Educ Asse Eval Acc
DOI 10.1007/s11092-017-9264-2

Might progress assessments hinder equitable progress? Evidence from England

Benjamin Alcott[1]

Received: 31 October 2016 / Accepted: 4 June 2017

Abstract Prior research has highlighted the importance of educational achievement throughout school in predicting subsequent progression to higher education in England. However, progress assessments may not only demonstrate students' prior academic achievement but also influence their future achievement. I compare students who have received different grades on one such assessment, despite performing almost identically, to see whether grade labels influence their progress to post-compulsory education. Further, I investigate whether any impact differs according to socio-economic status. Results indicate that grade labels received in eighth grade influence students' performance in school-leaving exams and enrollment in post-compulsory schooling. For lower socio-economic students, this impact is higher than for other students and extends to university enrollment.

Keywords Grade labels · University access · Equity · Regression-discontinuity design

1 Introduction and conceptual framework

Like many countries, England's recent governments have promoted higher education as a means to national economic growth. Enrollment rates have increased accordingly: the proportion of citizens aged 18–22 years enrolled in a degree course has risen from 5% in 1960 to 40% in 2013 (Callender 2006; Vignoles 2013).

However, this expansion has disproportionately served students from the higher socio-economic classes (Anders 2012a; Archer et al. 2003; Chowdry et al. 2013). By 2009, students from the most advantaged quintile of households were six times more likely to attend university than those from the least advantaged quintile (Vignoles and Powdthavee 2009). Thus, although England's three major political parties all support

✉ Benjamin Alcott
bma27@cam.ac.uk

[1] University of Cambridge, 184 Hills Road, Cambridge CB2 8PQ, UK

Published online: 05 July 2017

increasing undergraduate enrollment among poorer students (Conservatives 2015; Labour 2015; Liberal Democrats 2015), concerns about access persist.

While researchers have offered a range of theorizations for this disparity, the strongest inferential studies to date indicate that achievement in the early and middle years of schooling is crucial. Such research demonstrates that achievement measures in national examinations explain a great deal of the gap in undergraduate enrollment across socio-economic classes (see, for example, Anders 2012a; Chowdry et al. 2013; Marcenaro-Gutierrez et al. 2007; Vignoles and Powdthavee 2009). However, the mechanisms by which earlier achievement drives future enrollment behavior have not been fully clarified.

In this study, I test a key assumption in education policy, namely that national summative assessments, which in England are ostensibly low stakes until age 16 years, reflect students' academic ability in a transparent and innocuous manner. In contrast, I examine whether these assessment results play a role in shaping students' subsequent educational outcomes, and whether any impact may differ according to students' socio-economic class. I do this by separating the information that students receive about their academic achievement from the underlying achievement itself.

More specifically, I aim to distinguish the impact upon a student's educational progress of receiving a lower or higher grade at age 14 years, independent of achievement. For example, do students who receive a higher grade in mathematics attain at a higher level in the future than those who performed indistinguishably but received a lower grade? Through a regression discontinuity design, I compare the impact of receiving the average grade, versus receiving the grade below, on the transition to post-compulsory education. In addition, I test whether impacts differ according to students' socio-economic status, thus exploring whether this factor may help to explain disparities in university enrollment.

The exams used are the Key Stage 3 national assessments in English and mathematics that, until 2009, students took at the end of eighth grade.[1] The key grade boundary of interest lies between the average grade and the grade just below. Grades are defined according to cut-points in the underlying, continuous test score; students, parents, and teachers receive information on the grades but not on the continuous test scores. Examining authorities do not decide the cut-points until after students have sat the exams, and these cut-points change every year. The outcomes of interest are three of the criteria by which academic progress is often judged in English policy debates: (1) achievement in tenth-grade national examinations, (2) enrollment in the non-compulsory grades of high school, and (3) enrollment in a university degree course.

I find some evidence that grade labels influence students' educational progress years later: students who receive the higher grade in their eighth grade English assessment perform better in tenth-grade examinations and are more likely to enroll in the non-compulsory grades of high school. However, when exploring heterogeneous effects by socio-economic status, there is far stronger evidence of a bifurcation among low-SES students. For these students, the higher grade label leads to a more sizeable increase in future exam performance, and effects extend to college enrollment. In contrast, there is relatively limited evidence of an impact among high-SES groups, suggesting that grade

[1] These assessments are now commonly replaced by an alternative, Cognitive Assessment Tests, the implications of which I discuss later in this manuscript.

labels have a disproportionately large impact on those students least likely to progress to post-compulsory education.

1.1 England's education system

One key feature of England's national curriculum is that students begin to specialize in subjects from an early age. At age 14 years, students select about ten subjects—from approximately 40 options—on which they will be tested at age 16 years in national General Certificates of Secondary Education (GCSE) examinations. Schooling ceases to be compulsory after these exams; those who do not pass five or more GCSEs—around 40% of students (House of Commons Education Committee 2013)—tend to leave the standard high school system in favor of more vocational courses or employment.

Those that do pass at least five GCSEs are able to continue to the final 2 years of high school to take advanced levels (A-levels), which are generally seen as providing the most academically rigorous courses and a pre-requisite for university study. Students tend to choose three or four A-level subjects, can only study those for which they took GCSEs, and must select specific subjects in order to study a particular subject for university. For example, students hoping to study medicine at university are advised to take advanced science options at GCSE and have to take chemistry at A-level. In their university applications, students must specify which subject they plan to study. During their undergraduate degree programs, students do not take introductory classes across a range of subjects; instead, they only study courses in either a single- or dual-subject program from the outset.

All undergraduate applications are managed by a single organization: the Universities and Colleges Admissions Service (UCAS). The application process is largely uniform across institutions: universities have access to candidates' personal statements, anticipated A-level results (as predicted by schoolteachers), and GCSE results. Only a minority of institutions uses interviews to further screen applicants.

Students hoping to progress beyond the compulsory stages of education are thus required to choose appropriate GCSE subjects at age 13–14 years and perform well in these subjects at age 15–16 years. GCSEs provide a strong predictor of future university attendance (Chowdry et al. 2013), operating as a "symbolic and material currency in terms of future educational progression" (Davey and Fuller 2013). UCAS has been in place for 20 years, GCSE exams for 28 years, and A-levels for over 60 years. For two decades then, the national school curriculum and undergraduate admissions process for English universities have followed a consistent pattern with uniform processes. However, this system also puts pressure on students to envisage coherent academic trajectories and perform well in examinations from mid-adolescence.

1.2 The role of summative testing

Although GCSE results are the first examinations to formally influence students' subsequent opportunities, national assessments begin earlier in the lifespan. In fifth grade, students complete Key Stage 2 assessments in English, science, and mathematics. Until 2009, students would repeat these subjects in Key Stage 3 assessments during eighth grade. In 2009, the UK government abolished testing in Key Stage 3 exams. A

standardized alternative, Cognitive Abilities Tests (CATs), has become increasingly widespread in recent years; the producer of these tests claims that they are now used by two thirds of schools (GL assessment 2015). The UK government has not mandated CATs but does promote their use (Department for Education 2014). In the Sect. 4 of this manuscript, I will return to the implications of the use of these CATs instead in place of the Key Stage 3 assessments; for all empirical analysis, though, I focus on conditions in which students took the Key Stage 3 assessments.

For the Key Stage 3 assessment in each subject, a given student, their parent(s), and teachers received level scores, equivalent to discrete letter-grade boundaries. Consequently, this grading system allowed for direct comparisons among students. In theory, all national examinations prior to year 11 are low stakes for students in the sense that their primary role is to assess the performance of schools (Daugherty 1995), rather than to provide students with qualifications that influence their subsequent study and employment opportunities. Yet, Conner (1999) argued that Key Stage 3 exams had attained a de facto high-stakes status, and the motivational impact on students of knowing how one has been ranked in such assessments should not be overlooked (Broadfoot 1999).

As noted, the strongest inferential studies to date have established that strong predictors of post-secondary enrollment occur from childhood and early adolescence. Among these predictors, prior attainment in national examinations is the most powerful (Vignoles 2013). However, while such studies have established the predictive power of attainment measures, the mechanisms by which attainment measures might influence enrollment have not yet been confirmed through quantitative analyses. One starting point is to divide explanations between those for which pre-existing attributes define and are reflected in attainment measures—such as academic aptitude, self-discipline, and confidence—and those for which assessment feedback has some impact in itself, independent of underlying causes of a given level of attainment. In this study, I focus on the latter.

Among policymakers, a common rationale for summative testing is that it raises academic standards (Harlen and Deakin Crick 2003; Kellaghan and Greaney 2001), but this perspective may take too little account of the complexity of factors relating to motivation (Kellaghan et al. 1996). There is countervailing evidence suggesting that summative testing is detrimental to students who achieve lower grades, i.e., instead of general uplift, testing leads to greater polarization (Bourdieu 1998; Paris et al. 1991; Pollard et al. 2000). Correlational studies between students' self-esteem and achievement find an increase in this relationship for cohorts who have taken more national examinations (Davies and Brember 1998, 1999). Ethnographic studies find that classroom interactions changed in the wake of students receiving exam results, with the self-esteem of lower performers dropping (Leonard and Davey 2001; Reay and William 1999).

Prior research across the social sciences offers a range of potential theorizations for why the Key Stage 3 assessments might have a detrimental influence on those receiving lower grades. Among these, perhaps the most pervasive in educational research is attribution theory, a phenomenological approach that focuses on an individual's judgment of the cause underlying a negative experience. The three key dimensions of an individual's reasoning are whether an event was due to factors that the individual believed to be (1) internal to themselves, (2) controllable, and (3) stable (Weiner 2010).

For example, if a student attributes a disappointing exam result to insufficient effort, this cause is internal, unstable, and controllable, so they will experience guilt and regret, which are positive motivators. In contrast, a student who receives identical negative feedback but attributes this to insufficient aptitude, a cause that is internal, uncontrollable, and stable, will experience hopelessness, shame, and humiliation, and so will see little point in studying for future assessments. In this theorization, summative examinations such as those at Key Stage 3 are detrimental to motivation and achievement (Graham 1990). This is because they can consolidate what Dweck (2006) labels a "fixed mindset": in response to this form of feedback, children are likely to frame exam results as evidence of permanent ability traits, lessening their belief that ability can be developed through work and so deterring future effort (Dweck 1986, 2000).

There is evidence that students in the UK link school exam results to fixed traits that they then project into adult life. For example, on the basis of their observations in a London school, Reay and William (1999) depict a classroom climate where Key Stage examinations serve as a criterion by which students judge themselves and one another. When asked what a high Key Stage grade would say about a classmate, one student responds "that he's heading for a good job and a good life and it shows he's not gonna be living on the streets," whereas her own expectation of a low grade would say that "I might not have a good life in front of me and I might grow up and do something naughty" (pp. 346–347).

Of course, it is not students alone who might be influenced by exam grades. For example, policy demands for accountability may pressure teachers into allocating more resources to some students than others. In recent years, policymakers have treated the number of students achieving five GCSEs (including English and mathematics) at grade C as a key criterion in judging school performance (Vignoles 2013), and this metric figures prominently in governmental school inspections (see, for example, Taylor 2012; Wilshaw 2013). Teachers and school leaders may allocate resources accordingly, with less focus placed on those students projected to perform well above or below this cutoff (Ball et al. 2012; Gillborn and Youdell 2000).

As noted already, England's school system is characterized by early subject specialization. Within schools, ability streams may influence student achievement (Ireson et al. 2005), and only some students may be able to take higher-tier GCSE examinations, in which they can attain the higher grades, or more advanced GCSE subjects, such as triple science. Given the high correlation between students' earlier and later performance in national assessments (Alcott 2013), Key Stage 3 grade levels are likely to present an important tool for teacher and school leaders to help make such decisions. Teachers receive the exam results in terms of discrete grades rather than the continuous underlying scores, and may prefer this heuristic to some form of assessment marked on a continuous scale that they would have to generate in addition to the national examinations.

There are also grounds to hypothesize that such processes exacerbate socioeconomic disparities in educational outcomes. A body of literature claims that, even during pre-elementary and elementary education, formal schooling legitimizes the cultural practices and preferences more typical of the upper and middle classes over those more typical of the lower classes (see, for example, Reay 1995, 1998; Vincent et al. 2008; Walker and Clark 2010). From the earliest years of schooling, teachers label students as intelligent, average, or slow, and stream them accordingly; these

designations are often predicted by social class and have a lasting impact on students' academic confidence (Alcott, 2017; Steedman 1988; Thomas et al. 2012). In interviews, Reay (1995) found that middle-class parents were quick to raise concerns about group-reading activities for fear that these were aiding other children's progress at the expense of their own.

Such class disparities widen further when differential parental knowledge and confidence lead to asymmetrical competition to gain entry for their children to specific middle schools and, by extension, GCSE courses (Ball et al. 1996). England's school system makes the consequences of subject choice and performance in middle school for later opportunities considerable; as a result, early manifestations of the constraining factors identified within habitus are compounded by the time students reach the later years of schooling, well before their course choices and exam performance have easily visible implications for university applications (Vincent 2001). Hence, while many studies focus on students around the final years of the compulsory school system, this associated research indicates that the institutionalized constraints that students face are established from the earliest years of schooling.

Within this conceptualization of formal education, exams serve a key factor, as across schooling, they may serve to consecrate the advantages of more privileged students, serving as both cause and effect of greater separation (Bourdieu 1998, p. 104). Researchers of education in England have used Bourdieu s theories to argue not only that better aggregate performance on exams serves to separate social classes but also that differences in parental behavior according to social class further exacerbate disparities. This is because parents from lower social classes have less confidence in contesting practices within a school, such as placing their child in a lower ability stream or demanding more extensive feedback on classwork (Cochrane 2007, 2011; Giddens 1991; Pugsley 1998).

Consequently, working-class families may be especially prone to accepting summative judgments about their children's progress and ability levels. National exams are a key case in point, as parents from lower social classes are less likely to challenge their validity or encourage their children to accept them as incontestable judgments of ability. Following from attribution theory, learners who attribute success to effort, and who perceive ability to be changeable and controllable, are likely to deal with failure constructively and so persevere with future learning tasks (Schunk 1991). Hence, differential responses to exam results by class are important because it is plausible that this may serve to exacerbate pre-existing attainment gaps.

1.2.1 Aims of the present work

In this manuscript, I undertake an inferential verification of the presence, or lack thereof, of a labeling effect from Key Stage 3 exams, and so can initiate the type of generalizable analyses on these phenomena that are currently lacking for the English context. To date, those studies that are more critical of summative assessment in England have relied predominantly on qualitative research methods to discern the impact of examinations on student behavior (see, for example, Leonard and Davey 2001; Reay and William 1999). The major exceptions, by Davies and Brember (1998, 1999), provide quantitative representations of the impact of Key Stage 1 and Key Stage

2 assessments on self-esteem, but without providing the type of valid counterfactual scenarios necessary to make inferential claims.

Quasi-experimental methods have not yet been applied to estimate the impact, if any, of grades received during these assessments on students' progress to post-compulsory education. Such methods have already identified effects of performance feedback in different educational settings. In Swedish elementary schools, Sjögren (2010) found mixed impacts of the use of grading according to gender and parental education levels. Positive effects of feedback on subsequent educational progress have been found for high school students in Spain (Azmat and Iriberri 2010) and the USA (Papay et al. 2010). In the UK, Sartarelli (2011) has explored the impact of Key Stage scores on student behavior, finding no impact on most outcomes with the exception of bullying.

I aim to transfer such quasi-experimental methods to the study of the impact of low-stakes summative assessments in England on students' subsequent educational outcomes. Specifically, I attempt to identify the impact of receiving a low or high exam grade, independent of achievement. In addition, I investigate whether any effects differ for students of different socio-economic strata. Using a regression discontinuity design, I compare the impact of receiving the average exam level (versus below average) on the following outcomes:

1. GCSE performance at age 16 years.
2. Enrollment in A-levels by age 19 years.
3. Enrollment in a university degree course by age 20 years.

2 Method

2.1 Data

I use data from the Longitudinal Study of Young People in England (LSYPE). The LSYPE used a two-stage probability proportional to size sampling procedure, with the primary sampling units being schools. LSYPE respondents were born between September 1, 1989 and August 31, 1990. Interviews were conducted annually between the spring of 2004, when the youths were in eighth grade, and 2010, providing seven waves of data. As with most longitudinal surveys, the LSYPE is prone to sample attrition in later waves (Anders 2012a; Piesse and Kalton 2009). While the survey's first wave sampled 15,770 youths, sample sizes are reduced to 14,947, 11,186, and 8233 according to whether the outcome of interest related to GCSEs, A-levels, or university enrollment, respectively.

The key predictor variables in my analyses are respondents' Key Stage 3 test results in English and mathematics. I do not analyze respondents' Key Stage 3 test results in science for two reasons. First, unlike English and mathematics, it is not possible to assess students' subsequent progress in science because students do not follow a core curriculum with a common exam in this subject. Second, the predominance of English and math performance in school league tables could mean that science results factor less into teachers' planning for student interventions and ability streaming. Consequently, any impact of test results in science on the more general outcomes of

enrollment in A-levels and university study may differ from the respective impacts of mathematics and English.

LSYPE is linked to the UK Government's National Pupil Database, which holds administrative data on the test scores that students obtained in their Key Stage 3 subject tests. In each subject test, students could receive a grade of 3, 4, 5, 6, or 7 (as well as 8 for mathematics only) depending on the marks that they obtained, from within a range 0–100 for English and 0–150 for mathematics. Figure 1 provides an example of how continuous score mapped onto grade levels in one of these assessments.

Data on the first outcome, passing GCSE examinations, is obtained through the LSYPE's link to the National Pupil Database. This outcome is defined in two respects: subject-specific and general performance. Subject-specific performance compares students Key Stage math grades to whether they receive a grade of C or higher in GCSE math, and the equivalent for Key Stage English to GCSE English. General performance is assessed through the capped GCSE score, which comprises a continuous point score based on the eight best exam results for a given student. The capped GCSE score is a common measure in education policy in England (Vignoles 2013); it is based on students' performance in just eight, rather than all, exams to account for the tendency for students at fee-paying schools to take more GCSE examinations than do students in the state-provided system. The second and third outcomes, enrollment in A-levels and enrollment in university, rely on students' self-reports during the survey's fourth and subsequent waves.

To analyze heterogeneous effects according to socio-economic status, I use the three-class grouping—higher, intermediate, and lower—of the National Statistics Socio-economic Classification (henceforth NSSEC) offered by the UK government's Office for National Statistics (2014), with the caveat that the long-term unemployed are also included in the "lower" grouping due to insufficient sample sizes. Students for whom NSSEC data is missing are included in models of general effects but are not included in models of differential impacts according to socio-economic status.

Table 1 provides descriptive statistics on the label thresholds and outcome measures, first for all students and then according to socio-economic status. There is a clear socio-economic disparity for each of the label thresholds and outcome measures, with higher NSSEC students outperforming intermediate NSSEC students, who in turn outperform lower NSSEC students. Table 2 presents average outcomes for all sampled students according to whether they scored above or below the level 6 threshold in each subject. For each subject, those scoring above the threshold realize each of the outcomes at higher rates than those who scored below the threshold, and always by at least a factor of two. All differences are statistically significant at any conventional level.

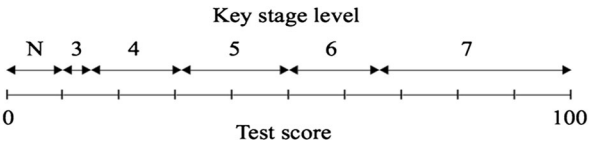

Fig. 1 An example of how test score maps to Key Stage level. Note: Example comes from 2003 English assessment, which LSYPE respondents took. Fewer than 1% of respondents score below level 3

Table 1 Test scores and outcomes for full sample and according to NSSEC level

		By NSSEC level:		
	All	Lower	Intermediate	Higher
English level 6 or above	0.34	0.26	0.35	0.55
Math level 6 or above	0.53	0.44	0.59	0.74
Mean capped GCSE point score	303	287	317	353
Attain C or higher in GCSE English	0.59	0.51	0.65	0.81
Attain C or higher in GCSE mathematics	0.54	0.46	0.60	0.76
Enroll in A-levels	0.56	0.51	0.47	0.73
Enroll in university degree	0.49	0.41	0.46	0.62

Results for the full sample ("All") include students with missing NSSEC data. Source: Longitudinal Study of Young People in England

2.2 Analytical approach

The use of regression-discontinuity designs (RDDs) in social science research dates back half a century (see Thistlethwaite and Campbell 1960; Campbell and Stanley 1963). Although the approach received limited attention in subsequent decades (Cook 2008), it has become increasingly popular in recent years (Lee and Lemieux 2010; McCall and Bielby 2012). In the field of education, studies have used the RDD approach to investigate a range of influences such as financial aid (Van der Klaauw 2002), scholarships (DesJardins and McCall 2014), remedial education programs (Jacob and Lefgren 2004), and pre-school interventions (Ludwig and Miller 2007).

RDD's popularity is likely due to its intuitive appeal as the most valid estimation strategy under specific allocation mechanisms, absent random allocation (Cook 2008; Lee and Lemieux 2010). The specific mechanisms are those in which treatment is assigned via a cutoff point on an observed continuous variable. So long as individuals are unable to manipulate whether they fall on one side of the

Table 2 Sample averages for outcomes by level 6 threshold

	Below level 6	Level 6 and above	*p* value
English			
Capped GCSE point score	265	376	<0.001
Attain C or higher in GCSE English	0.41	0.96	<0.001
Enroll in A-levels	0.40	0.84	<0.001
Enroll in university degree	0.32	0.73	<0.001
Mathematics			
Capped GCSE point score	239	359	<0.001
Attain C or higher in GCSE mathematics	0.15	0.89	<0.001
Enroll in A-levels	0.32	0.75	<0.001
Enroll in university degree	0.25	0.64	<0.001

Source: Longitudinal Study of Young People in England

cutoff point or the other, the variation in treatment around this point is essentially random (Lee and Lemieux 2010). The present study is well suited to an RDD approach because the assessment feedback device—discrete exam grades—varies at distinct cutoff points on the underlying continuous scores that students achieve in each test. Students are unable to manipulate which side of the grade cutoff their scores fall because grade boundaries are not defined until after each year's test. In addition, students, parents, and teachers are provided with discrete grades after the test but not the exact value on the underlying continuous mark scale. This is important because knowledge of the continuous mark might mitigate the response of any of these groups in relation to the grade; for example, teachers might see students who scored just below the threshold as more comparable to those scoring just above it than to those scoring many points below the threshold.

In order to get a higher grade on the Key Stage test, students must meet or exceed a given test score. Thus, a given student's receipt of a higher grade (G) depends on her test score (S). More specifically, the grade depends on their score in relation to a grade cutoff (C), whereby those students for whom $S \geq C$ receive grade G. Taking receipt of a higher grade as a treatment, for student i its relation to a future outcome (Y) can be denoted as

$$Y_i = \begin{cases} Y_i^1 \; if \; G_i = 1 \\ Y_i^0 \; if \; G_i = 0 \end{cases} \tag{1}$$

or

$$Y_i = Y_i^0 + G_i\left(Y_i^1 - Y_i^0\right) \tag{2}$$

However, the inferential challenge is that it is not possible to observe both Y_i^1 and Y_i^0 because student i cannot both receive and not receive the treatment G. Random allocation overcomes this challenge since, taking α as a vector of all variables that could influence Y_i prior to receipt of the treatment, in the equation

$$Y_i = \alpha + \tau G_i + \varepsilon_i \tag{3}$$

G_i and the error term ε_i are independent. Thus, the estimate of the treatment effect τ is obtained by subtracting the average of Y for all untreated students from the average of Y for all treated students.

The properties of the Key Stage grade boundaries make it possible to partially emulate this ideal randomized scenario through a "sharp" RDD design. In a sharp design, the cutoff variable perfectly predicts allocation of the treatment (Imbens and Lemieux 2008), so that

$$\begin{aligned} G_i &= 1\{S_i \geq c\} \\ G_i &= 0\{S_i < c\} \end{aligned} \tag{4}$$

The sharp design is possible because all students whose scores fall below the boundary receive the lower discrete grade, while all those whose scores fall at or above the boundary receive the higher grade.

Returning to Eq. 4, accepting that the expected value of the error term ε_i varies with S_i, but assuming that it does so as a continuous function of S, Hahn et al. (2001) have shown that

$$\tau = \lim_{x \downarrow c} E[Y_i | S_i = s] - \lim_{x \uparrow c} E[Y_i | S_i = s] \quad (5)$$

provides the treatment effect of G. In other words, the treatment effect can be found by taking the difference in outcome Y between those observations just above (specifically, within x points of) the cutoff point c, and those observations just below (again, within x points of) the same cutoff point c.

One key consequence is that this approach only uses observations from the arbitrarily selected range $[c - x, c + x]$ around the cutoff point (McCall and Bielby 2012), whereas all observations may be used under randomized allocation. Thus, the RDD provides an estimate of the local average treatment effect, but it does not allow for extrapolations to observations further from the threshold (DiNardo and Lee 2011; Shadish et al. 2002). In the present study then, any observed treatment effect of the grade label will pertain to students whose attainment level falls near to the grade cutoff point, but extrapolations to those with far higher or lower achievement levels are less plausible.

In order to limit bias, kernel-based polynomials provide a popular means to predict τ under the RDD design (Hahn et al. 2001; Lee and Lemieux 2010). Consequently, researchers face choices regarding three factors that have implications for a given model's validity: the kernel, the bandwidth, and the use of polynomial terms in the regression model (Mealli and Rampichini 2012). Among these, the choice of kernel is of lesser importance because model estimates are not so sensitive to different kernels as they are to different bandwidths and polynomial terms (Imbens and Lemieux 2008; McCall and Bielby 2012). I follow McCall and Bielby (2012), who use a Gaussian kernel, although all models were re-estimated with alternative kernels in order to check whether estimates are highly sensitive to this choice. Estimates using these alternative kernels were not substantively different to those using the Gaussian kernel.

As Mealli and Rampichini (2012) note, choice of bandwidth offers a tradeoff between precision and bias. In order to allow for bandwidths larger than those most commonly used in prior research (Calonico et al. 2014), I use the formulation offered by Calonico et al. (2013) to select data-driven bandwidths for each model in order to minimize the mean square error. In their formulation, Calonico et al. (2013) base their confidence intervals on a bias-corrected discontinuity estimator to account for the impact of large bandwidth choices, but depart from the prior literature by using an alternative formulation of standard errors in order to account for the greater variability in the calculation of a given t-statistic caused by estimated bias correction.

2.3 Limitations

A regression discontinuity design's internal validity depends primarily on whether the distribution of observations around the cutoff point is as good as random. For the current study then, it is important that students cannot manipulate their positioning around the grade threshold. For example, if more advantaged students were able to systematically ensure that their scores fell just above the grade cutoff, a difference in

subsequent outcomes between those just above and just below the grade boundary could simply be due to prior advantage, making spurious an attribution to the grade label itself. In recent years, McCrary s (2008) formal test of the distribution of the running variable has become perhaps the most common means for examining potential manipulations, by testing whether observations fall disproportionately on one side of the cutoff point. As Fig. 2 indicates, the running variables—i.e., the test score in each Key Stage 3 subject assessment—would not pass McCrary s test.

However, this can be explained by contextual factors. First, the examining authority does not define grade boundaries until after the test, and the central examination body already holds examination scores by the time that these boundaries are announced. Second, one key practice that examiners use for the Key Stage 3 assessments is "borderlining," which is the practice of re-marking those papers that fall three marks below the discrete level boundaries (QCA 2004; Quinlan and Scharaschkin 1999). Since this is done for those papers falling just below the boundary but not for those just above, and given the variation in examiners scores (House of Commons Children, Schools and Families Committee 2008), it is natural that the frequency of scores will drop just before the boundary and rise just after. While borderlining proved controversial and was eventually stopped (House of Commons Children, Schools and Families Committee 2008; National Audit Office 2008), it was still in use in 2004, the year in which LSYPE respondents sat these assessments.

To test whether the practice of borderlining benefitted some student groups more than others, I conduct an alternative sensitivity check for potential manipulation. Presented in Appendix 2.A, I regress whether students fall just above or just below each threshold on the key background characteristics of NSSEC grouping and Key Stage 2 performance. Of the 18 models that I estimate in this sensitivity check, just one has a coefficient significant at the 5% level, providing limited evidence of any systematic manipulation of positioning around the level 6 threshold in any of the exams.

Perhaps the greatest constraining factor on the research design's external validity is sample attrition. Whereas the proportion of English domiciled young people aged 17–19 years enrolled in university was 33% in 2008/09 and 34% in 2009/2010

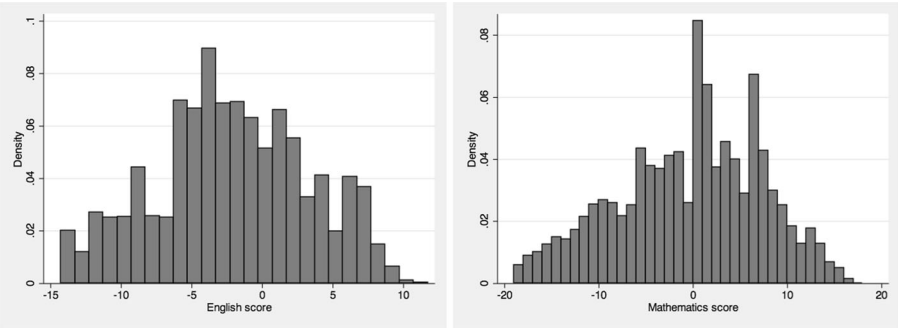

Fig. 2 Histograms of test scores by subject. Source: Longitudinal Study of Young People in England. On the *x*-axis, scores are standardized so that −6, 0, and 6 represent the minimum scores for levels 5, 6, and 7, respectively

(Department for Business Innovation and Skills 2012), enrollment among the LSYPE cohort averaged 43% across these years, even when using survey weights. This indicates not only that survey attrition occurred among students with lower achievement but also that attrition depended on characteristics for which LSYPE's sample weights cannot control (Anders 2012b).

This attrition has lessened the extent to which LSYPE is representative of the national cohort that it was intended to represent. Clearly, sample attrition is not occurring at random: higher attaining students were more likely to continue participating in LSYPE. In addition to this consideration, unobserved characteristics—such as intrinsic motivation or interest in education—might differ between the sample and the national population of students. It is plausible that such factors might influence not only survey participation but also reaction to a grade label. For example, if, as I think plausible, more intrinsically motivated students were more likely to continue participating in LSYPE, and were also less likely to be influenced by an extrinsic stimulus such as a grade label, then model estimates here might underestimate the impact of grade labels.

Another consequence of survey attrition is that the samples that I use for my three outcomes of interest differ slightly from one another. As presented in Table 3, it appears that respondents from higher socio-economic backgrounds and higher Key Stage 3 attainment were more likely to continue participating in the later LSYPE waves, during which the questions on A-levels and university enrollment were asked. This adds a further complicating factor because apparent changes in the impact of the grade label over the educational stages (for example, evidence of a dissipating impact or a compounding impact) may be partially attributable to differences between the subsamples analyzed at each stage.

Consequently, while findings are likely to be indicative of national trends, subsequent model estimates should not be presumed to be nationally representative. Nonrandom missing data problems are a common concern in social sciences (Allison 2002), and, because of attrition, are especially problematic for longitudinal data analysis (Alderman et al. 2001; Goldstein 2009; Molenberghs and Fitzmaurice 2008). This challenge is thus an important but necessary tradeoff for the ability to link phenomena at one educational stage to longer-term outcomes.

Table 3 Student characteristics by sub-sample

Observations	When the following outcome is observed:		
	GCSE 14,947	A-levels 11,186	University 8233
NSSEC level (%)			
Lower	27	27	25
Intermediate	30	28	28
Higher	43	45	47
Mean Key Stage 3 score (standard deviation)			
English	33.1(6.3)	33.7 (6.1)	34.4 (6.0)
Mathematics	35.5 (8.1)	36.3 (8.0)	37.2 (7.8)

Source: Longitudinal Study of Young People in England

3 Results

3.1 For all students

Model estimates for the impact of the level 6 label are presented in Table 4. An accompanying visual depiction is presented in Fig. 3, where scores are standardized so that the level 6 threshold is represented by zero. As noted in Sect. 2, estimates are based on non-parametric models that use Calonico et al. s (2013) estimation procedure in order to minimize standard errors. Consequently, bandwidth size varies across models; information about the bandwidth and sample size used for each model is presented in Appendix 2.B.

These findings provide moderate evidence of a labeling effect from the English examination, but essentially no evidence of the mathematics label having an impact. The level 6 label in the English assessment is associated with a 3.7-point increase in capped GCSE point score. It is worth noting though that this is only significant at the 0.1 level and that this coefficient equates to less than a one-grade increase (e.g., from a C to a B) on a single GCSE exam, which would be represented by a 6-point increase. In addition, the level 6 label in English is associated with a 9.4 percentage-point increase in enrollment in A-levels. This may seem a large increase given that 67% of the full sample enrolls in A-levels, although it is worth noting that the lower bound for this 9.4 percentage-point increase is just 1.4 percentage points. The level 6 label in English does not appear to have a statistically significant impact on a given student s likelihood of attaining a C or higher in GCSE English. In the mathematics examination, the level 6 label is not significant in relation to any of the outcomes. For university enrollment, although the label from mathematics is close, neither it nor the English label is significant at the 10% level.

Table 4 Estimated impact of level 6 label

Outcome	Subject	
	(1) English	(2) Mathematics
Capped GCSE point score	3.741*	4.788
	(2.738)	(3.852)
Attain C or higher in GCSE subject	0.028	0.017
	(0.025)	(0.039)
Enroll in A-levels	0.094**	0.037
	(0.041)	(0.026)
Enroll in university degree	0.039	0.061
	(0.033)	(0.038)

Coefficients for the first row represent the expected change in GCSE points score associated with the level 6 label. Coefficients for the second, third, and fourth rows represent the percentage point change in probability, associated with the level 6 label, of a student achieving the given outcome. Results include students with missing NSSEC data. Standard errors are reported in parentheses. Source: Longitudinal Study of Young People in England

*$p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

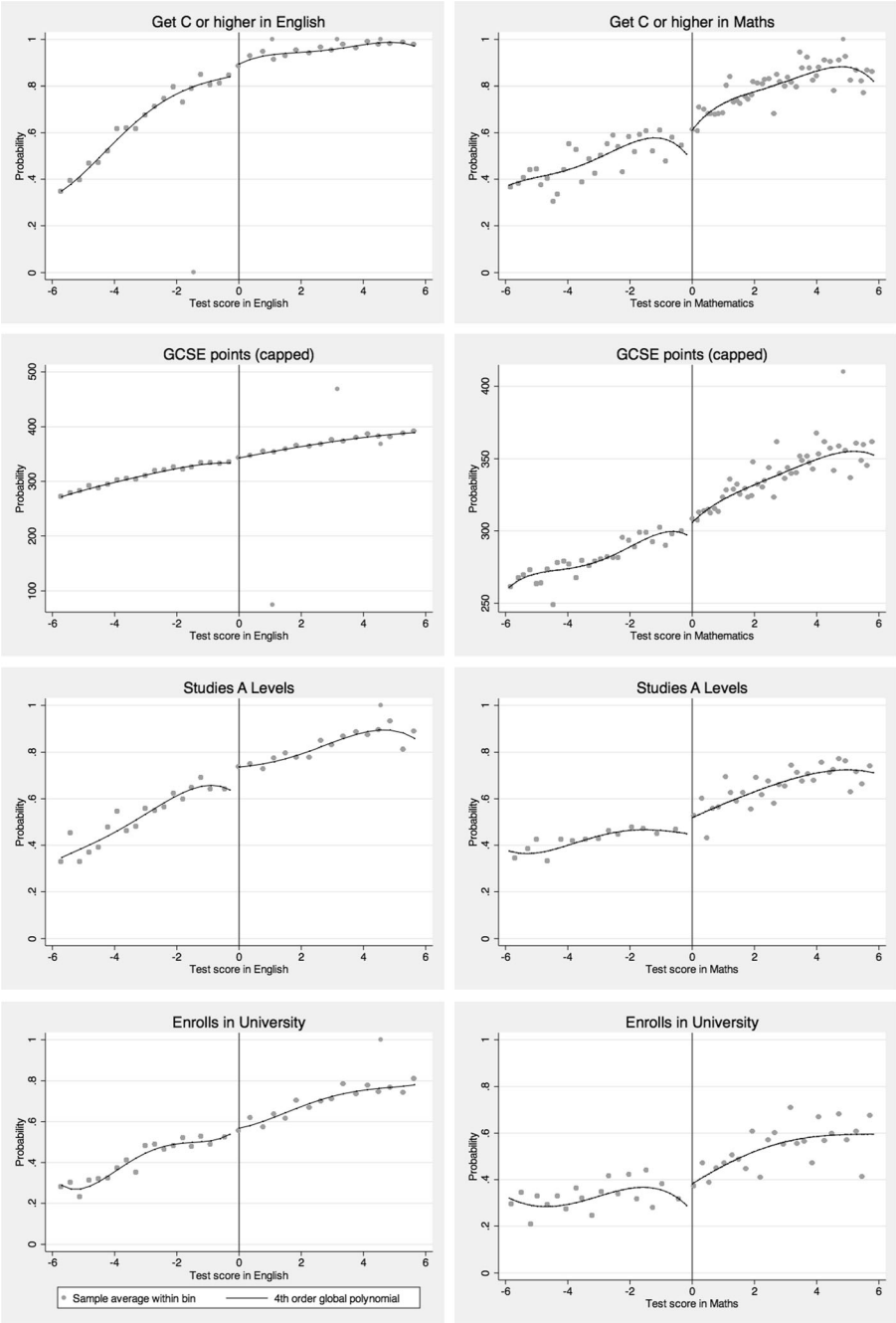

Fig. 3 Regression discontinuity estimates: all students. Source: Longitudinal Study of Young People in England. On the *x*-axis, scores are standardized so that −6, 0, and 6 represent the minimum scores for levels 5, 6, and 7, respectively

3.2 By socio-economic status

Estimates of the level 6 label according to socio-economic status are presented in Table 5. These indicate that labeling effects differ according to socio-economic status: while there is limited evidence of an impact for students from the higher and mid-NSSEC groupings, substantial labeling effects are visible for the lower NSSEC group.

For high-NSSEC students, three of the 11 models provide significant non-zero estimates, albeit each of these is at the 10% level. The English label is linked to improved performance in GCSE English, with a 4.6 percentage-point increase in the likelihood of attaining a C or higher at GCSE. The level 6 label in mathematics is linked to a 10.8 percentage-point increase in the likelihood of getting a C or above in GCSE mathematics and a 9.4 percentage-point increase in the likelihood of enrollment in A-levels. For mid-NSSEC students, there is evidence of a labeling effect for just a single outcome: the mathematics exam is positively associated with an increase of 14 percentage points in the likelihood of university enrollment by age 20, which is significant at the 5% level.

Table 5 Estimated impact of level 6 label by NSSEC level

	Subject	
	(1) English	(2) Mathematics
NSSEC: lower		
Capped GCSE point score	20.11**	28.21***
	(8.40)	(7.57)
Attain C or higher in GCSE subject	0.207***	0.057
	(0.062)	(0.066)
Enroll in A-levels	0.110**	0.082
	(0.054)	(0.056)
Enroll in university degree	0.205**	0.008
	(0.092)	(0.066)
NSSEC: intermediate		
Capped GCSE point score	−1.189	5.438
	(8.47)	(5.862)
Attain C or higher in GCSE subject	−0.020	0.055
	(0.035)	(0.071)
Enroll in A-levels	0.111	0.061
	(0.075)	(0.065)
Enroll in university degree	0.015	0.144**
	(0.083)	(0.071)
NSSEC: higher		
Capped GCSE point score	−1.062	9.494
	(4.200)	(5.160)
Attain C or higher in GCSE subject	0.046*	0.108**
	(0.037)	(0.054)
Enroll in A-levels	−0.068	0.094*
	(0.045)	(0.055)
Enroll in university degree	−0.030	0.039
	(0.052)	(0.048)

Standard errors are reported in parentheses. Source: Longitudinal Study of Young People in England

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

In contrast to the other two groupings, there is considerable evidence of a labeling impact on lower-NSSEC students. The English examination appears to have a positive impact on all four outcomes, with coefficients of 20 points for capped GCSE point score (equivalent to a boost of more than three grade levels), 21 percentage points for attaining at least a C in GCSE English, 11 percentage points for A-level enrollment, and 21 percentage points for university enrollment. The mathematics exam has a positive impact on capped GCSE point score of 22 percentage points, but no discernible impact on pass rates in GCSE mathematics or enrollment in either A-levels or a university degree.

4 Conclusions and discussion

At Key Stage 3, it appears that the level 6 grade label had a polarizing effect on otherwise similar students. This finding supports the hypothesis that feedback from summative testing, even when ostensibly low-stakes, has an important impact on behavior. It corroborates analyses conducted in non-English contexts on the impact of exam results more broadly (e.g., Papay et al. 2010) or on non-academic outcomes (e.g., Sartarelli 2011), and substantiates the claims made by a body of literature (e.g., Black and Wiliam 1998, 2006; Harlen and Deakin Crick 2003; Reay and William 1999) on the English context that has provided compelling phenomenological evidence but lacks the inferential studies necessary to establish a plausible counterfactual.

Moreover, these findings add weight to the claim that students' pathways through formal education depend, at least partially, on social class. Specifically, labeling effects appear to have been greatest for students from lower socio-economic classes providing further support for claims that England's school system does more to worsen than redress educational inequalities. It is worth noting though that magnitude of these estimated labeling effects, especially when considering the lower bounds of their associated confidence intervals, cannot explain the majority of the link between social class, school achievement and university enrollment that has been identified in the literature (e.g., Chowdry et al. 2013; Anders 2012a). Nonetheless, it seems that grade labels play some part both in the achievement-enrollment relationship and socio-economic disparities in enrollment.

4.1 Implications for policy and practice

Since the purpose of testing at Key Stage 3 was not even to award qualifications to students for their performance (and neither, implicitly, to punish those with relatively poor performance), these findings indicate that it may be worth at least reassessing current feedback procedures. Potential adjustments could lie on a spectrum at which the more extreme end would be dropping assessments entirely, should they provide little information about school quality. A less extreme change would be to not share results with students, or, should it be the case that teachers or school leaders are the source of labeling effects from the assessments, these tests might be anonymized. A more moderate approach would be to provide feedback scores on a continuous scale. Even if this were to complicate the interpretation of scores for children or parents, this would likely be compensated for by the benefits of moving away from the present mode of

over-simplification. Perhaps most moderate of all, greater care should be taken over students and parents interpretation of the grades so that lower performers do not see their grade levels as definitive judgments of academic ability, but rather as something that is malleable with greater effort in future. Such an approach would be consistent with the avowed policy objective of fostering "growth" mindsets (Dweck 2006) by enabling students to respond and to persevere with setbacks (UK Department for Education 2014).

With policy changes since the LSYPE cohort, there is already some greater flexibility regarding how schools make use of summative assessment. Whereas Key Stage 2 assessments in English and mathematics follow a similar format to the assessments taken by the LSYPE cohort, CATs, the common replacement for Key Stage 3 assessments, differ from their predecessor more drastically. As noted, even though the government does not require schools to use CATs, a majority of schools do so. Given this absence of central enforcement, the implementation of CATs varies substantially across schools, and our knowledge of this variation is limited. For instance, CATs provide schools with each student's underlying continuous score and band groupings, and schools are also able to define their own band groupings. Some schools use these tests to group children into ability bands, set achievement targets for both students and teachers in the subsequent tenth-grade exams, and share results with students' parents, either in the format of the continuous score, achievement bands, or a mixture of the two.

4.2 Implications for research

The uneven implementation of CATs constrains future inferential statistical research on this age group in England. Even if data on CAT use were to become more widely available, endogeneity would present a problem: there are likely to be unobservable factors influencing why certain schools implement CATs in a particular way that will also be linked to future educational outcomes. However, applications of the regression discontinuity design could still be used for Key Stage 2 assessments to identify whether labeling effects are visible from this earlier age too, and if so, whether their impact may be greater or lesser than those occurring in adolescence. In addition, the UK government has more secure datasets than the LSYPE that would enable more detailed analyses of any relationship between Key Stage 2 assessments and student outcomes in A-levels and higher education. For example, it would be possible to examine performance in A-levels and the type of subjects studied. Similarly, for higher education, it would be informative to analyze differences in institutional quality and subject major. Among other benefits, such research would enable more nuanced analyses of the relationship between Key Stage grade labels in a particular subject, e.g., mathematics, and subject-specific outcomes, such as performance in GCSE mathematics, or the study of STEM subjects at university.

A greater use of qualitative approaches would be especially beneficial to this research field. While the inferential analysis presented here serves to verify the presence of a grade label impact, it is incapable of unearthing who is responding to the grade label or why. When considering the divergent impact of labels on students of different social classes, prior research on social class in England could justify attributing this to the behavior of students (Ball et al. 2002; Reay and William 1999), parents (for

example, Cochrane 2007, 2011; Giddens 1991; Pugsley 1998), or teachers (for example, Steedman 1988; Thomas et al. 2012). The far stronger evidence of impact from the assessment of English than of math could be attributed to the foundational differences in how assessment influences teaching in each subject (Hodgen and Marshall 2005), although I would be inclined to focus on ability streaming practices in England. Namely, while nearly all students are streamed by ability after the Key Stage 3 assessments, 82% are already streamed by ability in math compared to 48% in English (Department for Education 2010), which suggests that Key Stage 3 grades mattered far more to ability streaming in English than to math.

However, such claims can be only speculative when based on the evidence available here. Qualitative work could thus explore the competing and interacting nature of the mechanisms underlying the impact of grade labels—such as whether such tests serve to confirm that teacher prejudice or working-class families are especially prone to accepting summative judgments about their children s progress and ability levels—in a manner that is not possible with the research design undertaken here. Important work has already been produced on the impact of Key Stage assessments on student esteem (Leonard and Davey 2001; Reay and William 1999). Alterations in behavior may also occur for students, peers, parents, teachers, or some other interested party. Further studies that use qualitative methods, such as ethnographic observations or interviews, could shed greater light on who else besides students is responding to grade labels, and the mechanisms by which their responses have an impact.

In contrast to the more uniform practices of the Key Stage 2 exams, the uneven implementation of CATs could be a boon for such research. The variety in conditions could enable researchers to tease out the impact of small differences in how schools choose to share and act upon test results. For example, more phenomenological analyses could investigate whether students respond differently to grade labels when they have direct consequences, such as defining ability streams, or the more general form of judgment that stems from being placed on an ability scale. Similarly, it would be useful to know whether parents still rely on discrete grade boundaries when presented alongside the underlying continuous scores. Thus, the very source of difficulty posed to inferential research could prove fruitful for qualitative research, providing us with more nuanced insights into the role of summative assessment in education.

Acknowledgements I wish to thank Brian McCall, Stephen DesJardins, Julie Posselt, Anna Vignoles, Susan Dynarski, Jake Anders, Tammy Campbell, Peter Keen, attendees at a University of Cambridge-University of Michigan exchange seminar, and two anonymous referees for their valuable feedback and thoughtful comments in the writing of this manuscript. Any errors remain my responsibility alone.

Appendix A: Sensitivity tests around the grade thresholds

This appendix tests the suitability of the Key Stage 3 data for the RDD approach by testing for discontinuities in prior student characteristics around the level 6 cutoff point in each assessment. Tables 6, 7, and 8 test for discontinuities in NSSEC status, and Tables 9, 10, and 11 test for discontinuities in performance at Key Stage 2, as measured by continuous test scores in those assessments.

Judging the models as a group, there is little evidence that students were able to systematically manipulate their position around the cutoff points. Of the 18 models, just one provides an estimate that is significantly different from zero. The significant estimate comes from the model that regresses students' positions around the level 6 threshold of the Key Stage 3 mathematics assessment on their test scores in Key Stage 2 mathematics for students (column (1) of Table 10). This finding seems to be counterintuitive, as it indicates that students scoring just above the level 6 threshold did significantly worse in Key Stage 2 mathematics than those who scored just below this threshold.

All models are created using Calonico et al.'s (2013a) bandwidth selection approach and construction of standard errors. Each model uses a uniform kernel and local-polynomial of order one, i.e., linear. For all tables, standard errors are provided in parentheses. Significance levels are denoted as follows: ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

Appendix B: Specifications for discontinuity models

In this appendix, the following tables provide more detailed information about the RDD models used in Sect. 3.

Table 6 Test for discontinuity in rate of low-NSSEC children by assessment

	(1) English	(2) Mathematics	(3) Science
RD estimate	−0.00808	−0.0163	−0.0110
	(0.0172)	(0.0135)	(0.0158)
Observations	4261	6070	3994

Table 7 Test for discontinuity in rate of mid-NSSEC children by assessment

	(1) English	(2) Mathematics	(3) Science
RD estimate	−0.0268	0.0103	−0.00504
	(0.0273)	(0.0181)	(0.0227)
Observations	3405	6033	4260

Table 8 Test for discontinuity in rate of high-NSSEC children by assessment

	(1) English	(2) Mathematics	(3) Science
RD estimate	0.0502	−0.0244	0.0138
	(0.0381)	(0.0254)	(0.0273)
Observations	2212	4051	3653

Table 9 Test for discontinuity in Key Stage 2 English score by assessment

	(1) English	(2) Mathematics	(3) Science
RD Estimate	−0.156	0.248	−0.0547
	(0.144)	(0.165)	(0.188)
Observations	4147	2744	3698

Table 10 Test for discontinuity in Key Stage 2 math score by assessment

	(1) English	(2) Mathematics	(3) Science
RD estimate	−0.780***	−0.153	−0.193
	(0.227)	(0.170)	(0.156)
Observations	2395	5481	3278

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table 11 Test for discontinuity in Key Stage 2 science score by assessment

	(1) English	(2) Mathematics	(3) Science
RD estimate	−0.138	0.108	0.0143
	(0.152)	(0.156)	(0.110)
Observations	2498	3312	5540

Table 12 Specification for models presented in Table 4, column (1)

	(1) Capped GCSE point score	(2) Attain C or higher in GCSE English	(3) Enrollment in A-levels	(4) Enrollment in university degree
RD estimate	3.741	0.0280	0.0938	0.0389
Observations	5342	3405	2235	3276
conventional S.E.	2.738	0.0250	0.0405	0.0328
Conventional *p* value	0.172	0.262	0.0207	0.236
Robust 95% CI	[−0.79; 11.96]	[−0.01; 0.1]	[0.02; 0.2]	[−0.03; 0.13]
Robust *p* value	0.0858	0.137	0.0128	0.197
Order loc. poly.	1	1	1	1
Order bias	2	2	2	2
BW loc. poly.	3.078	1.868	1.715	3.173
BW bias	5.832	3.607	3.654	5.530
Kernel type	Uniform	Uniform	Uniform	Uniform

Table 13 Specification for models presented in Table 4, column (2)

	(1) Capped GCSE point score	(2) Attain C or higher in GCSE math	(3) Enrollment in A-levels	(4) Enrollment in university degree
RD estimate	4.788	0.0170	0.0369	0.0610
Observations	3803	2751	4707	2368
conventional S.E.	3.852	0.0388	0.0261	0.0375
Conventional *p* value	0.214	0.661	0.158	0.104
Robust 95% CI	[−5.21; 12.44]	[−0.08; 0.08]	[−0.01; 0.11]	[−0.03; 0.14]
Robust *p* value	0.422	0.968	0.134	0.208
Order loc. poly.	1	1	1	1
Order bias	2	2	2	2
BW loc. poly.	2.511	1.698	4.611	2.904
BW bias	4.825	4.089	8.610	5.588
Kernel type	Uniform	Uniform	Uniform	Uniform

Table 14 Specification for models presented in Table 5, column (1) for low-NSSEC students

	(1) Capped GCSE point score	(2) Attain C or higher in GCSE English	(3) Enrollment in A-levels	(4) Enrollment in university degree
RD estimate	20.11	0.207	0.110	0.205
Observations	711	681	949	462
conventional S.E.	8.405	0.0619	0.0538	0.0925
Conventional *p* value	0.0167	0.000842	0.0414	0.0267
Robust 95% CI	[3.03; 42]	[0.11; 0.38]	[0.01; 0.26]	[0.02; 0.47]
Robust *p* value	0.0235	0.000405	0.0411	0.0311
Order loc. poly.	1	1	1	1
Order bias	2	2	2	2
BW loc. poly.	2.387	2.257	3.120	1.831
BW bias	3.917	4.813	5.661	3.243
Kernel type	Uniform	Uniform	Uniform	Uniform

Table 15 Specification for models presented in Table 5, column (2) for low-NSSEC students

	(1) Capped GCSE point score	(2) Attain C or higher in GCSE math	(3) Enrollment in A-levels	(4) Enrollment in university degree
RD estimate	28.21	0.0569	0.0819	0.00816
Observations	965	831	1073	786
conventional S.E.	7.570	0.0657	0.0556	0.0656
Conventional *p* value	0.000195	0.387	0.141	0.901
Robust 95% CI	[13.31; 48.06]	[−0.11; 0.18]	[−0.03; 0.23]	[−0.16; 0.14]
Robust *p* value	0.000536	0.655	0.144	0.857
Order loc. poly.	1	1	1	1
Order bias	2	2	2	2
BW loc. poly.	3.450	2.872	3.923	3.609
BW bias	6.947	6.311	7.162	6.951
Kernel type	Uniform	Uniform	Uniform	Uniform

Table 16 Specification for models presented in Table 5, column (1) for mid-NSSEC students

	(1) Capped GCSE point score	(2) Attain C or higher in GCSE English	(3) Enrollment in A-levels	(4) Enrollment in university degree
RD estimate	−1.189	−0.0202	0.111	0.0152
Observations	454	1059	632	579
conventional S.E.	8.467	0.0347	0.0752	0.0832
Conventional *p* value	0.888	0.561	0.141	0.855
Robust 95% CI	[−24.81; 14.87]	[−0.09; 0.07]	[−0.05; 0.31]	[−0.2; 0.2]
Robust *p* value	0.624	0.892	0.154	0.974
Order loc. poly.	1	1	1	1
Order bias	2	2	2	2
BW loc. poly.	1.437	3.069	1.850	2.007
BW bias	2.637	6.282	3.422	3.369
Kernel type	Uniform	Uniform	Uniform	Uniform

Table 17 Specification for models presented in Table 5, column (2) for mid-NSSEC students

	(1) Capped GCSE point score	(2) Attain C or higher in GCSE math	(3) Enrollment in A-levels	(4) Enrollment in university degree
RD estimate	5.438	0.0553	0.0605	0.144
Observations	910	729	842	600
conventional S.E.	5.873	0.0706	0.0652	0.0707
Conventional *p* value	0.354	0.434	0.354	0.0420
Robust 95% CI	[−8.07; 19.35]	[−0.12; 0.19]	[−0.1; 0.2]	[0.01; 0.33]
Robust *p* value	0.420	0.680	0.530	0.0427
Order loc. poly.	1	1	1	1
Order bias	2	2	2	2
BW loc. poly.	3.425	2.625	3.054	2.611
BW bias	6.034	5.327	5.833	4.888
Kernel type	Uniform	Uniform	Uniform	Uniform

Table 18 Specification for models presented in Table 5, column (1) for high-NSSEC students

	(1) Capped GCSE point score	(2) Attain C or higher in GCSE English	(3) Enrollment in A-levels	(4) Enrollment in university degree
RD estimate	−1.062	0.0459	−0.0680	−0.0301
Observations	1667	1121	1356	1311
conventional S.E.	4.200	0.0370	0.0448	0.0519
Conventional *p* value	0.800	0.215	0.129	0.562
Robust 95% CI	[−11.86; 8.14]	[−0.01; 0.15]	[−0.16; 0.05]	[−0.14; 0.1]
Robust *p* value	0.716	0.101	0.306	0.747
Order loc. poly.	1	1	1	1
Order bias	2	2	2	2
BW loc. poly.	3.375	2.400	2.869	3.147
BW bias	6.153	5.651	4.972	5.426
Kernel type	Uniform	Uniform	Uniform	Uniform

Table 19 Specification for models presented in Table 5, column (2) for high-NSSEC students

	(1) Capped GCSE point score	(2) Attain C or higher in GCSE math	(3) Enrollment in A-levels	(4) Enrollment in university degree
RD estimate	9.494	0.108	0.0935	0.0394
Observations	1140	1142	1150	1342
conventional S.E.	5.163	0.0539	0.0554	0.0477
Conventional *p* value	0.0659	0.0453	0.0914	0.409
Robust 95% CI	[−0.47; 22.52]	[−0.03; 0.21]	[−0.01; 0.24]	[−0.08; 0.14]
Robust *p* value	0.0602	0.131	0.0833	0.585
Order loc. poly.	1	1	1	1
Order bias	2	2	2	2
BW loc. poly.	3.396	3.463	3.485	5.225
BW bias	7.046	6.500	6.673	9.630
Kernel type	Uniform	Uniform	Uniform	Uniform

References

Alcott, B. (2017). The influence of teacher encouragement on educational persistence: evidence from England. *Research in Higher Education*. DOI: 10.1007/s11162-017-9446-2. URL: http://www.readcube.com/articles/10.1007/s11162-017-9446-2?author_access_token=SbOo7dP3dxxSD6ecHyEH9fe4RwlQNchNByi7wbcMAY5pXNWfHkowJP47_MWOrqswW3fAjzfDWq6DTmw9oHtUjGo3yY4KKLd_qi0rXM-Q7vyUsF2fKsQtM3mQwk-oBzPou3N3f2PTyfHaOUllq81LCA%3D%3D (forthcoming)

Alderman, H., Behrman, J. R., Kohler, H. P., Maluccio, J. A., & Watkins, S. (2001). Attrition in longitudinal household survey data: some tests for three developing-country samples. *Demographic research, 5*, 79-124.

Allison, P. D. (2002). Missing data. Thousand Oaks, CA: Sage Publications.

Anders, J. (2012a). The link between household income, university applications and university attendance. *Fiscal Studies, 33*(2), 185–210.

Anders, J. (2012b). *Using the Longitudinal Study of Young People in England for research into higher education access.* Department of Quantitative Social Science-Institute of Education report number 12–13, University of London.

Archer, L., Hutchings, M., & Ross, A. (2003). *Higher education and social class: issues of exclusion and inclusion*. London: Routledge Falmer.

Alcott, B. (2013). *Predicting departure from British education: Identifying those most at risk through discrete time hazard modelling.* Widening Participation and Lifelong Learning. doi:10.5456/WPLL.15.4.46

Azmat, G., & Iriberri, N. (2010). The importance of relative performance feedback information: evidence from a natural experiment using high school students. *Journal of Public Economics, 94*(7), 435–452.

Ball, S. J., Bowe, R., & Gewirtz, S. (1996). School choice, social class and distinction: the realization of social advantage in education. *Journal of Education Policy, 11*(1), 89–112.

Ball, S. J., Davies, J., David, M., & Reay, D. (2002). 'Classification' and 'judgement': social class and the 'cognitive structures' of choice of higher education. *British Journal of Sociology of Education, 23*(1), 51–72.

Ball, S. J., Maguire, M., & Braun, A. (2012). *How schools do policy: policy enactments in secondary schools*. Abingdon: Routledge.

Black, P., & Wiliam, D. (1998). Inside the black box: raising standards through classroom assessment. *Phi Delta Kappa*, 1–13.

Black, P., & Wiliam, D. (2006). The reliability of assessments. In J. Gardner (Ed.), *Assessment and learning* (pp. 119–131). London: SAGE.

Bourdieu, P. (1998). *The state nobility: elite schools in the field of power*. Stanford: Stanford University Press.

Broadfoot, P. (1999). *Empowerment or performativity? English assessment policy in the late twentieth century.* Paper presented at the British Educational Research Association Annual Conference, University of Sussex, 2–5 September.

Callender, C. (2006). The impact of tuition fees and financial assistance. In D. B. Johnstone, M. J. Rosa, H. Vossensteyn, & P. N. Teixeira (Eds.), *Cost-sharing and accessibility in higher education: a fairer deal?* (pp. 105–132). Dordrecht: Springer.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2013a). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal, vv*, ii, 1–34.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2013b). *Robust nonparametric confidence intervals for regression-discontinuity designs.* Retrieved from http://www.iza.org/conference_files/PolicyEval_2013/cattaneo_m9171.pdf.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust data-driven inference in the regression-discontinuity design. *Stata Journal, 14*(4), 909–946.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Chicago: Rand McNally.

Chowdry, H., Crawford, C., Dearden, L., Goodman, A., & Vignoles, A. (2013). Widening participation in higher education: analysis using linked administrative data. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(2), 431–457.

Cochrane, M. (2007). *Spoilt for choice? Pupil perceptions of the options process at Year 9*. Paper presented at the British Educational Research Association Annual Conference, London, 5–8 September. Retrieved from: http://www.leeds.ac.uk/educol/documents/165852.htm.

Cochrane, M. (2011). Children's university aspirations and the effects of cultural and social capital. In J. Adams, M. Cochrane, & L. Dunne (Eds.), *Applying theory to educational research: an introductory approach with case studies* (pp. 95–107). Hoboken: Wiley.

Conner, C. (1999). Through a glass darkly: assessing the future? *Education 3–13, 27*(3), 32–37.

Conservatives (2015). *The Conservative Party Manifesto*. Retrieved from http://www.bond.org.uk/data/files/Blog/ConservativeManifesto2015.pdf.

Cook, T. D. (2008). "Waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics, 142*(2), 636–654.

Daugherty, R. (1995). *National curriculum assessment: a review of policy 1987–1994*. London: Falmer Press.

Davey, G., & Fuller, A. (2013). Hybrid qualifications, institutional expectations and youth transitions: a case of swimming with or against the tide. *Sociological Research Online, 18*(1), 2.

Davies, J., & Brember, I. (1998). National curriculum testing and self-esteem in year 2—the first 5 years: a cross-sectional study. *Educational Psychology, 18*(4), 365–375.

Davies, J., & Brember, I. (1999). Reading and mathematics attainments and self-esteem in years 2 and 6-an eight-year cross-sectional study. *Educational Studies, 25*(2), 145–157.

Department for Business Innovation & Skills. (2012). *Participation rates in higher education: academic years 2006/2007–2010/2011 (provisional)*. London: Department for Business, Innovation & Skills Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/16203/12-p140-participation-rates-in-he-2010-11.pdf.

Department for Education (2010). *Response to freedom of information request. Reference number 2010/0058447*. Retrieved from https://www.whatdotheyknow.com/request/streaming_of_pupils_at_secondary?unfold=1#incoming-105016.

Department for Education (2014). *Schools win funds to develop and share new ways of assessing pupils*. Retrieved from https://www.gov.uk/government/news/schools-win-funds-to-develop-and-share-new-ways-of-assessing-pupils.

DesJardins, S. L., & McCall, B. P. (2014). The impact of the Gates Millennium Scholars Program on college and post-college related choices of high ability, low-income minority students. *Economics of Education Review, 38*, 124–138.

DiNardo, J., & Lee, D. S. (2011). Program evaluation and research designs. *Handbook of Labor Economics, 4*, 463–536.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*(10), 1040–1048.

Dweck, C. S. (2000). *Self theories: their role in motivation, personality and development*. Philadelphia: Psychology Press.

Dweck, C. (2006). *Mindset: the new psychology of success*. New York: Random House.

Giddens, A. (1991). *Modernity and self-identity: self and society in the late modern age*. Stanford: Stanford University Press.

Gillborn, D., & Youdell, D. (2000). *Rationing education: policy, practice, reform, and equity*. Buckingham: Open University Press.

GL assessments (2015). *Life after levels*. Retrieved from http://www.gl-assessment.co.uk/focus/life-after-levels.

Goldstein, H. (2009). Handling attrition and non-response in longitudinal data. *Longitudinal and Life Course Studies, 1*(1), 63–72.

Graham, S. (1990). Communicating low ability in the classroom: bad things good teachers sometimes do. In S. Graham & V. S. Folkes (Eds.), *Attribution theory: applications to achievement, mental health, and interpersonal conflict* (pp. 17–36). Hillsdale: Lawrence Earlbaum.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica, 69*(1), 201–209.

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice, 10*(2), 169–207.

Hodgen, J., & Marshall, B. (2005). Assessment for learning in English and mathematics: a comparison. *Curriculum Journal, 16*(2), 153–176.

House of Commons Children, Schools and Families Committee. (2008). *Testing and assessment: third report of session 2007–2008 (vol. 2)*. London: The Stationery Office Limited.

House of Commons Education Committee. (2013). *From GCSEs to EBCs: the Government's proposals for reform*. London: The Stationery Office Limited Retrieved from http://www.parliament.uk/documents/commons-committees/Education/EIGHTH-REPORT-GCSEs-to-ECBs-Reform-HC-808.pdf.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: a guide to practice. *Journal of Econometrics, 142*(2), 615–635.

Ireson, J., Hallam, S., & Hurley, C. (2005). What are the effects of ability grouping on GCSE attainment? *British Educational Research Journal, 31*(4), 443–458. https://pdfs.semanticscholar.org/2acb/6bff4d1414c743914f522966ae9b71eeb08d.pdf. Access date: 20 June 2015

Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: a regression-discontinuity analysis. *Review of Economics and Statistics, 86*(1), 226–244.

Kellaghan, T., & Greaney, V. (2001). *Using assessment to improve the quality of education*. Paris: Unesco, International Institute for Educational Planning https://pdfs.semanticscholar.org/2acb/6bff4d1414c743914f522966ae9b71eeb08d.pdf.

Kellaghan, T., Madaus, G., & Raczek, A. (1996). *The use of external examinations to improve student motivation*. Washington, DC: AERA

Labour (2015). *The Labour Party Manifesto*. Retrieved from http://www.labour.org.uk/page/-/BritainCanBeBetter-TheLabourPartyManifesto2015.pdf.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281–355.

Leonard, M., & Davey, C. (2001). *Thoughts on the 11 plus*. Belfast: Save the Children.

Liberal Democrats (2015). *Manifesto 2015*. Retrieved from http://www.libdems.org.uk/read-the-full-manifesto.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics, 122*(1), 159–208.

Marcenaro-Gutierrez, O., Galindo-Rueda, F., & Vignoles, A. (2007). Who actually goes to university? *The Economics of Education and Training, 32*(2), 333–357.

McCall, B. P., & Bielby, R. M. (2012). Regression discontinuity design: recent developments and a guide to practice for researchers in higher education. In J. C. Smart & M. B. Paulsen (Eds.), *Higher education: handbook of theory and research* (pp. 249–290). Dordrecht: Springer.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: a density test. *Journal of Econometrics, 142*(2), 698–714.

Mealli, F., & Rampichini, C. (2012). Evaluating the effects of university grants by using regression discontinuity designs. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 175*(3), 775–798.

Molenberghs, G., & Fitzmaurice, G. (2008). In Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Ed.) *Longitudinal data analysis* (pp. 395–408). Boca Raton, FL: CRC Press.

National Audit Office. (2008). *Third validation compendium report (vol. 2)*. London: The Stationery Office Limited.

Office for National Statistics (2014). *The National Statistics Socio-economic Classification (NS-SEC rebased on the SOC2010)*. Retrieved from http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/soc2010-volume-3-ns-sec–rebased-on-soc2010–user-manual/index.html#.

Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for low-performing urban students: evidence from Massachusetts. *Educational Evaluation and Policy Analysis, 32*(1), 5–23.

Paris, S. G., Lawton, T. A., Turner, J. C., & Roth, J. L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher, 20*(5), 12–20.

Piesse, A., & Kalton, G. (2009). A strategy for handling missing data in the Longitudinal Study of Young People in England (LSYPE). DCSF-RW086. London, UK: Department for children, schools and families.

Pollard, A., Triggs, P., Broadfoot, P., Osborn, M., & McNess, E. (2000). *What pupils say: changing policy and practice in primary education*. London: Continuum.

Pugsley, L. (1998). Throwing your brains at it: higher education, markets and choice. *International Studies in Sociology of Education, 8*(1), 71–92.

QCA (2004). *Report on key stage 3 English review of service delivery failure 2003–2004 to QCA Board*. Retrieved from http://dera.ioe.ac.uk/8228/1/10343_ks3_en_report_04.pdf.

Quinlan, M., & Scharaschkin, A. (1999). *National curriculum testing: problems and practicalities. Paper presented at the British Educational Research Association Annual Conference.* Retrieved from http://www.assessnet.org.uk/e-learning/file.php/1/Resources/Assessment_Today/2009_Assessment_Today/060709_Paper.pdf.

Reay, D. (1995). 'They employ cleaners to do that': habitus in the primary classroom. *British Journal of Sociology of Education, 16*(3), 353–371.

Reay, D. (1998). 'Always knowing' and "never being sure": familial and institutional habituses and higher education choice. *Journal of Education Policy, 13*(4), 519–529.

Reay, D., & William, D. (1999). I ll be a nothing : structure, agency and the construction of identity through assessment. *British Educational Research Journal, 25*(3), 343–354.

Sartarelli, M. (2011). *Do performance targets affect behaviour? Evidence from discontinuities in test scores in England*. London: DoQSS Working Paper No. 11–02.

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist, 26*(3–4), 207–231.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sjögren, A. (2010). *Graded children–evidence of longrun consequences of school grades form a nationwide reform*. IFN Working Paper No. 839. Uppsala: Institute for Labour Market Policy Evaluation. http://www.ifau.se/globalassets/pdf/se/2010/wp10-07-Graded-children-evidence-of-longrun-consequences-of-school-grades-from-a-nationwide-reform.pdf. Access date: 28 May 2016

Steedman, C. (1988). The mother made conscious: the historical development of primary school pedagogy. In M. Woodhead & A. McGrath (Eds.), *Family, school & society* (pp. 82–95). London: The Open University.

Taylor, C. (2012). *Improving attendance at school*. London: Department for Education.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: an alternative to the ex post facto experiment. *Journal of Educational Psychology, 51*(6), 309.

Thomas, L., Bland, D., & Duckworth, V. (2012). Teachers as advocates for widening participation. *Widening Participation and Lifelong Learning, 14*(2), 40–58.

UK Department for Education (2014). *England to become a global leader of teaching character*. Retrieved from https://www.gov.uk/government/news/england-to-become-a-global-leader-of-teaching-character.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: a regression–discontinuity approach. *International Economic Review, 43*(4), 1249–1287.

Vignoles, A. (2013). Widening participation and social mobility. In C. Callendar & P. Scott (Eds.), *Browne and beyond: modernizing English higher education* (pp. 112–129). London: Institute of Education Press.

Vignoles, A. F., & Powdthavee, N. (2009). The socioeconomic gap in university dropouts. *The BE Journal of Economic Analysis & Policy, 9*(1), 1-34.

Vincent, C. (2001). Social class and parental agency. *Journal of Education Policy, 16*(4), 347–364.

Vincent, C., Braun, A., & Ball, S. J. (2008). Childcare, choice and social class: caring for young children in the UK. *Critical Social Policy, 28*(1), 5–26.

Walker, M., & Clark, G. (2010). Parental choice and the rural primary school: lifestyle, locality and loyalty. *Journal of Rural Studies, 26*(3), 241–249.

Weiner, B. (2010). The development of an attribution-based theory of motivation: a history of ideas. *Educational Psychologist, 45*(1), 28–36.

Wilshaw, M. (2013). *The annual report of Her Majesty's Chief Inspector of education, children's services and skills 2012/2013*. London: The Stationery Office.