results for the overall impact of the tournament on the measure of student overall performance. Third, we present detailed results on the within-student analysis, including discussion of heterogeneous responses, robustness analysis and a focus on passing subjects.

a. Empirical Framework and Identification

The central idea is that the football tournaments dramatically raise the value of leisure time for some people, and correspondingly reduce the relative value of all other time uses, including the effort put into studying. We have no data on effort, so our analysis is reduced form linking the timing of football tournaments to subject grades. Identification comes directly from the two exogenous time frames with which students are faced. The timetables for the football tournaments are fixed, always around the same time of year and outside their control. Tournaments recur every other year so we have a series of alternating treatment and control years. The overall timetable for the tests is similarly fixed and outside individual students' control. Eligibility for treatment depends solely on a student's year of birth, whether they will take their summer exams in an even-numbered year (with tournaments) or an odd-numbered year. There is very little grade repetition, so treatment is essentially inescapable.

Of course, other things happen in our control years that will change the value of leisure for some students. On average, the distribution of personal events should be about the same each year. At a national level too, most events of note happen every year and so wash out of our analysis.[18]

The only room for variation at all is in the choice of optional subjects. It would be potentially problematic if some students avoided taking optional subjects with late exams in tournament years. Even so, our conditional within-student difference should take out any first order effect of differences in unobservable student characteristics. We compared the prior achievement and other characteristics of students picking late or early options in tournament and non-tournament years. We run a difference-in-difference, comparing the mean ability of students taking late options with those taking early options, and then difference that difference across tournament and non- tournament years. The results show no difference at all in the average KS2 score (prior ability). The diff-in-diff coefficient is positive, but only 0.001 of an SD and not significantly different from zero (even in a regression of 12.2 million observations). So the counter-story that unobservably smarter students switch out of late options in anticipation of the tournament/exam clash some two and half years ahead is not supported by the evidence on observed ability. Because optional subjects are chosen over two years ahead of the tests, and because the timing of those tests varies from year to year, it is very unlikely that students would be able to strategize their choices of optional subjects to minimise the overlap with the tournament.

Because the tournaments only partially overlap with the exam period, we can extend the analysis to study within-student differences in performance. We compare student performance in subjects by the percentage overlap with the tournament dates, between tournament years and non-tournament years. Identification in this analysis builds on that above (students facing two exogenous time frames and no real scope for choice), by taking within-student difference to remove the effect of unobserved individual characteristics. This approach, however, can only identify the late-early difference (the 'tilt' between scores in late subjects and early subjects), but not the overall effect.

b. Overall impact of the tournament

Table 2 presents the results. We take the binary overall outcome measure of whether the student achieved at least a C grade in at least 5 subjects, and estimate a logit model on our full set of personal characteristics and a dummy for whether the student's exam year was a tournament year[19]; we also include a time trend as the overall pass rate was trending up over this period[20]. Column 1 shows that the

[18] Following the sporting theme, we have looked at other potential events around this time. In tennis, the Wimbledon championships happens every year and start much later; sticking with football, the later stages of the Champions League have involved significant representation of English clubs throughout the period along a rising trend; while the Champions League final is typically at the start of the exam period, the other knock-out matches finish before exams.

[19] The advantage for logit in this context is that there are groups of students with chances of passing exams close to zero, and others with a chance close to one, and the functional form of logit does better at these tails.

[20] Omitting the time trend in fact strengthens the tournament effect.

average effect (odds ratio) is 0.883 and strongly significant. As we have emphasised, there is likely to be important heterogeneity around that average, and we display two ways of capturing that (we model heterogeneity more fully below in the within-student analysis). First, in column 2, we utilise our 'keen on football'/ 'not keen on football' composite measure interacted with the tournament dummy. We see that those likely to be keen on football show a substantially greater decline in their chances of hitting the benchmark, the interaction odds ratio being 0.815. Second, in column 3, we show a much greater impact of the tournament on poor, male students, and the least impact on non-poor female students. In column 4, we exclude the minority of students who only take 6 or fewer GCSE subjects, with no important effects on the results.

The size of the effect is substantial. For students likely to be keen on football, the substantially lower probability fall of achieving 5 good passes is noteworthy because they are already by far the lowest performing group, with only 21.3% achieving the benchmark (in non-tournament years) relative to 60.1% for non-poor female students. Overall achievement is lower and inequality in achievement is higher among cohorts of students taking their exams in tournament years.

c. Within-student analysis – results on student grades

The within-student analysis highlights the effect on a student's grades for subjects for which a lot of the tests are scheduled during the tournament period ('late'), relative to the grades for the same student in subjects for which most of the tests happen just before the tournament begins ('early'). Further, because we have a series of years alternating between tournament and non-tournament, we can track how this late-early difference evolves over time. We begin by simply illustrating this point graphically. For each year, we take the average of GCSE scores in 'late' subjects and subtract the average GCSE score in 'early' subjects and plot out this short time series in Figure 1. We see a clear pattern of higher late-early differences in non-tournament (odd-numbered) years than in tournament years, with the exception of 2006[21].

We now turn to regression analysis to quantify this effect more precisely and to discuss statistical significance.

Results on subject grades

We present first the main results, illustrating the average impact of the tournament on grades, subject-by-subject. We also investigate heterogeneous effects by observable student characteristics. We estimate for student *i* in year *t* in subject *z*:

[21] We note that the 2006 tournament was the height of the (mis-placed) "golden generation" hype around the England football team. It may well be that for this occasion far more than the other tournaments in our sample, the 'early' pre-tournament interest was as high as the in-tournament 'late' interest, thus reducing both the 'late' and 'early' scores by similar amounts and leaving the late-early difference at roughly non-tournament levels.

$$y_{itz} = \alpha + \beta.P_{tz} + \delta.I(year = T).P_{tz} + \gamma.X_i + \epsilon_{itz} \quad (1)$$

where y is the subject grade, P_{tz} is the proportion of subject z's exams that overlap the tournament in year t, and X_i is observable student characteristics. The term β captures the early – late gradient in exam scores in non-tournament years and $\beta + \delta$ is the gradient in tournament years. This relates the grades across subjects within-student to the timing of the exams relative to the tournament.[22] The source of variation is the timetabling of exams across subjects, and our main coefficient of interest, δ, is the comparison of this in tournament and non-tournament years. We also estimate a specification including student fixed effects, μ_i instead of student characteristics:

$$y_{itz} = \alpha + \beta.P_{tz} + \delta.I(year = T).P_{tz} + \mu_i + \epsilon_{itz} \quad (2)$$

To capture heterogeneous effects of the tournaments, we interact both β and δ with X:

$$y_{itz} = \alpha + \beta.P_{tz} + \pi.X_i.P_{tz} + \delta.I(year = T).P_{tz} + \theta.X_i.I(year = T).P_{tz} + \gamma.X_i + \epsilon_{itz} \quad (3)$$

and

$$y_{itz} = \alpha + \beta.P_{tz} + \pi.X_i.P_{tz} + \delta.I(year = T).P_{tz} + \theta.X_i.I(year = T).P_{tz} + \mu_i + \epsilon_{itz} \quad (4)$$

Here θ is the key parameter of interest. The full set of individual controls in X_i is: male dummy, FSM dummy, SEN dummies, within-year age whether English is an additional language, ethnicity dummies, KS2 English, KS2 Maths, and KS2 Science.

The results are in Table 3, reporting the key coefficients of interest, β and θ, omitting the coefficients on student characteristics. The dependent variable is a student's grade in an individual subject. As the error term is likely to be correlated across each students' exam results, we cluster at the student level; we discuss this assumption below. Column (1) reports the results of estimating equation (1) with student characteristics X plus year dummies; column (2) adds first, interactions of the proportion of late examinations with the full set of student characteristics, X, and second, interactions of the proportion of late examinations with whether the year is a tournament year and with student characteristics. For brevity, we only report selected interactions from the second set (θ) and none from the first set (π). Column 3 drops the student characteristics and introduces student fixed effects. Column 4 retains the student fixed effects and adds interactions of the proportion of late examinations with the full set of student characteristics, X, and interactions of the proportion of late examinations with whether the year is a tournament year and with student characteristics; again, we only report selected interactions from the tournament year set of interactions. Column 5 also retains the student fixed effects and replaces both sets of interactions with X, with interactions with our composite variables "Likely to be keen on football" and "Unlikely to be keen on football".

[22] Recall that potential differences in the difficulty of the subject are dealt with by normalising the scores by subject.

The key coefficient is the interaction of the proportion of late exams and the tournament year dummy. This is negative and well defined, significantly different from zero at the 1% level, and of consistent size across the columns. We discuss the magnitude of the effect below. The β coefficients describe the production technology in non-tournament years, and show that on average in non-tournament years, students tend to do better in subjects with more exams late on in the period. However, this time difference is very variable around this average rising performance. Figure 2 shows the distribution of mean late subject grade minus mean early subject grade across all students, differentiating tournament and non-tournament years. It confirms that the average is positive, and that the whole distribution shifts down in tournament years, but also shows a huge range of values indicating different performance technologies.

Heterogeneity of response

As we have emphasised throughout, not all students are transfixed by football tournaments and we expect significant heterogeneity in response. We address that first in a simple parametric fashion in this table. In columns 2 and 4 we interact the variable of interest with student characteristics (the main effects of these variables are subsumed in the fixed effect in column 4). As we might expect, the impact of many of the interactions is attenuated somewhat once we control for fixed effects, though many remain significant.

Part of the role played by these characteristics is capturing a high or low interest in the tournament. Students eligible for FSM suffer a substantial additional penalty. The estimates show that students for whom English is their first language also see a greater decline in their grades. We see that male students suffer a significantly more negative effect than females, though it is worth emphasising that the effect for females is significantly negative. Interactions with ethnicity markers, not reported here for brevity, go the same way (controlling for language). Relative to white pupils, Black Caribbean pupils have a greater penalty (-0.04 in col 4 specification), but Bangladeshi (+0.03), Indian and Pakistani students have positive coefficients. This finding fits well with those ethnic groups' typical sporting preferences having less of a focus on football. Thus poor boys are one of the hardest hit groups, white and Black Caribbean boys in particular.

These effects are plausibly interpreted as mainly picking up cultural differences, proxying a strong interest in watching the football tournament. The results in column 5 support that idea with our composite variable "likely to be keen on football" showing a strong additional negative effect, and the opposite "unlikely to be keen on football" showing an offsetting effect to the average, a small positive effect overall.

The results on differential impacts by prior ability are inconclusive: the coefficients change sign between columns 2 and 4 and are generally smaller than those just described. If we take the specification with student fixed effects as our main result, we see that high ability students face less of a penalty than

low ability ones. The impact of the tournament on exam scores is the product of fewer hours of study, depending on interest in football, and the effectiveness of those lost hours. On the former, ability is also likely to be correlated with cultural factors influencing tastes, and the latter will depend on the student's ability. In these results it seems that the cultural effect is stronger.

Often there are different assumptions about the appropriate level to cluster standard errors (see for example Abadie et al., 2017) and in this case there are plausible underlying mechanics and motivations for a number of different approaches. To us, the most natural approach is to cluster at pupil level, because we observe the same individual several times across multiple subjects and exams. This therefore accounts for common and unobserved levels of ability, motivation, focus, conscientiousness, interest in football and so on. Second, it is plausible to argue for clustering at school level. This is common in considering pupil-level grades, and accounts for commonality of teaching, school-level messages on dealing with distractions and the importance of the exams and so on. Third, the treatment is at year level – alternate years have a tournament or do not, and this is the source of the impact on exam scores. Following the idea of clustering at the assignment level suggests clustering at year level. Fourth, the treatment varies between subjects, depending on whether more of the exams happen within or before the tournament itself; this suggests clustering at exam-year level or subject-year level.

We show the effect of different assumptions about clustering the standard errors in Appendix Table 4, panel A. Given the size and structure of the dataset, this is a case where different assumptions make very big differences to the number of clusters. Clustering at pupil level (equivalently described as pupil*year level as each pupil only appears in one year), yields 3,651,667 clusters in our dataset of around 25m observations in Table 3. Clustering at school level gives 3,286 clusters, at year level just 7 clusters, and at subject*year level gives 161 clusters. Clearly, the number of clusters makes a very big difference, with standard errors 60 times larger when we assign our 25m observations to just a few clusters. While in general, as applied economists we would like to be as conservative as possible, there are cases when "as conservative as possible" is too extreme and unrealistic. Using just 161 clusters at subject-year level for 25m observations seems too conservative to us. In Appendix Table 4 panel B, we provide evidence to suggest that the correlation of residuals within pupils is far greater than the correlation within subject-year groups, also supporting the approach of clustering standard errors at pupil level. We have therefore chosen to adopt pupil level clustering as the main specification and use that in presenting our results.

Robustness tests

In Table 4 we report the results of testing various assumptions we make underlying our main the results. First, we can use TV viewing data instead of relying on timetabling information to distinguish disruption to study. Because of the ubiquity of football discussion in public consciousness over the tournament period, we prefer the broader timetabling measure, but column 1 of table 4 shows an alternative approach. We chose a high cut-off for viewership because there are popular games every day. We

created a variable at subject*year level, the proportion of exams in the subject that are the day after a match that gained 4 million or more viewers (this is sufficient to ensure it always appears in the Top 30 programmes for BBC1, BBC2 or ITV, the three largest free-to-air TV stations); obviously in non-tournament years the proportion is zero. The results fit well with our main results, yielding a strongly significant negative impact.

Second, we explore heterogeneity of response across subjects placing different weights on the summer exams relative to coursework in the overall grade for the subject.[23] The variation in weight is not great, but we distinguish subjects with a high weight on exams (English Language, English Literature, Geography, Mathematics, History) and those with a medium weight (Science, Design and Technology, French, German, IT studies, Media Studies). If students allocate their effort across subjects in a non-strategic way then we would expect a greater negative effect purely mechanically in subjects where the exams matter more. If students were to plan strategically, however, anticipating this outcome, they might spread their effort and this would dampen the mechanical effect of weighting. The results are in column 2 and show that the greater the weight on exams, the greater the impact on the final grade. This seems plausible to us, and suggests that students do not fine-tune their effort.

Third, it is possible in the English system to sit exams a year earlier than the age-correct year. This is not uncommon in maths (rarer in other subjects) and is obviously an endogenous decision. In column 3, we report the effect of dropping the students who sit subjects early (i.e. all their subject grades) makes no qualitative difference to the estimates. This supports the idea that few students are strategizing the timing of their exam taking with the tournament in mind.

Fourth, we counted the year 2008 as a non-tournament year for students in England as the national team did not qualify for the European Championships; clearly, that tournament generated less interest than the previous ones, it was still more than a non-tournament year. The simplest way to deal with the ambiguity here is to estimate just for the first six years of our sample, 2002 – 2007. This also makes the sample more symmetric through time between treatment and control years.[24] The results are in column 4 and show a much stronger effect of tournaments.

Fine-grained heterogeneity analysis

We briefly explore the diverse impacts of the tournament in a more flexible way, to exploit the size and richness of our dataset. We match groups within school (around 2000 schools), and grouping on student gender, FSM status, prior achievement group, ethnic group and quarter of birth.[25,26] Each student in a tournament year is matched with a student in a non-tournament year in the same school and defined by

[23] We are grateful to a referee for this idea.

[24] We are grateful to the Editor for this suggestion.

[25] We use three broad groups, working below the expected level (Key stage 2 score below 27), working at the expected level (KS2 of 27), or above the expected level.

[26] We use four aggregated groups: White, Black, Asian and Other.

the same set of observables. We take the difference between the mean score on late subjects and early subjects for each student, and then average this within each school*observables group, separately for tournament and non-tournament years, and analyse the difference.

Appendix Table 5 shows selected quantiles from the distribution of this difference-in-difference across around 15,000 school-groups with at least 40 students in. The focus of this table is the extremes not the average.[27] The table reveals that some experience a very substantial effect. For example, 10% of groups have an effect more negative than -0.26 SDs. The 10th percentile for male students is -0.28 SDs and for disadvantaged students is -0.35 SDs. These are difference-in-differences with all school and student factors accounted for, they do not simply reflect low levels of ability or performance. At this fine level of matching there is likely to be a good deal of noise and the assumption of the same noise structure for all groups may not be valid. So we should be appropriately cautious in interpreting these results, but they suggest that significant reductions in student effort and concentration will yield significant reductions in achievement.

d. Within-student analysis – results on passing subjects

While the grades on these subjects matter for the students' prospects it is important to look simply at whether students pass each subject, that is, get at least grade C. Arguably the impact of passing or not passing is much greater than that of just moving down a grade across the distribution. We therefore repeat the analysis of Table 3 at student*subject level, using a logit model to analyse a binary variable, whether the grade in the subject was at least a C. The results are in Table 5. The analysis differs from that of Table 3 in just two respects. First, we have omitted Science, because with a number of different variants of Science GCSE, establishing an overall grade by merging the qualification types is easier than establishing whether it has been passed or not. Second, we restrict the sample to students who pass at least one subject and fail at least one subject, so that the number of observations/students is the same across columns.

The main finding is confirmed. The interaction between proportion of exams being late and it being a tournament year is strongly and significantly negative. The impact is sizeable, an estimated odds ratio for passing of 81% in tournament years. Many of the heterogeneity terms are the same, but not all. FSM eligibility and English language status again show further negative effects; being male has a sizeable negative effect in column 2, but is small and positive in column 4. Both medium and high ability students have positive effects, suggesting that in terms of crossing this specific C grade threshold, the cultural/taste factors correlated with ability are the most important factor. Ethnic groups (not reported)

[27] These are school-group level quantiles of student level data, some groups being very significantly larger than others, and the quantiles are unweighted. So there is not an immediate read-across to the individual student*subject-level means implied by the regressions in table 2; they are based on exactly the same sample, data and calculations.

match the results in Table 3, Black Caribbean students have a negative coefficient and Asian ethnic groups typically positive. In column 5, the composite summary terms show a very consistent pattern – significantly negative for those likely to be keen on football, and a slight positive differential effect for those unlikely to be keen. Finally, we re-run the analysis just on a sub-sample of students who are marginal around the C/D border.[28] The results are given in Appendix Table 6 and show that the result is in fact stronger for this group.

5. Conclusion

We study the effect of a sharp, exogenous, and repeated change in the value of leisure on educational achievement. This arises from the fact that the world's most watched international football tournaments overlap with the high-stakes testing period in England. We analyse the overall impact of the tournament on pass rates and compare within-student variation in achievement over the exam period between tournament and non-tournament years. We have seven years of student*subject data on practically all the students in England, and these data allow us to bring out the heterogeneity of impact as well as quantifying the average effect.

We find a significant negative overall effect of the tournament on exam performance, more substantial for some groups. The key summary measure of performance for students in England is whether they achieve at least 5 good passes (i.e. at least a C grade). The results in Table 2 show that the odds ratio of achieving this is lower by 12% in tournament years. For students we define as likely to be keen on football it is reduced by 28% in tournament years. This result is made more noteworthy by the fact that it is already a low achieving group. This achievement benchmark is highly significant for students in terms of prospects for higher education, jobs and lifetime income[29]. Looking at within-student variation, we find that average effect[30] is a 0.017 SD fall in grades in late subjects relative to early in tournament years, and for our "keen on football" subgroup, likely to make a substantial switch of time to the tournament, achievement falls by 0.065 SD.[31]

[28] To identify marginal students, we created a binary indicator variable which takes the value 1 if the student has achieved a grade C or D, and 0 otherwise. For each exam year separately, we regressed this on all of the characteristics used here (age in months, FSM status, gender, English as a first language, SEN status, and ethnicity, using a linear probability model. These results were used to create a likelihood of being in that C/D range and the most likely 20% of students in each exam year were then selected.

[29] McIntosh (2006) and Dearden et al. (2002) estimate a rate of return of between 25% - 30%.

[30] This is measured in student-level SDs. We get to this from the subject-level SD in the tables as follows: the coefficient (from Table 3, col 3) multiplied by the mean across subjects proportion of late exams (0.22), multiplied by the mean number of subjects taken (7.80), multiplied by an adjustment to switch from subject level SD to pupil level SD (1.081).

[31] Note that the results described are impacts on the overall subject grades. As we have seen, the summer exams only constitute part of the overall grade, so the impact on the exams scores will be considerably higher. On average of about 60-70% weight on the final exams with the remainder coming from coursework throughout the year. The compulsory subjects of maths, English and Science average about 60%. Given this, the answer to the question "how much does a rise in the value of leisure reduce study time and therefore performance?" is about 1.67 times higher than the estimates just presented.

These are non-trivial changes. A comprehensive welfare calculation should also include the utility from watching the tournament, or its loss from focusing on study.[32] However, the utility loss from not watching one tournament will make only a minor offset to the gain in lifetime income from better qualifications, particularly when noting that tournaments will occur every other year for the rest of the student's life.

We interpret these results as arising from students watching and thinking about the football tournament, rather than studying over the key weeks of exams. We interpret the variation in impact as principally reflecting differing tastes for football, arising in turn from cultural norms and idiosyncratic factors.

Our results relate to two issues: a broad debate on the nature of educational achievement function, and a policy question specific to England about the timing of summer exams. Taking the second first, scheduling important exams during the tournament period reduces educational achievement, particularly for disadvantaged, low ability, male students. This is a group that has lower performance anyway, and our results show that the tournament has a substantial negative effect on their performance. This in turn will affect their likelihood of progression through the educational system and their lifetime income. Given this, the benefits of moving the exams just a few weeks earlier in tournament years are significant. The costs would depend on how the earlier exams were accommodated: starting the school year a little earlier, fewer revision days before the exams, or reducing the school year. The size of the benefits suggests a serious comparison of cost and benefit would be merited.

Our interpretation of the results as variations in effort carry implications for the understanding of educational achievement. First, unlike genetic characteristics, cognitive ability or non-cognitive traits, effort is almost immediately changeable. Our results suggest that this could have a big effect. This ties in with recent results on policies aimed at raising achievement. Fryer's (2010), Jackson's (2010) and Levitt et al.'s (2011; 2012) results suggest that directly paying students for greater effort has an impact on test scores. Second, the dramatic test score gains cited for "No Excuses" schools in the KIPP and HCZ or some Charter schools (Abdulkadiroglu et al., 2011; Angrist et al., 2011; Dobbie and Fryer, 2009) can plausibly be interpreted as those environments eliciting greater effort from the students. The fact that we find changes in student effort to be very potent in affecting test scores suggests that policy levers to raise effort through incentives or changing school ethos are worth considering seriously. Such interventions would be justified if the low effort resulted from market failures due to lack of information on the returns to schooling, or time-inconsistent discounting.

Third, an effort-based interpretation links to work on the importance of non-cognitive traits in educational achievement (see Cunha et al., 2010). Non-cognitive factors can be identified with personality traits (Heckman, 2011), and one of the 'big 5' personality traits is 'conscientiousness', with

[32] We thank a referee for this point.

the related traits of self-control, accepting delayed gratification, and a strong work ethic (Heckman, 2011, p. 5); it is also related to the rate of time preference (Daly et al., 2009). Conscientiousness has been shown to be an excellent predictor of educational attainment and course grades (Almlund et al. 2011, Borghans et al. 2011). These aspects of self-control and ability to concentrate are clearly related to the broad notion of effort we are using here.

References

Abadie, A., Athey, S., Imbens, G. and Wooldridge, J. (2017) When should you adjust standard errors for clustering? NBER Working Paper 24003, National Bureau of Economic Research, Cambridge, MA.

Abdulkadiroglu, A., Angrist, J., Cohodes, S., Dynarski, S., Fullerton, J., Kane, T. and Pathak, P. (2011) "Informing the Debate: Comparing Boston's Charter, Pilot and Traditional Schools." Forthcoming, *Quarterly Journal of Economics*

Almlund, M., A. Duckworth, J. J. Heckman, and T. Kautz (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, and L. Woessmann (Eds.), Handbook of the Economics of Education, Volume 4. Amsterdam: Elsevier. Forthcoming.

Angrist, J. and Lavy, V. (2009) The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial. *American Economic Review*, 99: 4, pp. 1384-1414.

Angrist, A., Dynarski, S., Kane, T., Pathak, P, and Walters, C. (2011) Who benefits from KIPP? IZA DP 5690

Angrist, A., Pathak, P, and Walters, C. (2011) Explaining Charter School Effectiveness. NBER WP 17332.

Angrist, J., Bettinger, E., Bloom, E., King, E. and Kremer, M. (2002) Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review*, 92: 5, pp. 1535-1558.

Baker, W. J. (1979) The Making of a Working-Class Football Culture in Victorian England. *Journal of Social History*, 13: 2, pp. 241-251.

Bettinger, E. P. (2010) Paying to Learn The Effect of Financial Incentives on Elementary School Test Scores *NBER working paper series no. w16333* (Cambridge, Mass., National Bureau of Economic Research).

Borghans, L., B. H. H. Golsteyn, J. J. Heckman, and J. E. Humphries (2011). IQ, achievement, and personality. Unpublished manuscript, University of Maastricht and University of Chicago (revised from the 2009 version).

Braun, Henry, Irwin Kirsch, and Kentaro Yamamoto. 2011. "An Experimental Study of the Effects of Monetary Incentives on Performance on the 12th-Grade NAEP Reading Assessment." Teachers College Record.

Clotfelter, Charles T. (2011). *Big-Time Sports in American Universities*. Cambridge University Press, Cambridge, Mass., USA.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010, May). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica* 78 (3), 883-931.

Daly, M., L. Delaney, and C. P. Harmon (2009). Psychological and biological foundations of time preferences. *Journal of the European Economic Association* 7 (2-3), 659-669.

Dearden, L., Emmerson, C., Frayne, C. and Meghir, C. (2009) Conditional Cash Transfers and School Dropout Rates. *Journal of Human Resources*, 44: 4, pp. 827-857.

Dearden, L., McIntosh, S., Myck, M. and Vignoles, A. (2002) 'The returns to academic, vocational and basic skills in Britain.' *Bulletin of Economic Research*, vol. 54, pp. 249–274.

Dee, T. S. (2011) Conditional cash penalties in education: Evidence from the Learnfare experiment. *Economics of Education Review*, 30: 5, pp. 924-937.

Dobbie, W. and Fryer Jr., R. G. (2009) Are high quality schools enough to close the achievement gap? Evidence from a social experiment in Harlem. NBER WP 15473

Falch, T., Nyhus, O. H. and Strøm, B. (2014) 'Causal effects of mathematics' Labour Economics vol. 31 pp. 174–187

Goldblatt, D. (2006) *The ball is round: a global history of football*. (London, Viking).

Heckman, J. J. (2011) Integrating personality psychology into economics. NBER WP 17378.

Henry, G. T. and Rubenstein, R. (2002) Paying for grades: Impact of merit-based financial aid on educational quality. *Journal of Policy Analysis and Management*, 21: 1, pp. 93-109.

Jackson, C. K. (2010) A Little Now for a Lot Later A Look at a Texas Advanced Placement Incentive Program. *Journal of Human Resources*, 45: 3, pp. 591-639.

Kilvington, D. and Saeed, A. (2011) 'British-Asians and racism within contemporary English football', *Soccer and Society*, 12, 5: 602-612;

Kremer, M., Miguel, E. and Thornton, R. (2009) Incentives to Learn. *Review of Economics and Statistics*, 91: 3, pp. 437-456.

Levitt, S., List, J. and Sadoff, S. (2011) The Effect of Performance-Based Incentives on Educational Achievement: Evidence from a Randomized Experiment, *mimeo*.

Levitt, S., List, J., Neckermann, S. and Sadoff, S.. (2012). The Behavioralist Goes To School: Leveraging Behavioral Economics to Improve Educational Performance. NBER Working Paper No. 18165.

Lindo, Jason M., Isaac D. Swensen, and Glen R. Waddell (2012). Are Big-Time Sports a Threat to Student Achievement? *American Economic Journal: Applied Economics*, 4 (4), 254-274

McGuire, B., Monks, K. and Halsall, R. (2001) 'Young Asian males: social exclusion and social injustice in British professional football?', *Culture, Sport and Society*, 4, 3: 65-80.)

Mcintosh, S. (2006) 'Further Analysis of the Returns to Academic and Vocational Qualifications.' *Oxford Bulletin of Economics and Statistics*, vol.68 (2), pp. 305-49

O'Neil, Harold F., Jr., Jamal Abedi, Charlotte Lee, Judy Miyoshi, and Ann Mastergeorge. 2004. "Monetary Incentives for Low-Stakes Tests." (CSE Rep. No. 625). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

O'Neil, Harold F., Jr., Brenda Sugrue, Eva L. Baker, and S. Golan. (1997). "Final report of experimental studies on motivation and NAEP test performance." (CSE Tech. Rep. No. 427). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Oswald, Y. and Backes-Gellner, U. (2014) Learning for a bonus: How financial incentives interact with preferences. *Journal of Public Economics* vol. 118 pp. 52–61

Pallais, A. (2009) Taking a Chance on College Is the Tennessee Education Lottery Scholarship Program

a Winner? *Journal of Human Resources*, 44: 1, pp. 199-222.

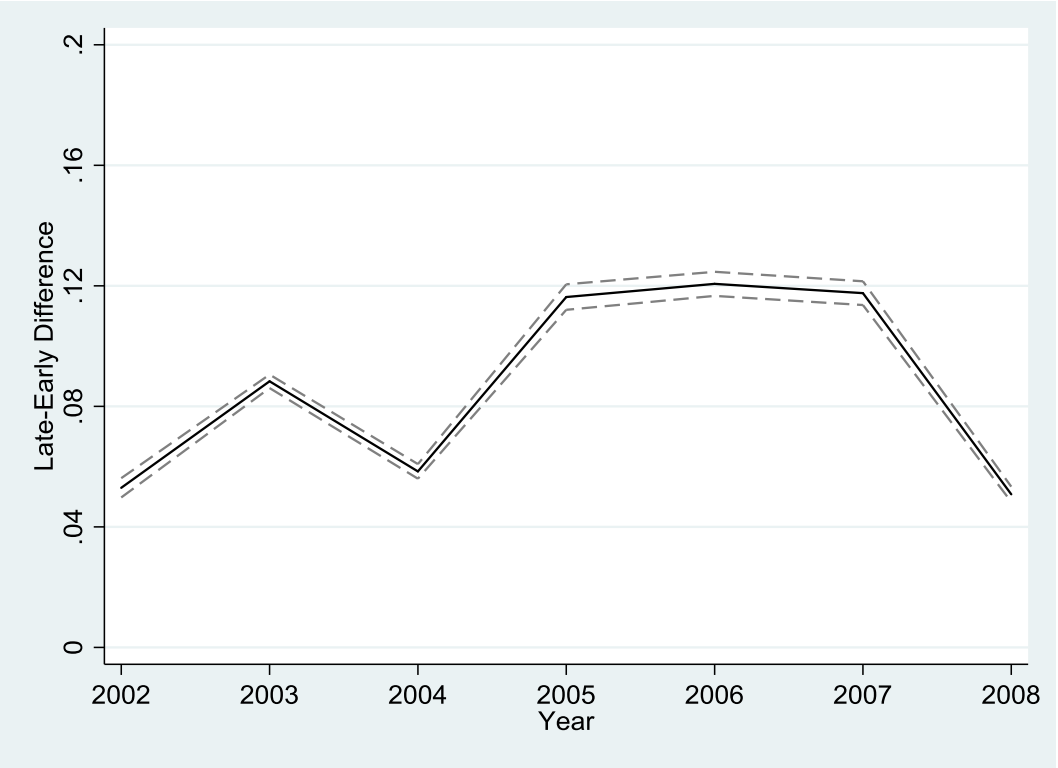
Rockoff, J. E., Staiger, D. O., Kane, T. J. and Taylor, E. S. (2012) Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. *American Economic Review*, 102: 7, pp. 3184-3213.

Sharma, D. (2010) The Impact of Financial Incentives on Academic Achievement and Household Behavior: Evidence from a Randomized Trial in Nepal *Working Paper*, The Ohio State University).

Stinebrickner, R. and Stinebrickner, T. R. (2008) The causal effect of studying on academic performance. *B E Journal of Economic Analysis & Policy*, 8: 1.

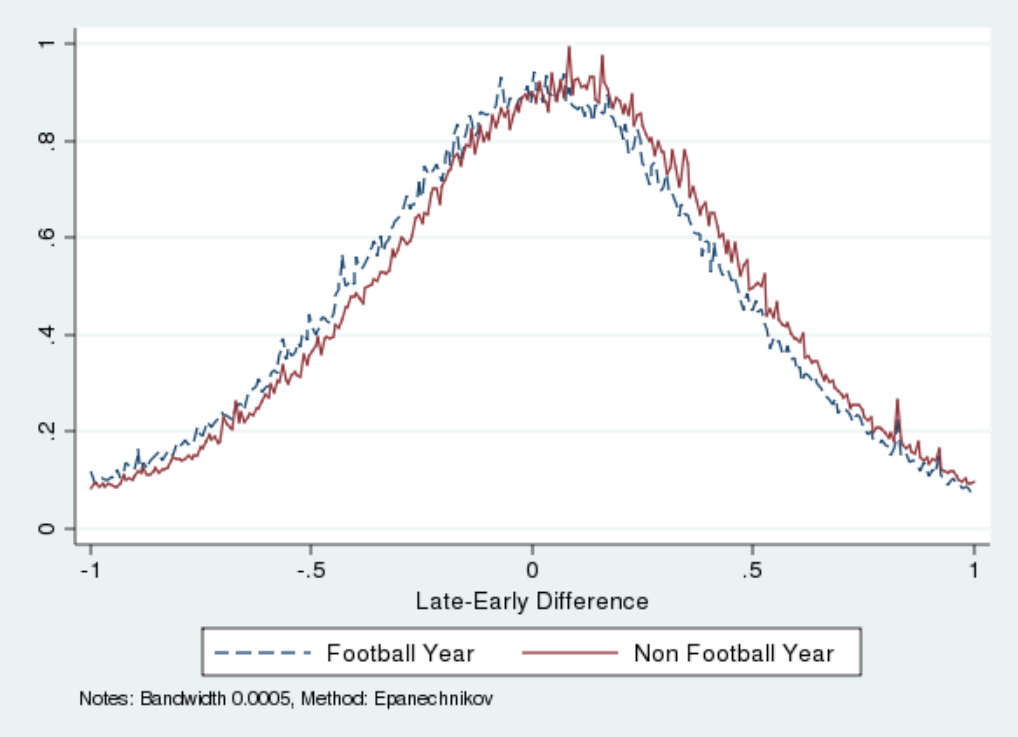
UEFA (2008) UEFA Euro 2008™ Review.

Figure 1: Illustrating the 'late-early' test score difference over tournament and non-tournament years.



We take the difference between the average GCSE score on 'late' subjects and the average GCSE score on 'early' subjects. We add 95% confidence intervals around the estimates of the difference.

Figure 2: Density functions for (late-early) subject score difference



Notes: An observation is a within-student difference between her score on all late subjects and her score on all early subjects. Late here is defined as having at least two thirds of the exams in the tournament period. All subject scores are normalised so the units are subject-level SDs. We distinguish tournament years from non-tournament years.

Table 1: Summary Statistics

	All students	Students with both "late" and "early" subjects
	%	%
Male	50.15	49.27
Eligible for FSM (Free School Meals)	12.05	11.03
Special Educational Needs (SEN), minor	13.48	11.40
Special Educational Needs (SEN), major	2.03	1.53
Selected ethnicities*		
White	84.64	84.05
Black Caribbean	1.34	1.38
Indian	2.33	2.47
Pakistani	2.28	2.37
Likely to be keen on football	4.09	3.47
Unlikely to be keen on football	2.21	2.38
Average Key stage 2 score	27.03	27.34
GCSE score, normalised	-0.041	0.014
Number of students	3,651,667	2,970,694
Total observations (subjects*students)	25,705,081	21,963,321

Seven years of data, 2002 – 2008 inclusive.
One cohort of students per year
Likely to be keen on football defined as male, FSM eligible, and white
Unlikely to be keen on football defined as female, not FSM eligible, and Asian
** Full set used in regressions.*

Table 2: Overall effect of the tournament

	(1)	(2)	(3)	(4)
Tournament year	0.883***	0.888***		
	(0.004)	(0.004)		
Tournament year*				
Likely to be keen on football		0.815***		
		(0.013)		
Unlikely to be keen on football		1.022		
		(0.023)		
Tournament year*				
Male, eligible for FSM			0.819***	0.816***
			(0.012)	(0.012)
Male, not eligible for FSM			0.862***	0.860***
			(0.005)	(0.005)
Female, eligible for FSM			0.888***	0.875***
			(0.012)	(0.013)
Female, not eligible for FSM			0.911***	0.903***
			(0.005)	(0.005)
Student characteristics	Y	Y	Y	Y
Time trend	Y	Y	Y	Y
Number of observations	3,651,594	3,651,594	3,651,594	3,060,545

An observation is a student; the dependent variable is a binary variable, equal to 1 if the student scored at least a C grade in at least 5 subjects. We report odds ratios.
Standard errors (e form) in parentheses; standard errors clustered at school level.
** significant at 10%; ** significant at 5%; *** significant at 1%*
Student characteristics included in all specifications are: gender, ethnicity, month of birth, poverty status, SEN status, English as additional language, prior ability measure (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score).
Also included in all specification is a time trend to account for gradually rising pass rate over this period.
Col (4) restricts the sample to students taking at least 7 GCSEs.

Table 3: Student fixed effect regression on subject grades

	(1)	(2)	(3)	(4)	(5)
Proportion of exams within subject scheduled "late"	0.103***	0.103***	0.126***	0.126***	0.126***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
(Proportion of exams within subject scheduled	-0.009***	-0.013***	-0.009***	-0.011***	-0.009***
"late") * (Year is a tournament year)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
(Proportion of exams within subject scheduled "late") * (Year is a tournament year) * (student characteristics): *(only selected interactions shown)*					
Male		-0.025***		-0.006***	
		(0.002)		(0.002)	
Eligible for FSM		-0.031***		-0.018***	
		(0.004)		(0.003)	
English as a first language		-0.016**		-0.000	
		(0.006)		(0.005)	
Middle ability		0.008***		-0.003	
		(0.003)		(0.002)	
High ability		-0.006**		0.014***	
		(0.003)		(0.002)	
Likely to be keen on football					-0.026***
					(0.006)
Unlikely to be keen on football					0.021***
					(0.005)
Student Characteristics	Y	Y			
Student fixed effects			Y	Y	Y
Number of observations	25,705,081	25,705,081	25,705,081	25,705,081	25,705,081
Number of students	3,651,667	3,651,667	3,651,667	3,651,667	3,651,667

*An observation is a student*subject; The dependent variable is the subject grade and the metric is subject-level SD*
*Standard errors in parentheses; standard errors clustered at student level. * significant at 10%; ** significant at 5%; *** significant at 1%*
"Late" is defined by calendar date for all years, coincides with the tournament dates in tournament years and mirrors those dates in non-tournament years.
Student characteristics are: gender, ethnicity, month of birth, poverty status, SEN status, English as additional language, prior ability measure (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score).
Year dummies are also included in cols (1) and (2). There is no variation within-student across years, so the specifications we include have student characteristics with year FE (in cols 1 and 2) and student FE, which implicitly mop up year effects (cols 3, 4 and 5).
Cols (2) and (4) also include the (proportion of exams late) interacted with gender, ethnicity, poverty status, SEN status, English as additional language, prior ability measure (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score). AND (proportion of exams late interacted) with (tournament year) interacted with: gender, ethnicity, poverty status, SEN status, English as additional language, prior ability measure (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score).
Col (5) includes the set of student characteristics above plus two binary composite variables interacted with (proportion of exams late) and with (proportion of exams late)(tournament year). The two composite variables are: Likely to be keen on football is equal to 1 if male, poverty status equals 1, and white ethnicity; unlikely to be keen on football is equal to 1 if female, poverty status equals 0, and Asian ethnicity.*

Table 4: Robustness tests

	Using TV viewership data: proportion of exams scheduled no more than 1 day before a game with 4 million+ viewers	Impact of a high weight on exams in subject grade	Omitting students who sit any subject early	Estimating sample 2002 – 2007 inclusive
	(1)	(2)	(3)	(4)
Proportion of exams within subject scheduled "late"	0.134***	0.214***	0.132***	0.139***
	0.001)	(0.001)	(0.001)	(0.001)
(Proportion of exams within subject scheduled "late") * (Year is a tournament year)		-0.017***	-0.016***	-0.023***
		(0.001)	(0.001)	(0.001)
Proportion of exams within 1 day of a game with 4 million+ viewers	-0.025***			
	(0.001)			
(Proportion of exams within subject scheduled "late") * (subject places high weight on exams)		-0.360***		
		(0.001)		
(Proportion of exams within subject scheduled "late") * (Year is a tournament year) * (subject places high weight on exams)		-0.008***		
		(0.001)		
Student Characteristics	No	No	No	No
Student fixed effects	Yes	Yes	Yes	Yes
Number of observations	25,705,081	21,251,767	20,519,618	22,154,179

*An observation is a student*subject; the dependent variable is the subject grade and the metric is subject-level SD*
*Standard errors in parentheses; standard errors clustered at student level. * significant at 10%; ** significant at 5%; *** significant at 1%*
Note that regressions in cols 2-4 have fewer observations than the main results in Table 3: in column 2 because we do not have exam weight data for all subjects; in column 3 because we omit all subjects for students who sit any subject early; and in column 4 because we omit year 2007.
The specifications and other variables included are otherwise the same as Table 3, column 4 which includes the (proportion of exams late) interacted with gender, ethnicity, poverty status, SEN status, English as additional language, prior ability measure (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score). AND (proportion of exams late interacted) with (tournament year) interacted with: gender, ethnicity, poverty status, SEN status, English as additional language, prior ability measure (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score).

Table 5: Student fixed effect regression results on passing subjects

	(1)	(2)	(3)	(4)	(5)
Proportion of exams within subject scheduled "late"	1.211***	1.242***	1.251***	1.303***	1.251***
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
(Proportion of exams within subject scheduled "late") * (Year is a tournament year)	0.839***	0.834***	0.819***	0.808***	0.819***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)
(Proportion of exams within subject scheduled "late") * (Year is a tournament year) * (student characteristics): *(only selected interactions shown)*					
Male		0.976***		1.025***	
		(0.006)		(0.010)	
Eligible for FSM		0.947***		0.902***	
		(0.010)		(0.0143)	
English as a first language		0.956**		0.981	
		(0.018)		(0.0272)	
Middle ability		1.131***		1.359***	
		(0.008)		(0.015)	
High ability		1.120***		1.335***	
		(0.009)		(0.017)	
Likely to be keen on football					0.921***
					(0.025)
Unlikely to be keen on football					1.016
					(0.034)
Student Characteristics	Y	Y			
Student fixed effects			Y	Y	Y
Number of observations	11,781,050	11,781,050	11,781,050	11,781,050	11,781,050
Number of students	1,811,378	1,811,378	1,811,378	1,811,378	1,811,378

*Odds ratios reported; An observation is a student*subject*
Note that there are fewer observations in this table as we have dropped students who either pass none or all of their subjects, and Science qualifications.
*Standard errors in parentheses; standard errors clustered at student level. * significant at 10%; ** significant at 5%; *** significant at 1%*
"Late" is defined by calendar date for all years, coincides with the tournament dates in tournament years and mirrors those dates in non-tournament years.
Student characteristics are: gender, ethnicity, month of birth, poverty status, SEN status, English as additional language, prior ability measures (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score).
Year dummies are also included in cols (1) and (2). There is no variation within-student across years, so the specifications we include have student characteristics with year FE (in cols 1 and 2) and student FE, which implicitly mop up year effects (cols 3, 4 and 5).
Cols (2) and (4) also include the (proportion of exams late) interacted with gender, ethnicity, poverty status, SEN status, English as additional language, prior ability measure (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score) AND (proportion of exams late interacted) with (tournament year) interacted with: gender, ethnicity, poverty status, SEN status, English as additional language, prior ability measure (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score).
Col (5) includes the set of student characteristics above plus two binary composite variables interacted with (proportion of exams late) and with (proportion of exams late)(tournament year).*
The two composite variables are: Likely to be keen on football is equal to 1 if male, poverty status equals 1, and white ethnicity; unlikely to be keen on football is equal to 1 if female, poverty status equals 0, and Asian ethnicity.

Appendix Figures and Tables

Appendix Table 1: Dates and tournaments

			Tournament			Exams		
Year	**Tournament year?**	**Host**	**Did England qualify?**	**Start date**	**End date**	**Examination start date**	**Examination end date**	**% of exams during football**
2002	World Cup	South Korea and Japan	Yes	31st May	30th June	13th May	28th June	61%
2003	-	-	-	-	-	12th May	27th June	-
2004	European Championship	Portugal	Yes	12th June	4th July	17th May	30th June	49%
2005	-	-	-	-	-	16th May	30th June	-
2006	World Cup	Germany	Yes	9th June	9th July	15th May	28th June	48%
2007	-	-	-	-	-	14th May	27th June	-
2008	European Championship	Austria and Switzerland	No	7th June	29th June	13th May	25th June	46%

Appendix Table 2: Football programmes in the top 10 most viewed programmes

Channel and programme rank for that week		Week ending and programme	Viewers (millions)
2002			
BBC1		w/e 9th June 2002	
	2	WORLD CUP 2002: ARGENTINA V ENGLAND (FRI 1230)	12
	5	WORLD CUP 2002: POST-MATCH (FRI 1420)	10.49
BBC1		w/e 16th June 2002	
	2	WORLD CUP 2002: ENGLAND V DENMARK (SAT 1230)	12.47
	4	WORLD CUP 2002: ENGLAND V NIGERIA (WED 0730)	12.22
	7	WORLD CUP 2002: POSTMATCH (SAT 1420)	8.85
	9	WORLD CUP 2002: SPAIN V IRELAND (SUN 1230)	7.77
BBC1		w/e 23rd June 2002	
	1	WORLD CUP 2002: ENGLAND V BRAZIL (FRI 0730)	12.46
	6	WORLD CUP 2002: POST-MATCH (FRI 0920)	9.77
BBC1		w/e 30th June 2002	
	4	WORLD CUP 2002: GERMANY V BRAZIL (SUN 1200)	10.08
	7	WORLD CUP 2002: POST MATCH (SUN 1350)	8.95
2004			
BBC1		w/e 13th June 2004	
	7	EURO 2004: SPA V RUS (SAT 1945)	6.4
	8	EURO 2004: PORT V GRC (SAT 1700)	6.19
ITV1		13th June 2004	
	1	EURO 2004 FRA V ENG (SUN 1944)	17.8
BBC1		w/e 20th June 2004	
	4	EURO 2004: SPA V PORT (SUN 1945)	8.78
	5	EURO 2004: GER V NETH (TUE 1945)	7.95
	6	EURO 2004: CRO V FRA (THU 1946)	7.55
	7	EURO 2004: POST-MATCH (THU 2135)	7.23
	8	EURO 2004: POST-MATCH (SUN 2135)	6.85
	9	EURO 2004: NETH V CZECH (SAT 1945)	6.74
ITV1		w/e 20th June 2004	
	1	EURO 2004 ENG V SWI (THU 1659)	14.31
BBC1		w/e 27th June 2004	
	1	EURO 2004: POR V ENG (THU 1945)	20.66
	2	EURO 2004: CRO V ENG (MON 1945)	18.28
	3	EURO 2004: POST-MATCH (MON 2136)	14.48
	4	EURO 2004: POST-MATCH (THU 2229)	14.22
	5	EURO 2004: PREMATCH (THU 1929)	11.71
	7	EURO 2004: PREMATCH (MON 1929)	9.83
ITV1		w/e 27th June 2004	
	4	EURO 2004 GER V CZE (WED 1944)	8.28
	9	EURO 2004 SWE V NETH (SAT 1945)	7.04
2006			
BBC1		w/e 4th June 2006	
	1	MATCH OF THE DAY LIVE (TUE 1958)	9.29
BBC1		w/e 11th June 2006	
	1	WORLD CUP 2006: ENG V PAR (SAT 1400)	12
	2	WORLD CUP 2006: POST-MATCH (SAT 1551)	9.29
	10	WORLD CUP 2006: GER V CRI (FRI 1701)	5.65
BBC1		w/e 18th June 2006	
	1	WORLD CUP 2006: BRA V CRO (TUE 2000)	9.64
	2	WORLD CUP 2006: GER V POL (WED 2000)	8.11
	4	WORLD CUP 2006: POST-MATCH (WED 2149)	6.74

	5	WORLD CUP 2006: ITA V GHA (MON 2000)	6.69
	8	WORLD CUP 2006: POST-MATCH (SUN 2151)	6.39
	9	WORLD CUP 2006: POST-MATCH (TUE 2151)	6.38
	10	WORLD CUP 2006: FRA V KOR (SUN 2000)	6.17
ITV1		w/e 18th June 2006	
	1	WORLD CUP 06: ENG V TRI (THU 1650)	13.67
	5	WORLD CUP 06: BRA V AUS (SUN 1658)	8.08
	10	WORLD CUP 06: SWE V PAR (THU 1959)	6.63
BBC1		w/e 25th June 2006	
	1	WORLD CUP 2006: ENG V ECU (SUN 1600)	16.29
	2	WORLD CUP 2006: POST-MATCH (SUN 1750)	13.45
	3	WORLD CUP 2006: ARG V MEX (SAT 2000)	8.46
	4	WORLD CUP 2006: JAP V BRA (THU 2000)	7.81
	10	WORLD CUP 2006: PREMATCH (SUN 1529)	7.44
ITV1		w/e 25th June 2006	
	1	WORLD CUP 06 (TUE 1950)	18.46
	3	WORLD CUP 06 (WED 1958)	8.74
	7	WORLD CUP 06 (SUN 1958)	7.43
	9	WORLD CUP 06: PREMATCH (TUE 1903)	6.7
2008			
BBC1		w/e 15th June 2008	
	10	EURO 2008: MATCH OF THE DAY LIVE (FRI 1929)	5.58
ITV1		w/e 15th June 2008	
	10	EURO 2008 LIVE (MON 1929)	5.74
BBC1		w/e 22nd June 2008	
	4	EURO 2008: MATCH OF THE DAY LIVE (SUN 1930)	7.21
	5	EURO 2008: MATCH OF THE DAY LIVE (TUE 1929)	6.29
	7	EURO 2008: MATCH OF THE DAY LIVE (FRI 1929)	5.64
ITV1		w/e 22nd June 2008	
	3	EURO 2008 LIVE (SAT 1929)	7.37
	5	EURO 2008 LIVE (THU 1929)	6.89
BBC1		w/e 29th June 2008	
	1	EURO 2008: MATCH OF THE DAY LIVE (SUN 1856)	8.84
	6	EURO 2008: MATCH OF THE DAY LIVE (WED 1929)	6.95
ITV1		w/e 29th June 2008	
	6	EURO 2008 LIVE (THU 1929)	6.77

We use weekly data on the viewing figures of the top 30 programmes per channel from the Broadcasters' Audience Research Board (www.barb.co.uk).

Appendix Table 3: Exam-tournament overlap and exam weighting by subject, 2002, 2004, 2006 and 2008

2002

Subject	Proportion of examinations *late*	Number of days with examinations scheduled	Number of days with examinations scheduled during tournament	Number of students
Core subjects				
Chemistry	1.00	2	2	39,079
Physics	1.00	2	2	38,626
Double award science	0.64	4	3	466,410
Single award science	0.64	4	3	51,581
English Language	0.50	2	1	561,092
Biology	0.40	2	1	39,923
Mathematics	0.40	2	1	572,510
English Lit	0.00	1	0	501,784
Optional subjects				
Classical civilisation	1.00	2	2	3,772
Economics	1.00	2	2	4,292
Modern Greek	1.00	4	4	589
History	1.00	2	2	197,531
ICT	1.00	2	2	148,730
Media Studies	1.00	1	1	26,738
Social Science	1.00	2	2	2,109
Sociology	1.00	2	2	11,837
Business Studies	0.80	3	2	85,118
Religious Education	0.57	6	3	194,248
Religious Studies	0.57	7	4	103,839
D&T	0.50	2	1	3,857
D&T Electronics	0.50	2	1	21,304
D&T Food	0.50	2	1	108,631
D&T Graphics	0.50	2	1	114,748
D&T Resistant Materials	0.50	2	1	121,355
D&T Textiles Technology	0.50	2	1	49,625
Psychology	0.50	2	1	681
French	0.25	3	1	321,636
Latin	0.17	4	1	9,462
D&T Systems and Control Technology	0.00	2	0	14,259
Drama	0.00	1	0	94,049
Geography	0.00	2	0	211,250
German	0.00	3	0	126,974
Humanities	0.00	2	0	20,376
Music	0.00	1	0	44,267
PE	0.00	1	0	116,005
Persian	0.00	2	0	261
Portuguese	0.00	2	0	561
Rural Science	0.00	1	0	809

Spanish	0.00	3	0	50,179
Turkish	0.00	3	0	1,173

2004

Subject	Proportion of examinations *late*	Number of days with examinations scheduled	Number of days with examinations scheduled during tournament	Number of students
Core subjects				
Chemistry	1.00	2	2	43,253
Physics	1.00	2	2	42,570
Biology	0.67	2	1	44,029
English Language	0.67	3	2	599,168
Science Double	0.64	5	3	480,236
Single award Science	0.58	4	2	55,758
Mathematics	0.40	2	1	609,499
English Lit	0.33	2	1	530,915
Optional subjects				
Classical Civilisation	1.00	2	2	4,064
Classical Greek	1.00	3	3	863
Economics	1.00	2	2	3,210
History	1.00	2	2	208,992
Media Studies	1.00	2	2	35,888
Psychology	1.00	2	2	1,346
Sociology	1.00	2	2	13,813
Business Studies	0.71	3	2	80,836
Religious Studies	0.53	9	5	362,775
D&T Electronics	0.50	2	1	19,069
D&T Food	0.50	2	1	107,596
D&T Graphics	0.50	2	1	105,897
D&T Resistant Materials	0.50	2	1	111,427
D&T Textiles Technology	0.50	2	1	54,149
French	0.33	3	1	298,185
Citizenship studies	0.00	1	0	26,565
D&T Systems and Control technology	0.00	2	0	13,673
Drama	0.00	1	0	93,889
Geography	0.00	2	0	199,790
German	0.00	3	0	118,862
Humanities	0.00	2	0	17,347
ICT	0.00	3	0	182,285
Latin	0.00	3	0	9,252
Music	0.00	1	0	51,209
PE	0.00	1	0	140,963
Persian	0.00	3	0	354
Portuguese	0.00	3	0	771
Spanish	0.00	3	0	55,721
Turkish	0.00	3	0	1,214

2006

Subject	Proportion of examinations *late*	Number of days with examinations scheduled	Number of days with examinations scheduled during tournament	Number of students
Core subjects				
Chemistry	1.00	2	2	49,485
Physics	1.00	2	2	49,064
Biology	0.67	2	1	51,300
Science Double	0.64	5	3	444,161
Single award Science	0.58	4	2	72,270
Mathematics	0.40	2	1	620,155
English Language	0.33	3	1	608,741
English Lit	0.00	2	0	528,495
Optional subjects				
Classical Civilisation	1.00	2	2	4,429
Classical Greek	1.00	3	3	964
Drama	1.00	1	1	94,112
Economics	1.00	2	2	2,663
Media Studies	1.00	2	2	52,738
Psychology	1.00	2	2	2,860
Sociology	1.00	2	2	16,396
Business Studies	0.71	3	2	77,542
Religious Studies	0.67	9	7	397,847
D&T Electronics	0.50	2	1	15,681
D&T Food	0.50	2	1	86,319
D&T Graphics	0.50	2	1	79,792
D&T Resistant Materials	0.50	2	1	93,830
D&T Textiles Technology	0.50	2	1	46,257
Geography	0.50	2	1	188,576
History	0.50	2	1	210,739
French	0.33	3	1	214,878
Citizenship studies	0.00	1	0	52,423
D&T Systems and Control technology	0.00	2	0	10,624
German	0.00	3	0	86,335
Humanities	0.00	2	0	15,364
ICT	0.00	2	0	192,595
Latin	0.00	3	0	9,602
Music	0.00	1	0	54,649
PE	0.00	1	0	164,098
Persian	0.00	3	0	390
Portuguese	0.00	3	0	899
Spanish	0.00	3	0	53,515
Turkish	0.00	3	0	1,188

2008

Subject	Proportion of examinations *late*	Number of days with examinations scheduled	Number of days with examinations scheduled during tournament	Number of students
Core subjects				
Physics	1.00	3	3	67,516
Biology	0.67	3	2	73,971
Science Double	0.67	3	2	455,739
Mathematics	0.50	4	2	586,988
Chemistry	0.33	3	1	68,457
English Language	0.00	2	0	580,845
English Lit	0.00	1	0	503,348
Optional subjects				
D&T Systems and Control technology	1.00	1	1	7,321
D&T Textiles Technology	1.00	1	1	40,975
Drama	1.00	1	1	92,098
Media Studies	1.00	1	1	59,576
Sociology	1.00	1	1	17,069
Business Studies	0.80	3	2	75,016
D&T Food	0.50	2	1	72,108
D&T Graphics	0.50	2	1	60,725
Geography	0.50	2	1	177,932
History	0.50	2	1	204,830
French	0.33	3	1	165,532
Latin	0.33	3	1	8,579
RS	0.25	3	1	371,901
Citizenship	0.00	1	0	73,867
D&T Resistant Materials	0.00	1	0	78,726
German	0.00	1	0	67,986
ICT	0.00	2	0	129,213
Music	0.00	1	0	52,160
PE	0.00	1	0	153,984
Spanish	0.00	2	0	55,127

Proportion of assessment via examination

Subject	%	Subject	%	Subject	%
English Language	60	English Literature	70	Maths	65
Science	50	Child Development	50	IT	40 – 50
Languages	50	Geography	75	History	75
Design & technology	40	Law	80	Media Studies	50
Music	25	Physical Education (gym)	50	Religious Education	75 – 100

The fraction of the mark from the final summer exams varies by course. Information from exam board sources.

Appendix Table 4: Different clustering approaches

Panel A: Impact of different choices of clustering

	Student (*year)	School	School*year	Subject*Year
Table 3, col. 1 specification				
Proportion of exams within subject scheduled "late"	0.103*** (0.001)	0.103*** (0.003)	0.103*** (0.002)	0.103*** (0.036)
(Proportion of exams within subject scheduled "late") * (Year is a tournament year)	-0.009*** (0.001)	-0.009*** (0.003)	-0.009** (0.004)	-0.009 (0.059)
Number of observations	25,705,081	25,705,081	25,705,081	25,705,081

All the columns replicate the analysis of Table 3, column 1, and simply vary the basis of the residual clustering
*An observation is a student*subject; The dependent variable is the subject grade;the metric is subject-level SD*
*Standard errors in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%*
"Late" is defined by calendar date for all years, coincides with the tournament dates in tournament years and mirrors those dates in non-tournament years.
Student characteristics are: gender, ethnicity, month of birth, poverty status, SEN status, English as additional language, prior ability measure (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score). Year dummies are also included.

To further support the use of clustering at student level, we take the following approach. We take the fitted residuals from the specification

$$KS4_{ist} = \beta_0 + \beta_1 X_{it} + \beta_2 \text{prop late}_{st} + \beta_3 \text{prop late}_{st} * \text{football}_t + u_{ist}$$

where X includes our standard individual controls. We regress the fitted residuals against different sets of dummy variables; the higher the R-squared, the higher the correlation between the residuals and the groups, indicating stronger correlation within the units defined by those groups.

Panel B

Clustering group	**Number of groups**	**Adjusted R-squared**
Subject-year	161	0.0195
School ID	3,286	0.0545
Student (year)	3,651,667	0.4631

The results suggest that error correlation is far higher within students than it is within subject-year units, and so that is the major issue to deal with through clustering.

Appendix Table 5: Quantiles of differences-in-differences for matched school-groups

	p5	p10	p25	p50	p75	p90	p95
All Students	-0.3307	-0.2570	-0.1531	-0.0486	0.0577	0.1571	0.2150
Male	-0.3571	-0.2846	-0.1756	-0.0628	0.0489	0.1531	0.2134
FSM	-0.4215	-0.3556	-0.2247	-0.0339	0.1017	0.2251	0.2546
Low ability	-0.4006	-0.3170	-0.1854	-0.0495	0.0814	0.1965	0.2632
Middle ability	-0.3380	-0.2711	-0.1579	-0.0457	0.0731	0.1792	0.2407
High ability	-0.2987	-0.2374	-0.1444	-0.0491	0.0477	0.1364	0.1912

We match students within schools using student gender, FSM status, prior achievement group, ethnic group and quarter of birth. Each student in a tournament year is matched with a student in a non-tournament year in the same school and defined by the same set of observables. For all students, there are 14,940 school-groups. School-groups are only included if there are at least 20 students within the school-group in both tournament and non-tournament years.

*We take the mean (late subject score – early subject score) difference for each student, and then average this within each school*observables group, separately for tournament and non-tournament years, and analyse the difference. Quantiles of the distribution of the following statistic are reported*

$\Delta_{sg} = \overline{(y_{isg,late} - y_{isg,early})}_{tournament} - \overline{(y_{isg,late} - y_{isg,early})}_{non-tournament}$

Metric is subject-level SD

Appendix Table 6: Student fixed effect regression results on passes, C/D marginal students

	(1)	(2)	(3)	(4)
Proportion of exams within subject scheduled "late"	1.219***	1.379***	1.345***	1.498*
	(0.007)	(0.009)	(0.009)	(0.037)
Proportion of exams within subject scheduled "late" * Year is a tournament year	0.880***	0.776***	0.864***	0.740***
	(0.007)	(0.007)	(0.009)	(0.009)
Proportion of exams within subject scheduled "late" * Year is a tournament year *:				
Male		1.159***		1.585***
		(0.020)		(0.038)
Eligible for FSM		0.656***		1.089
		(0.078)		(0.160)
English as a first language		0.880***		0.752***
		(0.031)		(0.036)
Middle ability		1.032		0.993
		(0.028)		(0.038)
High ability		0.895***		0.835***
		(0.026)		(0.032)
Student Characteristics	Y	Y		
Student fixed effects			Y	Y
Number of observations	4,854,411	4,854,411	2,204,623	2,204,623
Number of students	732,032	732,032	330,853	330,853

Odds ratios reported
*An observation is a student*subject; note that there are fewer observations in this table as we have dropped Science qualifications.*
Dependent variable is binary, equal to 1 if the subject grade is at least C, estimated by linear probability model.
*Standard errors in parentheses; standard errors clustered at student level. * significant at 10%; ** significant at 5%; *** significant at 1%*
Proportion of exams within subject scheduled "late", where "late" is defined by calendar date for all years, coincides with the tournament dates in tournament years and mirrors those dates in non-tournament years.
Student characteristics are: gender, ethnicity, month of birth, poverty status, SEN status, English as additional language, prior ability measures (based on Key stage 2 English score, Key stage 2 maths score, Key stage 2 Science score). Year dummies are also included in cols (1) and (3).
In cols (3) and (4), sample size drops because some students have only all Ds, or all Cs and so are dropped when we introduce student fixed effects.