Centre for Market and
Public Organisation

Raising student attainment:

- School inspections
- Student ‘effort’ and engagement

Simon Burgess

The Leverhulme Trust

Introduction

The questions we always start with:

- How to improve educational attainment?
- And reduce educational inequalities?

Why education?

- Human capital, cognitive and non-cognitive skills, growth and prosperity
- Earnings, inequality, social mobility, equality of opportunity, "escape"
- Personal fulfilment, realising potential, …

- Schools are arguably the most important stage of education for policy purposes. Universities, pre-school, families, …

Two policy areas

1. School Inspections

- Is there a role for school accountability?
- Schools are entrusted with two resources:
  - The talent and potential of the nation's children
  - Public money (schools spend over £30bn a year)
- Schools should be accountable for what they do with these.

Two policy areas

2. Student ‘effort’ and engagement

- Class size, teacher quality, school resources, peer groups, family income, …
- Less attention on student effort – really the only thing under the student’s control.
- Does studying hard pay off?
- We quantify how much student effort matters to educational attainment

Background: School system in England

- State-funded schools = 93% students
- About 550,000 students per cohort
- Compulsory schooling from age 5 – 16
- Primary Education, to age 11, compulsory secondary education to age 16.
- National Curriculum, four Keystages
- Keystage 2 exams at age 11, Keystage 4 exams at age 16 (also called GCSEs)
- GCSEs are high stakes exams for students – access to higher education and to jobs – and for schools.

Centre for Market and Public Organisation

How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England

Rebecca Allen (IOE) and Simon Burgess (CMPO)

- Accountability is not straightforward because school and teacher ‘effort’ cannot be perfectly measured.
- Level of effort
- Focus of effort:
  - wider learning versus qualifications
  - professional independence
  - “closed doors”
- Standard approach – provide incentives for agent to achieve principal’s aim
- Incentives often deemed infeasible in schools, so:
  - Provision of information on outcomes, providing indirect and non-monetary incentives.
  - Inspection to acquire detailed information

School inspections in England

- OFSTED – Office for Standards in Education
- Inspects schools, nursery provision, children and family services …
- Large organisation:
  - spends about £200m each year.
  - Reports directly to Parliament
  - Is independent of Department for Education
- School inspection regime is data driven, not all schools are inspected equally often.

Dealing with the outcomes of inspection

- What is the best policy for dealing with schools judged to be under-performing?
- What happens to schools that are judged unsatisfactory by Ofsted (between 2002 and 2009)?
- In principle, the effects of failing an Ofsted inspection could go either way:
  - inducement to focus on academic performance
  - spiral of decline

The policy treatment

- We compare those who 'just' fail and are given a *notice to improve* with those judged as satisfactory

> *"the school requires significant improvement because either: it is failing to provide an acceptable standard of education, but is demonstrating the capacity to improve; or it is not failing to provide an acceptable standard of education but is performing significantly less well in all the circumstances reasonably be expected to perform"*
>
> (Ofsted, 2011a, page 12)

The policy treatment

- ‘Light-touch’ judgement, although publicly humiliating?
- No operating restrictions
- Monitoring inspection within the year and full inspection after a year
- Opportunity to attend a school improvement seminar
- Expected to amend school plans

Identification problem

- We aim to estimate the causal impact of being judged by Ofsted as unsatisfactory on school performance
- Endogeneity of failure: underperforming schools have different levels and trajectories of achievement, regardless of inspections
- Estimation approach: regression discontinuity design (RDD) in a panel data context, comparing the performance for schools that are designated as just failing with those just passing
- Intuition is that schools around the failure threshold are very similar, except for random measurement of quality by inspectors
- A running variable based on sub-criteria judgements captures continuous variation between schools, on top of which is the discontinuity of a discrete judgement of ‘fail’ or ‘pass’

Ofsted inspections data

Year	02/03	03/04	04/05	05/06	06/07	07/08	08/09
Number of school visits	476	560	452	926	1,106	971	638
Number of sub-criteria	19	33	33	55	41	58	65
Rating = Excellent	18	10	13	n/a	n/a	n/a	n/a
Outstanding/Very good	117	98	109	98	165	180	151
Good	202	264	190	358	438	417	283
Satisfactory	114	130	107	348	415	300	167
Unsatisfactory	18	44	24	122	88	74	37
Poor	5	14	9	n/a	n/a	n/a	n/a
Very poor	2	0	0	n/a	n/a	n/a	n/a
Proportion failing (%)	5.3	10.4	7.3	13.2	8.0	7.6	5.8

Selecting ‘just’ passers and ‘just’ fails

From multiple sub-criteria to a continuous, uni-dimensional measure of failure

Role of rating variable:

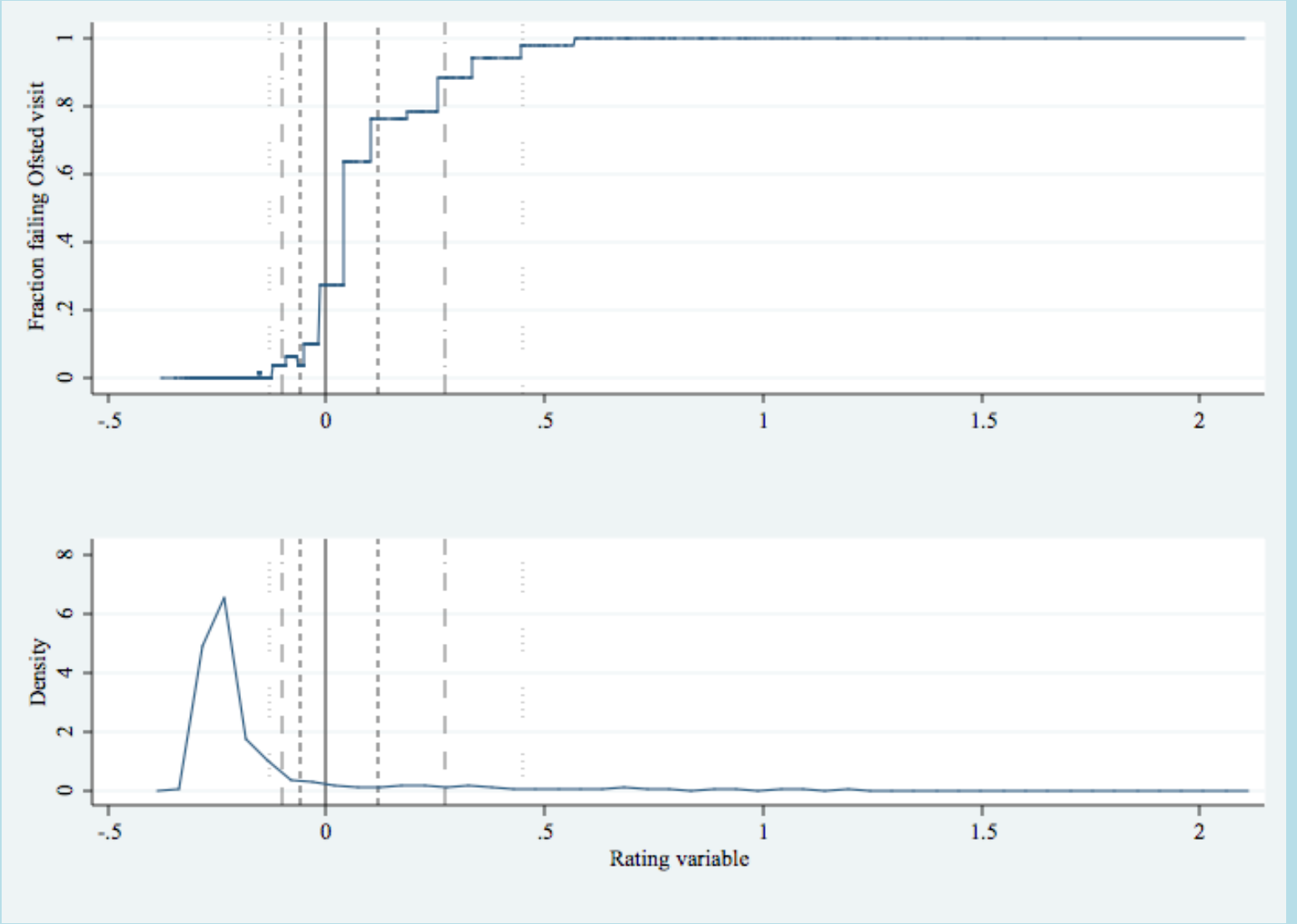
- Divides schools into those that actually passed and failed reasonably well
- Has enough variation to distinguish between bad fails and very bad fails

Our rating variable is prediction from the sub-criteria:

$$fail_s = \beta_0 + \beta_1 * \%fail_s + \beta_2 * \%satisfactories_s + \varepsilon_s$$

(centred around zero)

Rating variable and bandwidth



National Pupil Database (02-11)

- National Pupil Database from 2002 onwards, aggregated to school-level variables
  - The **achievement** of the year 11 cohorts is measured using:
    - 'Capped GCSE' = average score across all pupils in their best 8 subjects at GCSE, standardised across all pupils as a z-score
    - '%5AC GCSE' = proportion of pupils achieving five or more 'good' GCSEs at grades A*-C
    - average school grades in English and maths measured on a scale of 0 (=U) to 8 (=A*)
  - **Control variables** include free school meals, ethnicity, gender, English mother tongue proportions and average deprivation and prior attainment for cohort

Fuzzy regression discontinuity

Change in school GCSE outcomes at t+1, t+2, t+3, t+4 *minus* t-1

Level and change in control variables:
-Prior attainment x 3
-FSM and deprivation
-EAL, ethnicity, female

Inspection year dummies

$$\Delta_\tau Y_s = \alpha + \beta\, fail_s + \gamma \Delta_\tau X_s + \pi . X_{st-1} + \lambda . (Y_{st-1} - Y_{st-3}) + \delta . inspyear_s + \varepsilon_s$$

Instrumented using threshold dummy (rating>0) and quadratic of rating variable on each side of threshold

Prior trend in GCSE outcome variable

Fuzzy RDD results

Difference in GCSE:	**(t+1) – (t-1)**	**(t+2) – (t-1)**	**(t+3) – (t-1)**	**(t+4) – (t-1)**
All observations	0.057***	0.104***	0.112***	0.123***
	(0.012)	(0.015)	(0.016)	(0.022)
N	4004	3966	3313	2359
Broad bandwidth	0.043*	0.069**	0.092***	0.135***
	(0.022)	(0.027)	(0.031)	(0.043)
N	467	466	421	325
Narrow bandwidth	**0.046**	**0.102*** **	**0.121** **	**0.140** **
	(0.032)	**(0.036)**	**(0.044)**	**(0.055)**
N	**315**	**314**	**283**	**232**
V narrow bandwidth	0.022	0.021	0.135	0.082
	(0.061)	(0.067)	(0.084)	(0.100)
N	156	156	139	119

How is the change achieved?

- Do schools:
  - simply try to raise teaching effectiveness, or
  - Game by introducing a lot of GCSE-equivalents?
- Do schools focus:
  - Just on marginal pupils, or
  - All pupils?

Different outcome variables

Difference in GCSE:	**(t+1) – (t-1)**	**(t+2) – (t-1)**	**(t+3) – (t-1)**	**(t+4) – (t-1)**
Capped mean GCSE score	0.046	0.102***	0.121**	0.140**
	(0.032)	(0.036)	(0.044)	(0.055)
N	315	314	283	232
Fraction with least 5 A*-C GCSE	0.024	0.037*	0.050**	0.058**
	(0.019)	(0.021)	(0.025)	(0.029)
N	315	314	283	232
Mean English GCSE score	0.139**	0.164**	0.141*	0.094
	(0.062)	(0.068)	(0.079)	(0.082)
N	315	314	283	232
Mean Maths GCSE score	0.146**	0.114*	0.106	0.074
	(0.059)	(0.064)	(0.074)	(0.081)
N	315	314	283	232

Marginal pupils versus others

Difference in GCSE:	(t+1) – (t-1)	(t+2) – (t-1)	(t+3) – (t-1)	(t+4) – (t-1)
Lower ability students	0.010	0.075*	0.117**	0.118*
	(0.037)	(0.045)	(0.048)	(0.062)
Marginal students	0.082*	0.093*	0.113**	0.157**
	(0.044)	(0.048)	(0.053)	(0.064)
Higher ability students	0.085*	0.106**	0.095*	0.216***
	(0.044)	(0.050)	(0.055)	(0.069)

Conclusions

- Findings
  - Schools failing their Ofsted inspections improve their subsequent performance relative to the score in the pre-visit year
  - The magnitudes are quantitatively very significant: around 0.10σ
  - The main impact arises two years after the visit in this data
  - Effects are consistent across individual subjects
- Why do they improve?
  - Threat from re-inspection?
  - Information about relative performance?
  - Public stigma of failure?
- Policy implications
  - Cost compared to alternatives
  - What to do with merely satisfactory schools?

Centre for Market and
Public Organisation

Student effort and educational attainment: Using the England football team to identify the education production function

Robert Metcalfe (Oxford), Simon Burgess (CMPO), Steve Proud (CMPO)

What we do:

- We use a sharp, exogenous and repeated change in the value of leisure to identify the impact of student effort on educational performance
- The treatment arises from the fact that the world's major international football tournaments overlap with the exam period in schools in England, a nation obsessed with football
- Performance is measured using the universal high-stakes tests that students in schools in England take at the end of compulsory schooling

Motivations

- Education production function
  - Is effort important for attainment?
  - Impact of substantial decline in effort on exams 0.2σ
  - Implications for incentives and schools policies
  - Implications for interpretation of other results
- Local policy issue
  - Bring forward summer exams a few weeks?
  - Raise average attainment (by 0.02σ) and reduce inequality (raise by 0.03σ for poor, male students)
  - *(cf. Impact of "Literacy Hour" was 0.06σ)*
  - Transitional costs. Anything else?

Identification strategy (1)

- Treatment is well-suited to a causal study
- Exposure to the treatment is random: whether a particular student is born in an even year or an odd year
- Neither students nor schools can affect the timing of the exams, scheduled for the same weeks each year
- The maximum potential treatment is very strong
  - The competition always completely dominates TV, radio and other media during the weeks it takes place
- Actual treatment depends on an individual's interest in football – expect to model considerable heterogeneity.

Identification strategy (2)

- The key high-stakes examinations in England (GCSE) are taken at the end of secondary schooling (at age 16), and are always scheduled for May and June
- We obtained data on exam timetables for each subject, and compare with the tournament dates
- A proportion of exams overlap with these major football tournaments, and this generates within-student variation in tournament years.

Identification strategy (3)

- Tournaments occur every other summer, so each year is sequentially either a treatment year or a control year
- We can implement a clean difference-in-difference design:
  - We 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
- Address whether there is differential selection away from late exams in tournament years.

Outline

- Modelling framework
- Data
  - Timing of tournaments and exams
  - Student data
- Results
  - Aggregate
  - Differences
  - Differences-in-differences
  - Robustness
  - Quantifying the effects
- Conclusions

Model of student effort

- Attainment depends on effort and ability
- Attainment is valued because of higher lifetime income
- Students exert effort when revising for exams, which has a cost of lost leisure time

Model 2

- **Cost of effort, ie. value of time**:
  - Major cost is value of the leisure time forgone
  - Will depend on observable and unobservable individual characteristics
  - Key factor is that value of leisure increases for some individuals with a major football tournament.
  - Allow the impact of the tournament on the value of leisure to vary by individual, ϕ_i.
  - Distinguish pre-tournament and in-tournament
- **Valuation of attainment, ie. rate of return:**
  - Will also depend on observable and unobservable individual characteristics

Model 3

- **Attainment, ie. converting effort into grades:**
  - Attainment technology will vary by observable and unobservable student characteristics, and possibly by school
  - Allow for the possibility that the exam setting and marking may vary year-by-year by including year effects, *t*.
  - Allow student performance to vary through the exam period. Many possibilities …
  - In any case, we allow for unrestricted, idiosyncratic within-period time dummies, *m*. That is, this pattern can vary individual by individual.

Model to estimate:

$$q_{itm} = \beta_0 + \beta_1 a_i + \beta_2 Z_i + v_i + \eta_{itm} +$$
$$\sum_\tau \alpha_\tau I(t=\tau) + \sum_n \pi_{in} I(m=n) +$$
$$f\left(\phi_i^0, a_i\right).\{I(t=T)\} + f\left(\phi_i^1, a_i\right).\{I(t=T).I(m=T)\}$$

- Individual factors, observed and unobserved, error term
- Year dummies, date-of-exam dummies
- Impact of year of tournament, impact of month of tourname
- Allow impact to depend on taste for football and ability

Data

- Every four years (on even years) the FIFA World Cup takes place in June and July
  - Eg. 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
- 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.
  - Eg. 2008 Euro tournament was watched live by at least 155 million TV viewers, and the final round of the tournament was shown in a total of 231 countries.

Data – timings of football and exams

Football tournaments 2002-2008

Year	Host country	Tournament	Did England qualify?	Start date	End date
2002	South Korea and Japan	World Cup	Yes	31st May	30th June
2004	Portugal	European championships	Yes	12th June	4th July
2006	Germany	World Cup	Yes	9th June	9th July
2008	Austria and Switzerland	European championships	No	7th June	29th June

Data – timings of football and exams

Examination dates from 2002-2008

Year	'Football' year	Examination start date	Examination end date	% of exams during football
2002	Yes	13th May	28th June	61%
2003	No	12th May	27th June	-
2004	Yes	17th May	30th June	49%
2005	No	16th May	30th June	-
2006	Yes	15th May	28th June	48%
2007	No	14th May	27th June	-
2008	Yes	13th May	25th June	46%

Timing data from Cambridge Examinations.
The exams of different boards for the same subject across the country are on the same day.

Pupil Data

- Administrative data
- National Pupil Database (NPD), Pupil Level Annual Schools Census (PLASC)
- Covers all state schools in England (93% all pupils), over 0.5m pupils per cohort.
- Use data from PLASCs 2002 – 2009
- Focus on pupils that are identifiable within the state-system throughout this period (90% of the cohort)
- Final sample about 3.5m students

Structure of attainment data

- Students << Subjects << Exams
- Students:
  - typically take 7 – 8 subjects, of which 3 are compulsory (English, maths, science).
- Subjects
  - we know the overall grade for each subject. Subjects are assessed by mixture of exams and coursework, and we know fraction of coursework.
- Exams
  - we know dates of each exam, but not the mark for each exam.

Variables

- Pupil data: gender, ethnicity, within-year age, FSM, SEN, EAL; test scores; school attended.
- Dependent variable is the pupil's score in high-stakes exams at the end of compulsory schooling at age 16, GCSEs.
- We have this data for each subject that each student takes.
- GCSE scores are measured using National Curriculum points.
- We normalise the scores separately for each subject to remove any differences in subject difficulty
  - normalisation is done over all the years together as our focus is on across-year within-subject variation.

T1: Data Description

	All	With both "late" and "early" subjects
	%	%
Male	50.15	49.27
FSM Eligible	12.05	11.03
SEN – non-statemented	13.48	11.40
SEN – statemented	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
GCSE score, normalised	-0.041	0.014
Keystage 2 score	27.03	27.34
Number of students	3,651,667	2,970,694
Total observations (subjects*students)	25,705,081	21,963,321

Results

- Aggregate data
- Simple differences
- Within-individual (late – early) differences and compare the distribution of these between tournament and non-tournament years.
- Robustness checks
- Quantifying the effect sizes

Results 3

- These differences may be confounded by any other year to year effects: use difference-in-difference analysis
- Define ‘late’ subjects and ‘early’ subjects:
  - In tournament years, late subjects are those in which at least two thirds of the exams are on dates overlapping the tournament.
  - In non-tournament years, take the same calendar dates in the tournament years to define late subjects.
- Examine within-student differences in performance between late and early exams.
- Likely that there will differences in performance on subjects late in the exam period versus early in the period for a number of reasons.

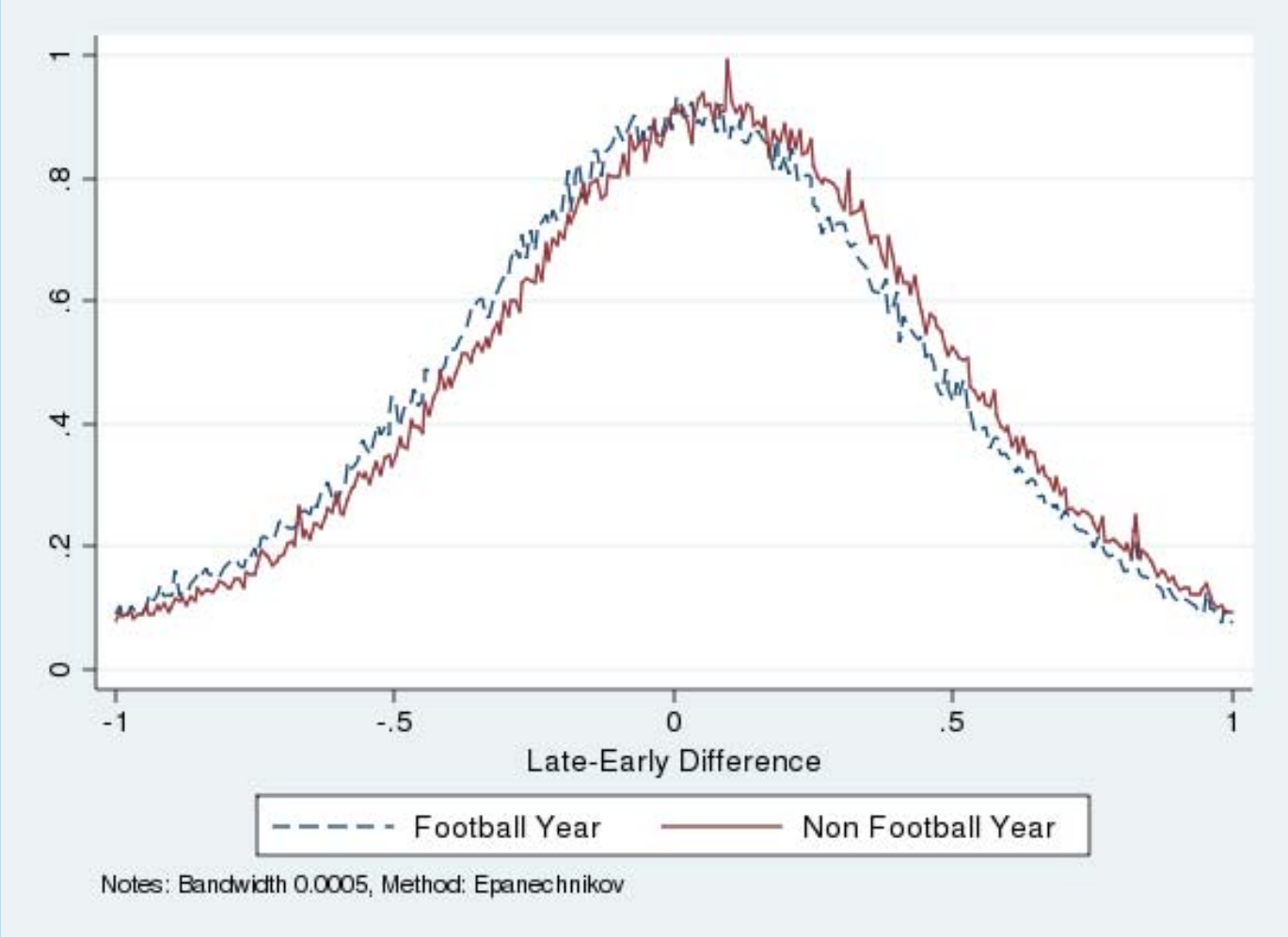
- For each pupil in each year:
  - Define a late – early difference as the student's mean score over her/his late subjects minus her/his mean score over the early subjects.
  - From our model, in non-tournament years :

$$\bar{q}_{i,t=NT,m=late} - \bar{q}_{i,t=NT,m=early} = \pi_{i,late} - \pi_{i,early}$$

  - all observed and unobserved individual characteristics drop out, the year effect drops out leaving only that student's idiosyncratic performance change through the exam time.
  - In tournament years, it is the same plus the impact of the tournament whilst it is in progress

$$\bar{q}_{i,t=NT,m=late} - \bar{q}_{i,t=NT,m=early} = \pi_{i,late} - \pi_{i,early} - f_i\left(\phi_i^1, a_i\right)$$

F3: Density functions for (late-early) difference



T3: Regression on (late – early) difference

	(1)	(2)	(3)	(4)
Tournament year	-0.063***	-0.054***	-0.063***	-0.050***
	(0.001)	(0.002)	(0.001)	(0.002)
Tournament year interacted with:				
Male		-0.026***		-0.025***
		(0.002)		(0.002)
FSM		-0.021***		-0.019***
		(0.002)		(0.002)
Low prior attainment		-0.009***		-0.009***
		(0.002)		(0.002)
High prior attainment		-0.011***		-0.011***
		(0.002)		(0.002)
School Fixed effects			Y	Y
Observations	2970694	2970694	2970694	2970694
R-squared	0.03	0.03	0.03	0.03

Unit = Individual student; Metric =subject level SD; Dependent variable is student's (late-early) difference; other student chars included as main effects and interactions

T4: Difference in differences

Prior Attainment	Not Eligible for FSM		Eligible for FSM		All pupils
	Female	Male	Female	Male	
Lowest	-0.0584***	-0.0679***	-0.0649***	-0.1077***	-0.0680***
	(0.0032)	(0.0035)	(0.0057)	(0.0065)	(0.0029)
Middle	-0.0253***	-0.0740***	-0.0208***	-0.0993***	-0.0495***
	(0.0027)	(0.0031)	(0.0060)	(0.0074)	(0.0025)
Highest	-0.0343***	-0.0661***	-0.0385***	-0.0755***	-0.0507***
	(0.0027)	(0.0028)	(0.0075)	(0.0082)	(0.0023)
All Pupils	-0.0385***	-0.0680***	-0.0471***	-0.0991***	-0.0556***
	(0.0022)	(0.0025)	(0.0043)	(0.0050)	(0.0021)

Metric is subject-level SD; The normalisation is by subject.

Matching

- Exact match:
  - We match within school
  - observables of student gender*FSM status*prior attainment group (3)*broad ethnic group*quarter of birth.
- So 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.
- This procedure generates a difference for each of 190 groups times about 2500 schools.
- We take the mean GCSE score within each school*observables group and difference this between tournament and non-tournament years.
- Display quantiles from this distribution

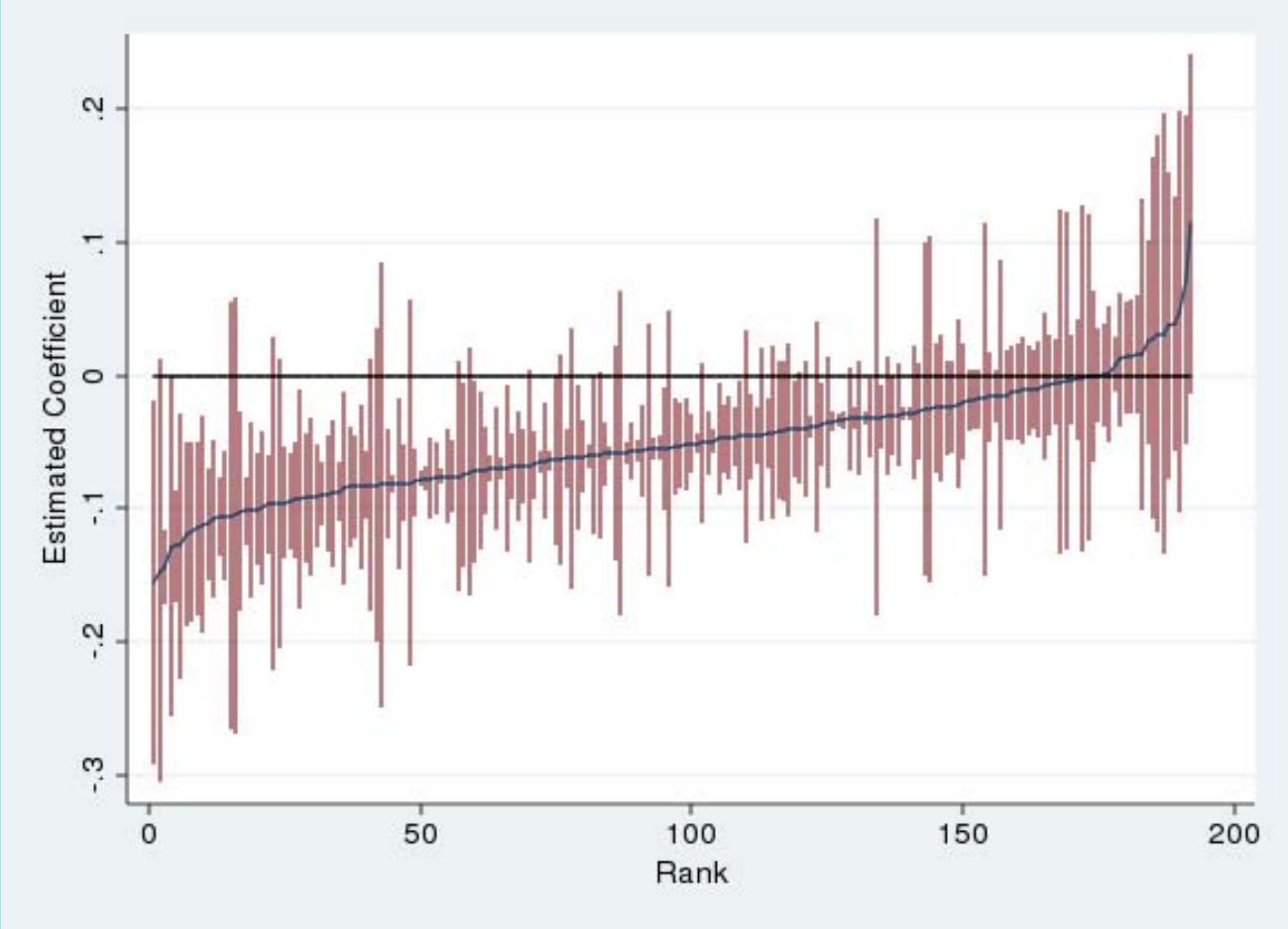
T5: Quantiles of Differences-in-differences

	{(Late – early) in tournament} - {(Late – early) in non- tournament}						
	p5	**p10**	**p25**	**p50**	**p75**	**p90**	**p95**
All Pupils	-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

These figures are based on school-group matching. For all pupils, 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. Quantiles of the distribution of the following statistic are reported:

$$\bar{Y}_{sg,late} - \bar{Y}_{sg,early})_{tournament} - (\bar{Y}_{sg,late} - \bar{Y}_{sg,early})_{non-tournament}$$

F4a: Difference-in-differences by matched groups



Metric is subject-level SD units.

Threats to identification

- In the variable?
- Through the matching?
  - the mean unobservable characteristics within a school for a group

    (say, poor, white, middle ability boys born in the first quarter of the year)
  - differed significantly between the years (2002, 2004, 2006) and the years (2003, 2005, 2007, 2008), and differed in a way correlated with tournament years across the 400,000 school*groups.

Selection away from late exams?

- Do some specific (low ability) students avoid selecting options with late exams in tournament years?
- This seems unlikely:
  - Some subjects see significant changes to the timing of their exams from year to year. Exams varied between 'early' and 'late' for half of subjects over this period.
  - Optional subjects are chosen about two and a half years before the summer exams – timing unlikely to be a major factor.

Testing for differences in observables

- We ran a difference-in-difference, comparing average prior ability of those taking late options with those taking early options, across T and NT years.
- For prior ability (mean KS2), coefficient is 0.001 of an SD, not significantly different from zero (even with 12.2m observations!)

Results – robustness checks

- Alternative definition of “late”: half of exams overlap with tournament period (rather than two thirds)
  - Average effect is still strongly negative but as the definition is less sharp, this produces a lower estimated effect.
- Some students sit exams (typically maths) a year early:
  - omitting these the results are largely unchanged.
- Counting 2008 as tournament year
  - Reduced negative effect for boys; positive effect for girls
  - 2008 is unfortunately a strange year in that science switches to late that year.
- Extend dichotomous early/late subject variable and construct a continuous variable from the exam timetable information
  - Table 6
- Time series impact by group:
  - Just use the time series variation

Results – Quantifying the effect

- Estimates of:
  - Effect on late exams relative to early
    - Cleanly identified but not necessarily the whole story
  - Effect on overall mean score
    - Is whole effect but may be confounded
- Note: on average, coursework about 50% total
- Look at:
  - Impact of effort on exam scores = diff-in-diff coefficient, doubled.
  - Effect on overall pupil mean score, converting to pupil mean SD units
  - Convert to GCSE (letter) grades

T7: Quantifying the Results

	Impact of effort on exams	Overall Effect Metric: SD of pupil mean score	Overall Effect Metric: GCSE grades
Difference in difference			
Table 3			
Mean (col. 3)	-0.126	-0.015	-0.208
Poor, male, white, low		-0.025	-0.347
attainment (col. 4)	-0.206		
Table 5			
All pupils, (median)	-0.116	-0.014	-0.194
All pupils, (p10)	-0.202	-0.025	-0.347
Male pupils, (median)	-0.140	-0.017	-0.236
Male pupils, (p10)	-0.216	-0.026	-0.361

Column 1 = coefficient *2
Column 2 = coefficient* (1.75/7.80)*(11.54/10.68) {share of late exams}*{converting subject sd to pupil sd}
Column 3 = column 2*(10.68/6)*7.80 {converting to gcse points} {converting to letter grades} multiplying by the number of exams

Comparison effect sizes

- Lowering class size from 24 to 16 students per teacher = 0.22 standard deviations on combined mathematics and reading scores (Krueger, 1999)
- A one-standard deviation increase in teacher quality = 0.15 - 0.24 deviations on mathematics and reading achievement (Rockoff, 2004; Aaronson et al, 2007; Kane & Staiger, 2008 for US; Slater et al for England, 2009)
- "No Excuses" Charter schools = 0.10 - 0.40 standard deviations increase per year in mathematics and reading (Abdulkadiroglu et al, 2009; Angrist et al, 2010)
- UK Literacy Hour = 0.06 deviations increase in reading attainment (Machin & McNally, 2008)

Conclusions

- We used a sharp, exogenous and repeated change in the value of leisure to identify the importance of effort for student performance.
- We compared within-student variation in the exam period between tournament and non-tournament years.
- We used seven years of high-stakes subject-level data on 92% of all students in England.

Education production function

- Student effort has a big effect on test scores:
  - Big reduction in effort reduces exam scores by 0.2SD
- This matters:
  - Effort is manipulable, incentives for effort can work and can produce big effects
  - Potentially high-value interventions after “early years”
  - It may be that the strong results in KIPP schools, “No Excuses” schools, Charters arise through eliciting greater effort
  - Some suggestions from neuroscience that high levels of effort directly affects cognitive development.

PS …

- Intervention to examine different ways of incentivising student engagement and effort:
  - Financial treatment (T1)
  - Event treatment (T2)
- Aim to generate exogenous variation in dimensions of “effort” and engagement
- RCT design

Large scale

- 10,000+ kids, 63 schools in ITT sample
- * 3 subjects each so approx 30,000 outcomes
- 7,500 kids with behaviour data
- 40 items of behaviour data per kid, matched with demographics and GCSE performance data
- Paid out > £0.5m in financial incentives

Scientific issues

- 'Effort' and 'engagement' are similar concepts to non-cognitive attributes, and to one of the 'big 5' psychological traits (conscientiousness).
  - But are more variable; amenable to incentivisation?
  - Can learn the value of conscientiousness? Can induce it?
- Long-run effects and implicit motivation
  - Negative? Positive? Can track through A levels (and beyond)
- Inputs not outputs:
  - Incentivise behaviours not outcomes
- Immediacy:
  - Reward 'immediately', rather than 6 months down the line
- Loss aversion:
  - Not in reality (b******** banks)
  - Using framing

Design issues

- School-year level rather than pupil level
  - Power
  - Compliance and fairness
  - Importance of friendships for adolescents
- Not all subjects: English, Maths, Science
  - Diversion
  - Feasibility and cost
- Time unit is a half-term (5 weeks)
  - Immediacy
  - Long-term learning
- Target of incentive
  - Inputs: behaviour (in E, M, S); homework (in E, M, S); classwork (in E, M, S); attendance.
- Threshold design:
  - Pros and cons
- One year programme, two year course with some fraction done
  - Underestimate of effect

Initial results

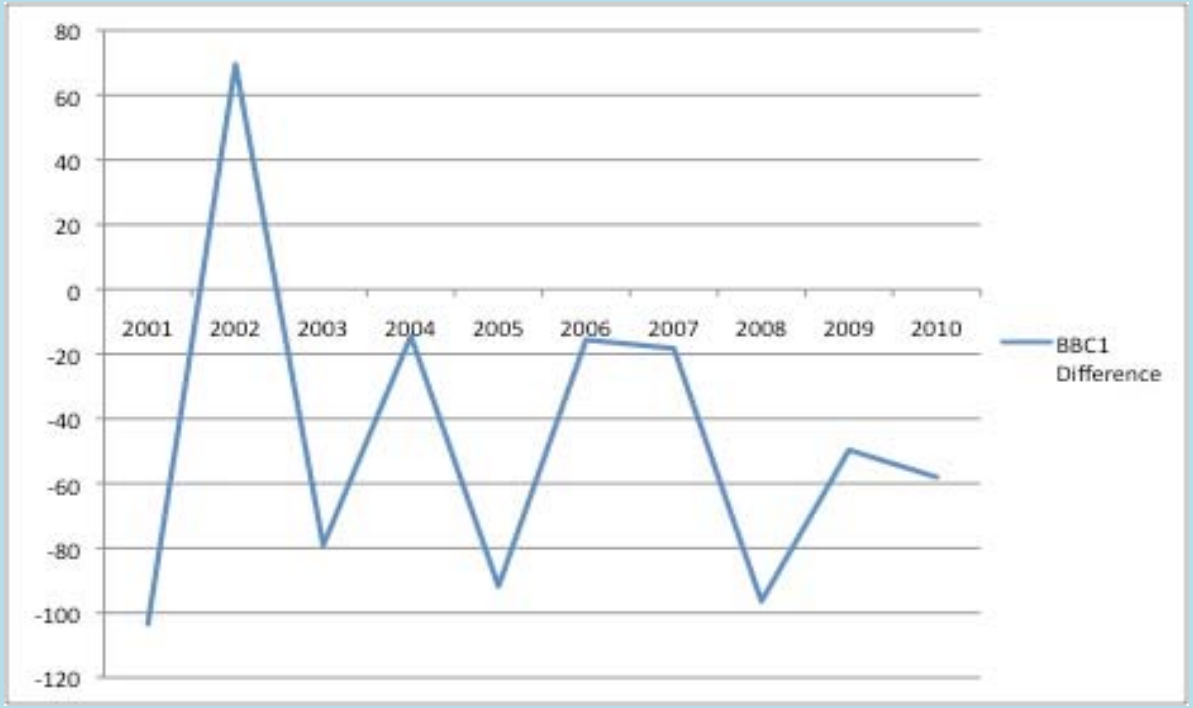
- Under-powered, not enough schools
- On test scores:
  - Nothing on English (not unusual)
  - Effect sizes on maths good in Event treatment
  - But doesn't reach significance
  - Science still to do
- On behaviour:
  - Overall: T2 > T1 > 0
  - Conduct: T2 > T1 = 0
  - Classwork: T2 = T1 >> 0
  - Homework: T2 > T1 = 0
  - Attendance: T1 > T2 = 0

- Lots of data generated, yet to explore

Extras

Television data

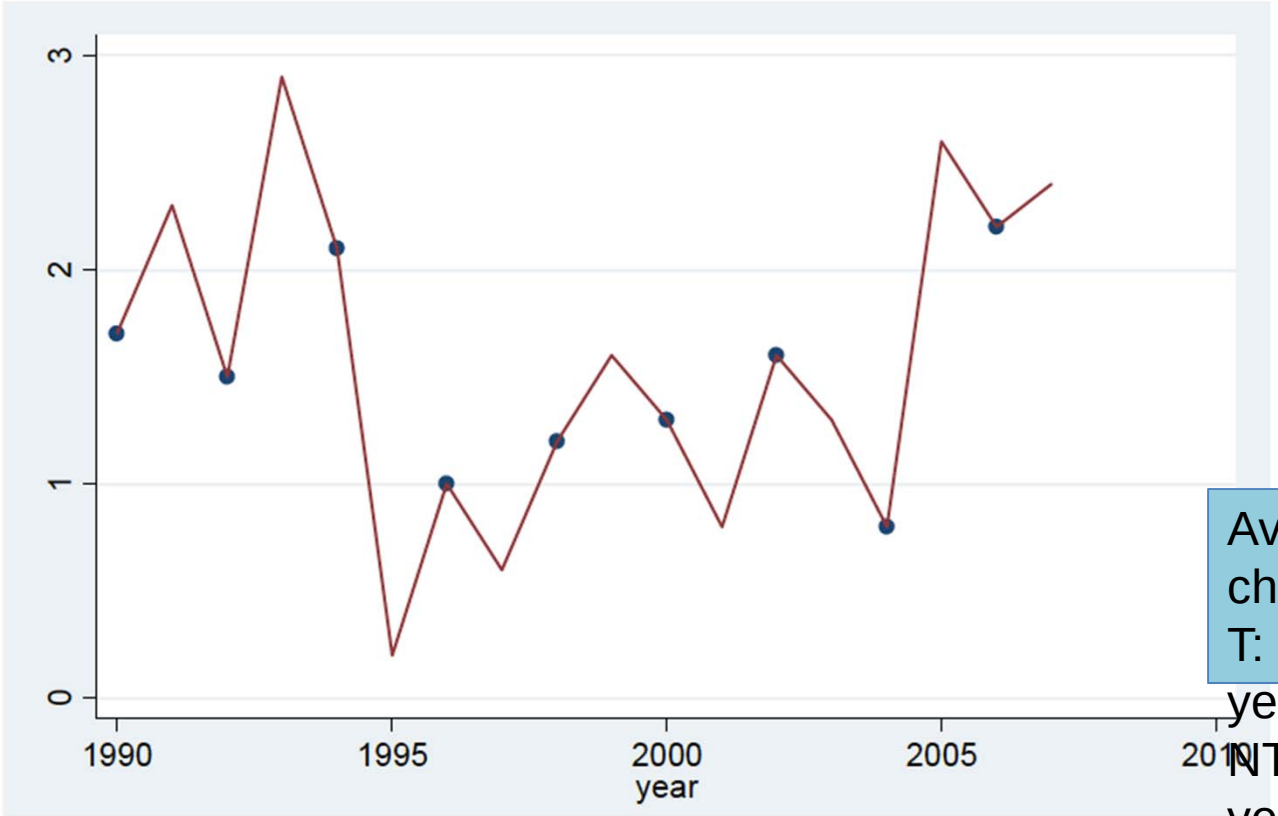
- Difference in monthly (June – April) TV figures (top 30 programmes) for BBC1, millions of viewers.
- Spikes during 2002, 2004, and 2006, and troughs in 2001, 2003, 2005.



Results 1

Annual change in % of pupils obtaining 5GCSEs

Tournament years highlighted.



Average annual change:

T: 1.49 ppts per year

NT: 1.63 ppts per year

Results 2

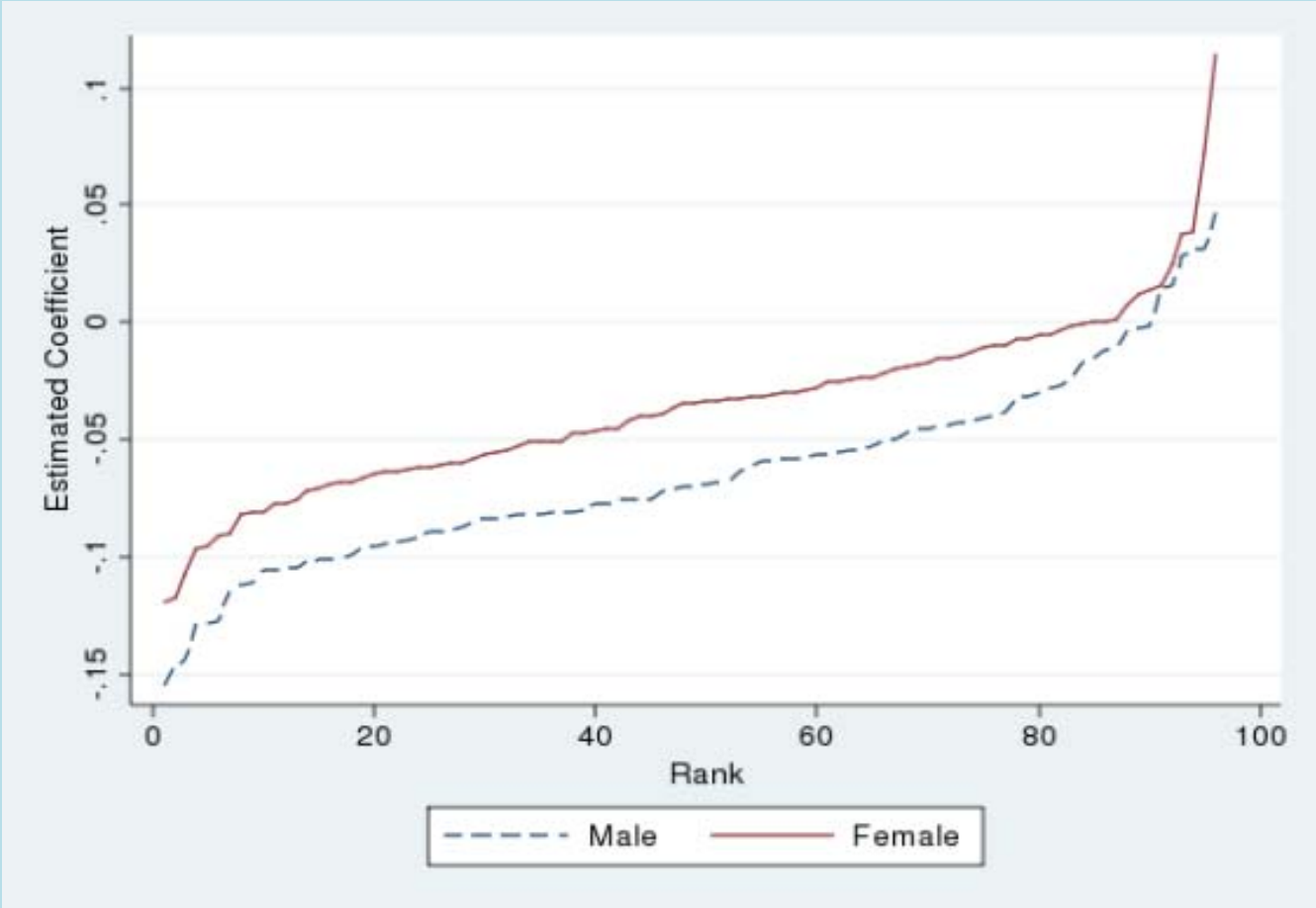
- Simple differences:
  - how students perform in tournament years against a similar set of students in non-tournament years
- Simple difference incorporates:
  - the pre-tournament build-up effect and the effect during the tournament itself.
  - the possibly-differing populations in tournament and non-tournament years,
  - any differences in the general year dummies,
  - So NOT cleanly identified

T2: Simple Average Differences

Prior Attainment	Not Eligible for FSM		Eligible for FSM		All pupils
	Female	Male	Female	Male	
Lowest	0.0508***	0.0357***	0.0210***	0.0146**	0.0369***
	(0.0030)	(0.0030)	(0.0058)	(0.0061)	(0.0024)
Middle	0.0312***	0.0151***	-0.0012	-0.0105	0.0211***
	(0.0023)	(0.0025)	(0.0070)	(0.0076)	(0.0020)
Highest	-0.0206***	-0.0487***	-0.0890***	-0.1225***	-0.0419***
	(0.0022)	(0.0025)	(0.0094)	(0.0103)	(0.0019)
All Pupils	0.0133***	-0.0026	-0.0178***	-0.0258***	-0.0014
	(0.0021)	(0.0022)	(0.0051)	(0.0054)	(0.0018)

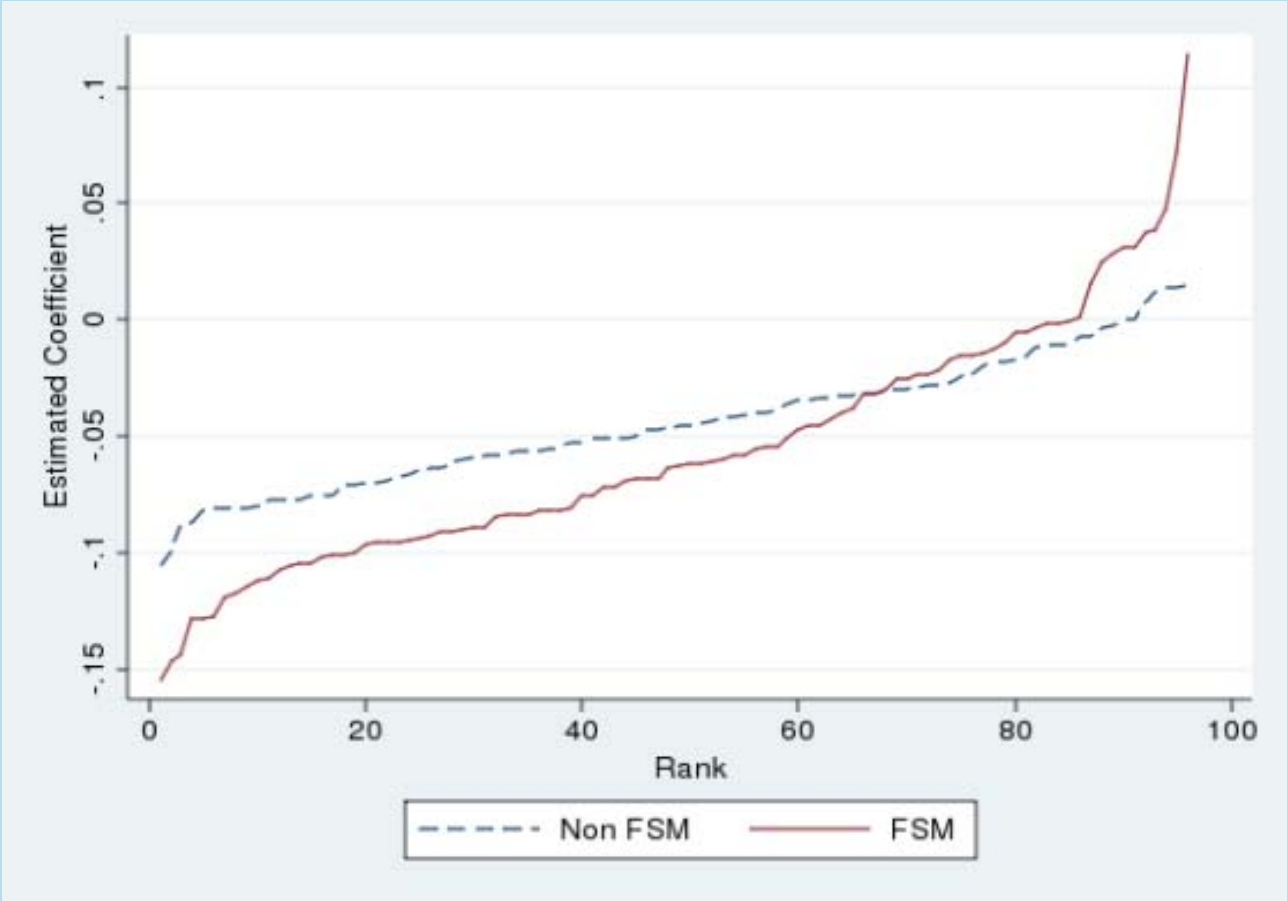
Metric is SD of student average score ; The number in each cell is: {mean (over pupils in that cell) of the pupil-mean of (normalised GCSE scores) in football years} – {mean (over pupils in that cell) of the pupil-mean of (normalised GCSE scores) in non-football years}.

F4b: D-in-Ds by Gender



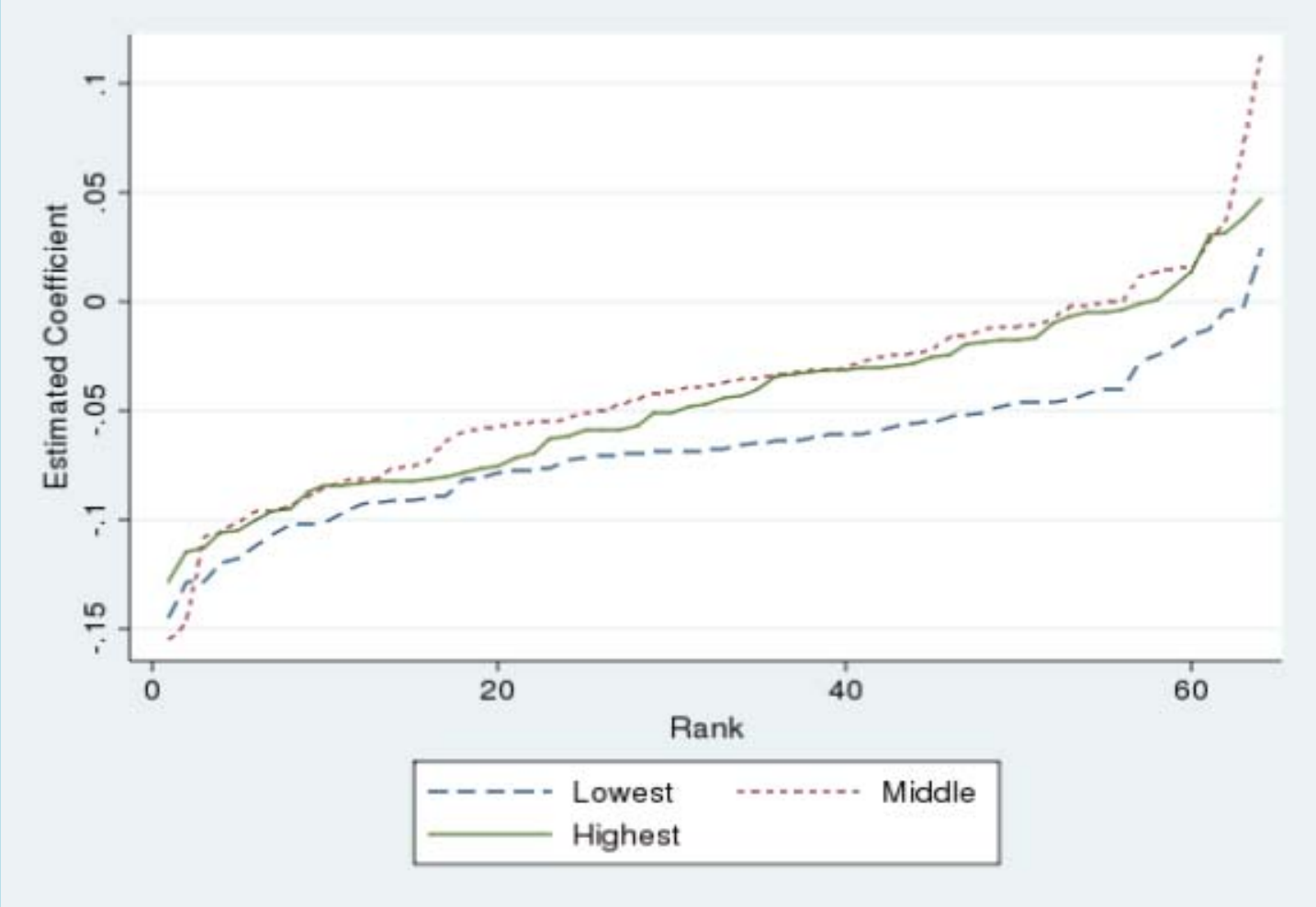
Groups ranked within gender

F4c: D-in-Ds by Poverty status



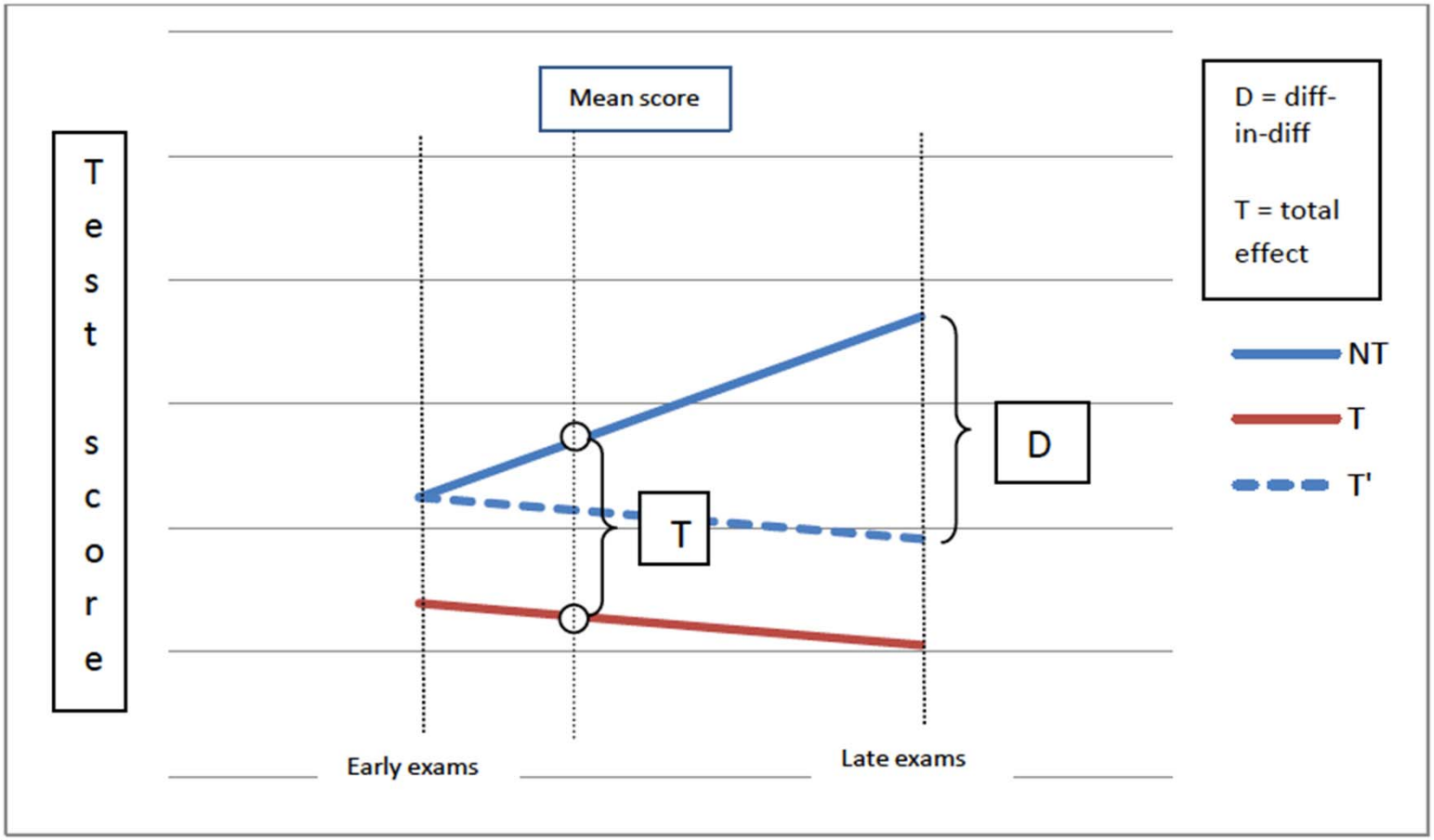
Groups ranked within poverty status

F4d: D-inDs by Ability Level (KS2)



Groups ranked within ability level

F5: Comparing the Difference in difference and the total effect



T6: Student*subject fixed effect results

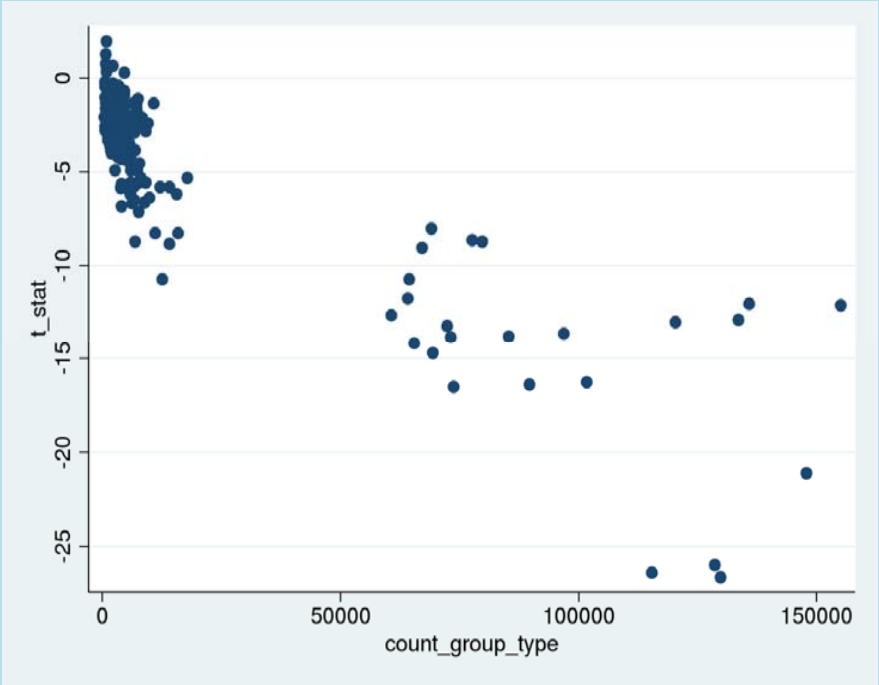
	(1)	(2)	(3)
Proportion of exams within subject which are "late"	0.068***	0.103***	0.126***
	(0.001)	(0.001)	(0.001)
Proportion of exams within subject which are "late" * Year is a tournament year	-0.007***	-0.009***	-0.009***
	(0.001)	(0.001)	(0.001)
Year dummies	Y	Y	
Student Characteristics		Y	
Student fixed effects			Y
Number of observations	25,705,081	25,705,081	25,705,081
Number of pupils	3,651,667	3,651,667	3,651,667

Observation = student*subject; Metric is subject-level SD; Standard errors in parentheses; standard errors clustered at student level.
Student characteristics are: gender, ethnicity, month of birth, poverty status, SEN status, English as additional language, prior ability measures (Keystage 2 English score, Keystage 2 maths score, Keystage 2 Science score)

Time series variation

Run a regression of group mean (late – early) difference on time trend, tournament dummy and constant for 7 annual observations, 2002 – 2008. Run for each of our 192 {gender*ability group*poverty status*qtr of birth*ethnic group} groups. Report the distribution of p-values ....

Decile of P-value	Mean Size of group	P-Value
1	99,184	0.00
2	22,686	0.00
3	7,666	0.00
4	3,893	0.00
5	2,956	0.00
6	3,249	0.01
7	3,493	0.04
8	2,925	0.09
9	3,143	0.22
10	1,766	0.60



Local policy issue

- Bring summer exams forward a few weeks
  - Concentrate exams in early weeks of period.
  - Shift whole school year a few weeks?
- Benefits:
  - Average effect on pupil mean GCSE score: 0.015σ
  - Greater effect on disadvantaged students, male students, black Caribbean students.
  - So would raise the average and reduce inequality.
- Costs:
  - Transitional costs

Further ideas …

- Impact on A levels, degree performance, wages, …
- Other countries with important exams overlapping the tournament period:
  - Football-loving countries
  - Non-football-loving countries
- Collect time use information from students.
- Field experiments of student incentives.