Blissful Ignorance? A Natural Experiment on the Effect of Feedback on Students' Performance*

Oriana Bandiera† Valentino Larcinese‡ Imran Rasul§

February 2012

Abstract

We present a theoretical framework and empirical strategy to measure whether and how providing university students with feedback on their own past exam performance affects their future exam performance. Our identification strategy exploits a natural experiment in a leading UK university where different departments have historically different rules on the provision of feedback to their students. Our theoretical framework makes precise that if feedback provides students with a signal of their marginal return to effort in generating test scores, then the effect of feedback depends on the balance of standard substitution and income effects, and on whether students over or under estimate the return to their effort. Empirically, we find the provision of feedback has a positive effect on student's subsequent test scores: the mean impact corresponds to 13% of a standard deviation in test scores. The impact of feedback is stronger for more able students and for students who have less information to start with, while no students appear to be discouraged by feedback. Our findings add to a growing literature on feedback in organizations more generally, and specifically in this setting the results suggest that the provision of feedback might be a cost-effective means to increase students' exam performance.

Keywords: feedback, incentives, students' performance, university education.

JEL Classification: D82, I23.

*We thank John Antonakis, Martin Browning, Hanming Fan, Nicole Fortin, Joseph Hotz, Michael Kremer, Raphael Lalive, Gilat Levy, Cecilia Rouse, Rani Spiegler, Christopher Taber, and seminar participants at Bocconi, CEU, Duke, Erasmus, IFS, Lausanne, LSE, NY Federal Reserve, Oxford, UBC, Washington, the Wellcome Trust Centre for Neuroimaging, and the Tinbergen Institute Conference for useful comments. We thank all those at the university involved in providing the data. This paper has been screened to ensure no confidential information is revealed. All errors remain our own.

†Department of Economics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom; Tel: +44-207-955-7519; Fax: +44-207-955-6951; E-mail: o.bandiera@lse.ac.uk.

‡Department of Government, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom; Tel: +44-207-955-6692; Fax: +44-207-955-6352; E-mail: v.larcinese@lse.ac.uk.

§Department of Economics, University College London, Drayton House, 30 Gordon Street, London WC1E 6BT, United Kingdom. Telephone: +44-207-679-5853; Fax: +44-207-916-2775; E-mail: i.rasul@ucl.ac.uk.

1 Introduction

This paper exploits a natural experiment in a leading UK university to shed light on whether providing university students with feedback on their own past exam performance affects their future exam performance. A growing literature in economics has begun investigating the causes and consequences of feedback provision. Much of this research has focused on the theory of optimal feedback provision [Lizzeri *et al.*. 2002, Ertac 2006, Ederer 2010], while empirical research has so far focussed on the impact of relative performance feedback [Hannah *et al.* 2008, Eriksson *et al.* 2009, Azmat and Iriberri, 2010].

Recent student surveys among undergraduates in the UK have consistently shown that in spite of being satisfied overall with their learning experience, most students demand more feedback from their faculty teachers. While more than 80% of respondents would agree that "teachers are are good at explaining things" and that they are "enthusiastic about what they are teaching" only about half of respondents would agree that "feedback on my work has helped me clarify things I did not understand" or that "feedback on my work has been prompt".[1] Since there are considerable opportunity costs of faculty members devoting time to providing students' feedback, it is important to first establish what is the causal impact of feedback on students' performance before resources could optimally be targeted to the provision of more feedback.

In this paper we develop a theoretical framework and empirical strategy to identify the causal effect of feedback on students' performance in a leading UK university. We use administrative records on the performance of 7,738 students enrolled full time on one-year graduate (M.Sc.) degree programs, over the academic years 1999/00-2003/04. The academic year in this university can be divided into two periods, and students are evaluated at the end of each. The combination of period one and period two test scores yield the individual's final M.Sc. degree classification.

The natural experiment we exploit relies on the fact that different university departments have historically different policies regarding the provision of feedback to their students. The feedback policies we observe differ according to whether individual academic performance in period one is revealed to students at the start of period two or not. Some departments provide students with their individual period one test score *before* they begin exerting effort towards their period two test score. We refer to these departments as being in the 'feedback regime'. Other departments only reveal test scores from period one at the end of the academic year, after period two test scores have also been generated. We refer to these departments as being in the 'no-feedback regime'.

A key feature of our setting is that we observe the performance of the same student in the same department before and after feedback is provided. This allows us to identify the effect of feedback from the difference-in-difference between period one and period two exam performance of students in different feedback regimes, thus controlling for time invariant unobserved sources of department and student heterogeneity that might create a spurious correlation between feedback

[1] Data and methodology are available at http://unistats.direct.gov.uk/.

regime and exam performance.

Our theoretical model makes precise the effect of feedback on period two performance and how this is heterogeneous across students. A defining feature of the university we study is that the majority of students are new to the institution, and to the UK education system as a whole. Hence they are unfamiliar with this university's style of teaching, grading, examinations, essay writing, and are uncertain about the production function that translates effort into test scores. This suggests that feedback on period one test scores can have a first order effect of providing a signal for the marginal return to effort in period two. We model this intuition by assuming that in period one all students are uncertain about the returns to their effort, and that students in the feedback regime receive a perfect signal of their return before choosing period two effort, whereas students in the no-feedback regime do not.

This framework is very similar to a standard labor supply model where the return to effort is akin to the wage rate, and effort is akin to hours worked. Hence changes in the return to effort will produce both 'income' and 'substitution' effects. These potentially offsetting effects will be one familiar source of heterogeneous responses to feedback. However, our framework departs from the standard labor supply model in that, with feedback, individuals update their prior beliefs as they learn their true return to effort. Hence a second source of heterogeneous responses will be individuals' prior beliefs regarding their return to effort. As discussed throughout, this relates closely to whether students over- or under-estimate the returns to effort in the absence of feedback.

The empirical analysis proceeds in three stages. First, we measure the average effect of feedback on individual test scores. As we observe the same student in both periods, we estimate the effect of feedback on the difference in performance of the *same* student across periods and feedback regimes, conditional on time invariant unobserved heterogeneity across students and departments. The identifying assumption is that the choice of feedback policy is orthogonal to unobservables that cause systematic differences in test scores across periods. Second, we test whether, in line with our assumptions, the effect of feedback depends on the informativeness of the signal and is heterogeneous across individuals. Third, we tie the empirical analysis to the theory by making precise which combinations of preferences and prior beliefs can be reconciled with the evidence.

Our main results are as follows. First, controlling for unobserved heterogeneity across students and departments, the difference-in-difference in test scores across periods and feedback regimes is significantly greater than zero. The magnitude of this effect corresponds to 13% of a standard deviation in the difference in test scores across periods in the no-feedback regime. The implied effect size of feedback is at the lower end of estimates of the effect size on test scores of both class sizes in primary, secondary, and tertiary education [Angrist and Lavy 1999, Krueger 1999, Bandiera *et al.* 2010], and of the effect size of teacher quality on test scores [Aaronson *et al.* 2007, Rockoff 2004, Rivkin *et al.* 2005]. It is reasonable to suppose a one standard deviation change in class sizes or teacher quality will be orders of magnitude more costly to achieve than the communication of feedback to students. Hence the provision of feedback appears to be a relatively

cost-effective means by which to raise students' learning and achievement in university education.

Second, on heterogeneous impacts we find that, in line with our assumption that feedback acts as a signal for an unknown parameter of the production function for test scores, the effect of feedback is entirely driven by students who are new to this institution. Moreover, quantile regression analysis reveals that the provision of feedback has a close to zero effect on students at the left tail of the return to effort distribution, has a significantly larger effect on students at the middle of the return to effort distribution, and has the most pronounced effect on students at the right tail of the return to effort distribution.

Finally, we exploit the fact that some students take period one courses in departments other than the one they are enrolled in to devise a placebo test that allows us to disentangle the effect of feedback from the effect of having period two scores assigned by a department that has chosen to provide feedback. Reassuringly, we find that students' period two performance is only affected by the actual feedback received, and not by the feedback policy of the department they are enrolled in. Taken together, the results of the placebo test and the heterogeneous impacts described above, are in line with our identifying assumption. Therefore, while we cannot, and do not, claim that the choice of feedback policy is randomly assigned across departments, the evidence suggests that departmental policies over the provision of feedback by departments are more accidents of history and therefore orthogonal to unobservables that cause systematic differences in test scores across periods, rather than being chosen as a endogenous response to how test scores are generated across periods in this setting.

In a closely related paper, Azmat and Iriberri [2010] provide evidence on the impact of relative performance feedback in a high school setting. When provided with information about both their own grades and the average grade in the class, the average impact across students was an increase in grades of 5%. Like us, they find that the effect is positive for all students. Differently from us, they find a strong effect for students at the bottom of the grade distribution, which could be a consequence of relative comparisons which are absent in our setting. Finally, they find that the positive effect is concentrated in Level 1 (the first year) and Level 4 (the last). While the last year effect may be driven by the importance of signaling for university admission, a positive effect in the first year (and not in the second and third year) is compatible with the idea that feedback helps students to evaluate the marginal return to their effort: compared with students in Level 2 and 3, first year students have less information about both their skills and their relative position in the class.

The paper is organized as follows. Section 2 develops a stylized model of students' behavior through the academic year and allows us to compare outcomes across feedback regimes. Section 3 describes the empirical setting and administrative data. Section 4 presents the empirical analysis and Section 5 concludes with a discussion of the external validity of our results. The Appendix tests alternative explanations and reports further robustness checks.

2 Theoretical Framework

2.1 Setup

We develop a model of students' behavior tailored to our empirical setting. We assume the academic year consists of two periods, $t = 1, 2$. At the beginning of each, students choose the effort to exert, e_t, in producing a test score or grade, $g_t(e_t)$. Students have a per period time endowment T_t to allocate between effort and leisure, l_t. Utility is assumed to be increasing in per period test scores and leisure. The student's optimization problem is,

$$\max_{e_1,e_2} \quad U(g_1(e_1), g_2(e_2), l_1, l_2) \; subject \; to \; l_t + e_t = T_t \; for \; t = 1, 2. \tag{2.1}$$

Feedback regimes differ according to whether period one performance, g_1, is revealed at the start of period two or not. Intuitively, for feedback to have an effect on the choice of effort two conditions must hold – (i) in the absence of feedback students do not know g_1, i.e. there is uncertainty in the production function that transforms e_1 into g_1; (ii) period one performance affects the marginal return to period two effort. Economic theory highlights two broad classes of mechanisms through which the latter can occur. First, period one and period two performance might be complements or substitutes in the utility function, so that feedback on g_1 affects the marginal utility of g_2 (and hence the marginal return to e_2) directly through a preference effect [Lizzeri *et al.* 2002]. This would occur if, for instance, students have a final average test score target, so that the marginal utility of g_2 is decreasing in g_1. Second, period one performance can act as a signal for an unknown parameter of the grade production function, so that feedback on g_1 affects the marginal utility of g_2 indirectly, by providing information on the marginal return to effort [Aoyagi 2007, Ederer 2010].

In our empirical setting, the signaling mechanism is key because a defining feature of the university we study is that the majority of students are new to this institution, and to the UK education system as a whole. Only 17% of students were undergraduates at the same university, and a similarly low proportion are British citizens.[2] Overall, the majority of students are unfamiliar with this university's style of teaching, grading, examinations, essay writing, and generally what is expected of students. In short, students are initially uncertain over the production function governing the relationship between effort and test scores. We capture this by assuming that test scores are a deterministic function of period specific effort e_t and a parameter a, through the following production function, $g_t = ae_t$. This multiplicative specification captures that the returns to effort depend on a parameter, which is unknown at the beginning of period one, although individuals know it is drawn from some continuous distribution, $f(a)$. The parameter a can be interpreted as the individual student's return to effort, so that for a given level of effort a more

[2]In our administrative records, in the feedback regime students come from 110 countries and 657 undergraduate institutions. In the no-feedback regime, students come from 132 countries and 851 undergraduate institutions.

able student performs better.

At the start of period two, students' beliefs on the returns to their effort depend on whether they receive feedback on their period one test score. If feedback is provided so that at the end of period one g_1 is revealed, then given the deterministic production function, such feedback is a fully informative signal of the student's return to effort, a. Hence, with feedback, the student can choose e_2 knowing a. If no feedback is provided, both g_1 and g_2 are revealed at the end of the second period which implies students choose e_2 based on their prior belief about the return to effort.[3] Of course, if the production function for generating test scores across periods is very different or rely on different types of effort, then even if an individual receives feedback on the return to effort for period one test scores, this will have little or no effect on period two scores. Whether feedback is effective is therefore ultimately an empirical question. In terms of the stylized model we present here, even if the test scores in the two periods were generated by using different production functions, our predictions would remain the same as long as the marginal return to effort in period one is positively correlated with that in period two.

For analytical clarity we assume that utility is additively separable across the time periods so $U(.) = u_1(g_1, l_1) + u_2(g_2, l_2)$ where $u_1(.)$ and $u_2(.)$ are continuous and concave in each argument. This assumption effectively precludes the possibility that feedback affects effort through preferences and focuses the analysis on the effect of feedback as a signal. Section 2.3 below discusses how the assumption of separable utility provides empirically testable implications, which are then taken to the data in the Appendix.

It is useful to note that our framework is very similar to a standard labor supply model where the return to effort is akin to the wage rate, and effort is akin to hours worked. Hence there will be 'income' and 'substitution' effects. These potentially offsetting effects will be one familiar source of heterogeneous responses to feedback. However, our framework departs from the standard labor supply model in that, with feedback, individuals update their priors as they learn their true return to effort. Hence a second source of heterogeneous responses will be individuals' prior beliefs regarding their return to effort. As discussed throughout, this relates closely to whether students are initially optimistic or pessimistic with regards to their true return to effort.

We now solve the model for period one and two efforts in both feedback regimes. We then compare effort levels across feedback regimes to generate testable implications for how the provision of feedback affects effort and hence grades, and to make precise why and how the provision of feedback can have heterogeneous effects across students.

[3]Two points are of note. First, the results are robust to assuming that feedback is an imperfect signal for a [Aoyagi 2007, Ederer 2010]. Second, we are assuming individuals are not subject to hindsight or recall bias over their period one effort and so update their beliefs regarding their true return to effort. Evidence of hindsight biases in other contexts was first presented by Fischhoff and Beyth [1975].

2.2 Second Period Effort: The Effect of Feedback

If feedback on g_1 is provided, then individuals know their true return to effort when choosing period two effort. Hence the student's period two optimization problem is, $\max_{e_2} u_2(e_2 a\ , T_2 - e_2)$, yielding the first order condition,[4]

$$\frac{\partial u_2/\partial l_2}{\partial u_2/\partial g_2} = a. \tag{2.2}$$

If no feedback on g_1 is provided then individuals do not know their true return to effort when choosing e_2 but know it is drawn from some continuous distribution, $f(a)$. Hence the student's period two optimization problem is,

$$\max_{e_2} u_2(e_2 a\ , T_2 - e_2) f(a) da. \tag{2.3}$$

The weighted mean value theorem for integrals then guarantees there exists an $\widehat{a}$ in the support of a such that,

$$\int u_2(e_2 a\ , T_2 - e_2) f(a) da = u_2(e_2 \widehat{a}, T_2 - e_2). \tag{2.4}$$

Hence in the no-feedback regime, individuals behave *as if* their return to effort is $\widehat{a}$. The first order condition for effort in the no-feedback regime is then,

$$\frac{\partial u_2/\partial l_2}{\partial u_2/\partial g_2} = \widehat{a}. \tag{2.5}$$

Denoting the optimal second period effort function in the feedback regime as $e_2^F(.)$, and that under the no-feedback regime as $e_2^{NF}(.)$, it then follows that $e_2^{NF} = e_2^F(\widehat{a})$. Other things equal, the comparison of effort choices in the two regimes thus depends on: (i) whether $a \gtreqless \widehat{a}$, and, (ii) the sign of de_2/da. The first factor relates to whether, in the absence of feedback, a student whose return to effort is a behaves as if her return to effort is higher or lower than $\widehat{a}$.

The second factor depends on the balance of the standard income and substitution effects. To see these effects we totally differentiate the first order condition to derive the standard Slutsky decomposition,[5]

$$\frac{de_2}{da} = \frac{\partial u_2/\partial g_2}{P} + e\frac{a\partial^2 u_2/\partial g_2^2 - \partial^2 u_2/\partial l_2 \partial g_2}{P}, \tag{2.6}$$

where $P > 0$. This highlights that two effects are at play. The first is a substitution effect, that is, holding total utility constant, an increase in return to effort implies it is more costly for the individual to engage in leisure and she will therefore increase her effort, all else equal. We refer to

[4]This first order condition is analogous to that in the standard labor supply model, so that the marginal rate of substitution between leisure and grades is set equal to their ratio of relative prices, where the price of grades is normalized to one. This makes clear that an alternative interpretation of the model is that students are learning the marginal rate of substitution between their effort and leisure. This is reasonable given that many of them are new to the UK so they may be unaware of the local amenities on offer or how much they value them.

[5]Note that $P = -\left[a^2 \frac{\partial^2 u_2}{\partial g_2^2} - a\frac{\partial^2 u_2}{\partial l_2 \partial g_2} - \frac{\partial^2 u_2}{\partial l_2 \partial g_2} a + \frac{\partial^2 u_2}{\partial g_2^2}\right]$, and this is guaranteed to be positive.

this effect as the "motivation effect". The second term captures the effect that, assuming leisure is a normal good, an increase in return to effort makes it possible to reach the same grade with lower effort, all else equal. We call this the "slacker effect". The net impact of a on e_2 is therefore theoretically ambiguous and depends on which effect dominates overall.

Given that in the no-feedback regime, $e_2^{NF} = e_2^F(\widehat{a})$, the difference in efforts under the regimes is $e^F(a) - e^F(\widehat{a})$, the sign of which depends on whether $a \gtreqless \widehat{a}$, namely whether the student is pessimistic or optimistic with regards to the returns to effort, and whether, at a, the motivation effect dominates the slacker effect or *vice versa*. Table 1 summarizes the four possible combinations and highlights how the effect of feedback is heterogeneous across students depending on their initial level of pessimism/optimism, and their preferences.

2.3 First Period Effort: Testable Implications of Separable Utility

We now turn to consider the effect of feedback on behavior in period one. In both the feedback and no-feedback regimes, in period one students choose their effort according to their prior beliefs so the first order condition for e_1 under both regimes is,

$$\frac{\partial u_1/\partial l_1}{\partial u_1/\partial g_1} = \widehat{a}. \tag{2.7}$$

Hence in this framework the provision of feedback has no effect on period one effort [Aoyagi 2007, Ederer 2010]. This is due to two assumptions. First, the individual's utility function is additively separable. Second, the precision of the signal on the return to effort is independent of the level of effort e_1. If either of these assumptions were not satisfied and students knew whether feedback would be provided, then the *anticipation* of feedback would affect behavior in period one [Lizzeri *et al.* 2002].

More importantly, this suggests an empirical test for the assumption of separable utility and hence the existence of preference effects of feedback in this setting. In particular, if utility is non-separable, the anticipation of feedback necessarily affects period one effort because the second period decision would enter in (2.7) [Lizzeri *et al.* 2002]. In the Appendix we test whether period one effort varies across feedback regimes to shed light on the validity of the assumption of separability.

3 Context and Data Description

3.1 Institutional Setting

Our analysis is based on the administrative records of individual students from a leading UK university. The UK higher education system comprises three tiers – a three-year undergraduate

degree, a one or two-year M.Sc. degree, and Ph.D. degrees of variable duration. Our working sample focuses on 7,738 students enrolled full time on one-year M.Sc. degree programs, over academic years 1999/00-2003/04. These students will therefore have already completed a three year undergraduate degree program at some university and have chosen to stay on in higher education for another year.[6]

We have data for 20 out of the 22 academic departments in the university all of which relate to disciplines in the social sciences. Together these departments offer around 120 M.Sc. degree programs. Students enroll onto a specific degree program at the start of the academic year and cannot change program or department thereafter. Each degree program has its own associated list of core and elective courses. Electives might also include courses organized and taught by other departments. For instance, a student enrolled on the M.Sc. degree in economics can choose between basic and advanced versions of the core courses in micro, macro, and econometrics, and an elective course from a list of economics fields and a shorter list from other departments.

Over the academic year each student must obtain a total of four credits. The first three credits are obtained upon successful completion of final examinations related to taught courses. As each examined course is worth either one or half a credit, the average student takes 4.4 examined courses in total. The fourth credit is obtained upon the subsequent completion of a research essay, typically 10,000 words in length. In all degree programs, the essay is worth one credit, is compulsory, and must relate to the primary subject matter of the degree program.

Students have little contact with faculty while working on their essay. While they typically have one meeting with a faculty member, whose role is to approve the essay topic, it is not customary for students and this faculty member to meet thereafter. There is, for example, no requirement for faculty members to meet with students while they are writing their essay. Upon completion, the essay is double marked by two faculty members, neither of whom is the faculty member that previously approved the essay topic. There are no seminar presentations during which students receive feedback on their essays. Finally, the essay is not accorded more weight than examined courses in the overall degree classification. It is for example, possible to be awarded a degree even if the student fails the long essay, if their examined course marks are sufficiently high.

3.2 Test Scores

Our main outcome variable is the test score of student i in her examined courses and the long essay.[7] Test scores are scaled from 0 to 100 and these translate into classmarks as follows: an

[6]Students are not restricted to only apply to M.Sc. degree programs in the same field as that in which they majored in as an undergraduate. In addition, the vast majority of M.Sc. students do not transfer onto Ph.D. programs at the end of their M.Sc. degree.

[7]The incidence of dropping out is extremely rare in this university. In our sample period we observe less than 1% of students enrolling onto degree programs and then dropping out before the examinations in June. We observe 0.6% of students taking their examined courses and dropping out before completing the long essay. These dropouts have no worse exam marks on average, and are equally likely to come from either feedback regime. Hence it is

A-grade corresponds to test scores of 70 and above, a B-grade to 60-69, a C-grade to 50-59, and a fail to 49 or lower. In this setting test scores are a good measure of students performance and learning for two reasons.

First, test scores are not curved so they reflect each individual's absolute performance on the course. Guidelines issued by each department indeed illustrate that grading takes place on an absolute scale.[8] The distribution of grades within a course-year reveals that, in line with absolute marking, there are some courses on which *all* students obtain the same classmark. On some courses this is because all students obtain a B-grade, and on other courses all students achieve an A-grade. In 23% of course-years, not a single student obtains an A-grade. In addition, the average classmark varies widely across course-years and there is no upper or lower bound in place on the average grade or GPA of students in any given course-year.[9]

Second, there are no incentives for faculty to strategically manipulate test scores to boost student numbers or to raise their own student evaluations [Hoffman and Oreopoulos 2009]. Indeed student numbers do not affect the probability of the course being discontinued and students fill in teaching evaluations two months before sitting the exam.[10] In addition, manipulating test scores is difficult to do as exam scripts are double blind marked by two members of faculty, of which typically only one teaches the course.

3.3 Timing and Feedback Regimes

Figure 1 describes the timing of academic activities for students over the year. Students enroll onto their degree program in October. Between October and May they attend classes for their taught courses. These courses are all assessed through an end of year sit down examination. All exams take place over a two week period in June. The long essays are then due in September, and final degree classifications of distinction, merit, pass, or fail, are announced thereafter.

The key feature of this setting is that departments have historically different policies over whether students are given feedback on their individual performance in the examined courses *before* they begin working on their essay. We observe two feedback regimes. There are 14 departments that only reveal their exam and essay grades at the end of the 12 month period, after the essay due date has passed. We refer to these departments as being in the 'no-feedback regime', as shown

unlikely that they leave because they are discouraged by feedback.

[8]For example, the guidelines for one department state that an A-grade will be given on exams to students that display "a good depth of material, original ideas or structure of argument, extensive referencing and good appreciation of literature", and that a C-grade will be given to students that display "a heavy reliance on lecture material, little detail or originality".

[9]An alternative check on whether test scores are curved is to test whether the mean score on a course-year differs from the mean score across all courses offered by the department in the same academic year. For 29% of course-years we reject the hypothesis that the mean test score is equal to the mean score at the department-year level. Similarly for 22% of course-years we reject the hypothesis that the standard deviation of test scores is equal to the standard deviation of scores at the department-year level.

[10]More precisely, we find that the existence of course c in academic year t does not depend on enrollment on the course in year $t-1$, controlling for some basic course characteristics and the faculty that taught the course in $t-1$.

in the upper half of Figure 1. In contrast, there are 8 departments that provide each student with her own examined course marks after all the exams have been taken and graded, and typically, before students start exerting effort towards their essay. We refer to these departments as being in the 'feedback regime', as shown in the lower half of Figure 1. All degree programs within the same department are subject to the same feedback policy.[11] *A priori*, theory suggests the effects of feedback are not uniform across individuals and so we would not therefore expect one regime to always be preferred to the other.

To identify which department belonged to which regime, we conducted our own survey of heads of department. Most heads reported the feedback rules as having been in place for some time and rarely updated.[12] In line with this, Table A1 shows that the feedback regime appears orthogonal to most departmental characteristics, with any reported differences being quantitatively small. Importantly, there is little evidence to suggest departments in the feedback regime systematically display more 'pro-feedback' attitudes in general, nor that departments in the no-feedback regime systematically try to compensate for this lack of feedback along other margins. Using information from student evaluations, that are conducted in April before the final examinations take place, we find that students are equally satisfied with their courses in both regimes.[13]

As made precise below, however, our identification strategy does not require feedback policy to be orthogonal to other determinants of test scores. Indeed, we exploit information on the test score of the *same* student in their examined and essay courses, and then estimate the causal effect of feedback on the difference in test scores across these two types of course. The estimates therefore take account of unobserved heterogeneity across students, across degree programs, and across departments.

Three further points are of note. First, since test scores are not curved, each student in the feedback regime receives information about her absolute performance in all courses, rather than her relative standing compared to her classmates.[14] Second, at the time when students begin working on their essays, the information available to faculty in both regimes is identical – faculty

[11] A small fraction of courses are partially assessed through coursework, although typically no more than 25% of the final mark stems from the coursework component. Although this coursework is conducted through the year, feedback is not provided before the final exam.

[12] We also note that the same departments in a neighboring university provide different feedback to their students, than the corresponding department in the university we study. This further suggests departments do not set their feedback policies predominantly on the basis of the characteristics of their students.

[13] Departments in the feedback regime appear smaller with fewer teaching faculty and larger class sizes on average. To the extent that larger class sizes are detrimental to student performance, we would expect their performance on examined courses to be lower, all else equal, which is easily checked for. There is no difference in the length of long essays, as measured by the word limit, required across feedback regimes. The behavior of faculty and tutorial teachers along a number of margins is not reported by students to differ between regimes. The only significant differences (at the 9% level) are that tutorial teachers are more likely to be well prepared in no-feedback departments, and more likely to return work within two weeks in feedback departments. Finally we note that all the departments we analyze are related to disciplines in the social sciences and humanities. This reduces the likelihood that very different skills are required to produce test scores across departments or regimes.

[14] Much of the theoretical analysis of feedback has assumed a tournament structure where agents receive feedback on the performance of all participants [Aoyagi 2007, Gershkov and Perry 2009, Ederer 2010].

always have the opportunity to find out the test scores of individual students and to act on this information. Key to identification is that faculty do not behave systematically differently in this regard on the basis of whether feedback is formally provided to the student or not. Third, students are provided with various types of feedback throughout the academic year, and this is true in both feedback regimes. For example, students hand in and receive back graded work during class tutorials, receive informal feedback from faculty, and mock exams take place in all departments (although there are no mid term exams in any department). In addition, students always have some subjective feedback, having sat their exams. The key difference across the regimes is that students in the feedback regime are provided with precise feedback on how well they actually performed on their examined courses. Our empirical analysis therefore measures the effect of this objective feedback on students' subsequent effort, conditional on all other forms of feedback students receive or seek throughout the academic year.

3.4 Descriptive Evidence

In the empirical analysis, period one covers the first nine months of the academic year during which students exert effort on their examined courses. Period one performance is measured by the student's average test score obtained on her examined courses. Period two corresponds to when students exert effort on their essay. Period two performance is measured by the student's essay test score. The primary unit of analysis is student i, enrolled on a degree program offered by department d, in time period $t = 1, 2$.

Table 2A provides descriptive evidence on test scores by period and feedback regime. The first panel shows the mean test score in examined courses for students enrolled in the no-feedback regime is 62.2. This is not significantly different to the mean test score for students in the feedback regime. Moreover, the standard deviation of period one test scores is no different across feedback regimes, as is confirmed in Figure 2A which shows the unconditional kernel density estimate of period one test scores by feedback regime.[15]

This descriptive evidence suggests – (i) any differences in departmental characteristics documented in Table A1 do not on average translate into different examined test scores; (ii) the anticipation of feedback does not affect behavior in period one, consistent with preferences being additively separable or students not knowing that feedback is provided.

The next rows of Table 2A present similar descriptive evidence for period two test scores across feedback regimes. Test scores are significantly higher in the feedback regime. Moreover, the standard deviation in test scores is also significantly higher in the feedback regime. The kernel density estimate in Figure 2B shows the dispersion increases because the distribution of test scores

[15]The left tail of the kernel density shows that we observe a low failure rate – 3% of observations at the student-course-year level correspond to fails. This is in part because degree programs in this university are highly competitive to enter as shown in Table A1. In the no-feedback regime, even such failing students are not informed that they have failed their examined courses.

becomes more left skewed. Students in the middle and right tail of the distribution of test scores appear to be positively affected by the provision of feedback, with little effect – either positive or negative – on students at the left tail of the test score distribution.

Table 2B presents the unconditional estimates of the effect of feedback on test scores across periods. Two points are of note. First, regardless of the feedback regime, test scores are always significantly higher in period two than period one – this likely reflects that the production function for generating test scores is different for essays than for examined courses. Second, the difference in test scores across periods is significantly higher in the feedback regime. The difference-in-difference (DD) is .760, significant at the 1% level, and corresponds to 13% of a standard deviation in the difference in test scores across periods in the no-feedback regime.[16]

4 Empirical Analysis

4.1 Method

We estimate the following panel data specification for student i, enrolled on a degree program offered by department d, in time period t,

$$g_{idt} = \alpha_i + \beta \left[F_d \times T_t\right] + \gamma T_t + \sum_{d'} \mu_{d'} T D_{id'} + \varepsilon_{idt}, \tag{4.1}$$

where g_{idt} is her test score, and α_i is a student fixed effect capturing time invariant characteristics of the student that affect her test scores equally across time periods, such as her underlying motivation and t, and labor market options upon graduation. Since each student can only be enrolled in one department or degree program, α_i also captures all department and program characteristics that affect test scores in both periods, such as the quality of teaching and grading standards. F_d is a dummy variable equal to one if department d is in the feedback regime, and zero if department d is in the no-feedback regime. T_t is a dummy variable equal to one in period two corresponding to the essay, and is equal to zero in period one corresponding to the examined courses.

The parameter of interest, β, measures the effect of feedback on the within student difference in test scores over periods. This is consistently estimated if $cov(F_d \times T_t, \varepsilon_{idt}) = 0$, hence identification concerns stem from the possible existence of unobservables that are correlated to the feedback regime in place and cause there to be differences in test scores across periods. These might be at the student level, e.g. if students who are better at essay writing sort into departments that provide feedback, or at the department level, e.g. if departments that provide feedback are also

[16] An alternative metric by which to gauge the effect of feedback is in terms of the probability a student obtains a test score of 70 and above, corresponding to an A-grade. On this metric, we find the difference-in-difference estimate to be .049, relative to baseline probability of .168 in the no-feedback regime.

more generous when marking research essays. We present evidence to allay these concerns in Sections 4.3 to 4.5 below.

The parameter γ captures any level differences in test scores across periods that may arise from there being different production technologies for generating test scores in examined courses and essays, as Table 2B suggests. The inclusion of student fixed effects α_i do not allow us to estimate the direct effect of feedback on test scores because a given student can only be enrolled in one department and the feedback rule is department specific.

To account for differences in test scores due to students taking courses in other departments, we control for a complete series of departmental dummies – $TD_{id'}$ is equal to one if student i takes any examined courses offered by department d' and is zero otherwise. Finally, ε_{idt} is a disturbance term which we allow to be clustered at the program-academic year level to account for common shocks at this level to all students' test scores, such as those arising from the degree program's admission policy in each academic year.

4.2 Baseline Estimates

Table 3 presents the results. The first specification in Column 1 controls for student characteristics rather than student fixed effects, α_i, to allow us to estimate the direct effect of feedback on test scores. The student's characteristics are gender, whether she attended the same university as an undergraduate, whether she is a student registered for UK, EU, or outside EU fee status, and the academic year of study. The result shows that – (i) the DD in test scores across periods and feedback regimes is $\widehat{\beta} = .882$, which is significantly different from zero and slightly larger than the unconditional DD reported in Table 2; (ii) consistent with Table 2, test scores are always significantly higher for essays than for examined courses, $\widehat{\gamma} = 1.28$; (iii) there is no significant difference in period one test scores between students in the feedback and no-feedback regimes.

Column 2 shows these results are robust to controlling for a complete series of dummies for each teaching department the student is exposed to. In the next two Columns we control for departmental or degree program fixed effects and so we can no longer estimate the effect of feedback on period one test scores as the feedback regime does not vary within a department, hence, *a fortiori*, within a program. In Columns 3 and 4 we see that the previous results are robust to conditioning out unobserved heterogeneity across enrollment departments or degree programs. This casts doubt on the concern that $\widehat{\beta}$ merely captures differences in grading styles between exams and essays that are correlated with the provision of feedback.

Column 5 presents estimates of the complete specification in (4.1). Controlling for unobserved heterogeneity across students we find the DD in test scores across periods and feedback regimes to be positive and significant at .797. Moreover, the point estimate on the difference in test scores across time periods is smaller than in the previous specifications, suggesting that differences in production technology between examined courses and essays are less pronounced when we account

for heterogeneity across students. This specification shows that over 60% of the variation in test scores is explained by the inclusion of student fixed effects, suggesting considerable underlying heterogeneity in student return to effort. We exploit this variation later to shed light on heterogeneous responses to feedback and to tie together the evidence with theory.[17]

We note the implied effect size of feedback is at the lower end of estimates of the effect size on test scores of both class sizes [Angrist and Lavy 1999, Krueger 1999] and teacher quality [Aaronson *et al.* 2007, Rockoff 2004, Rivkin *et al.* 2005] at other tiers of the educational system. Perhaps the most relevant benchmark for comparison is the class size effect we have estimated using exactly these administrative records. More precisely, in Bandiera *et al.* [2010] we document the existence of a robust and significant class size effect size of $-.10$ in this university setting. A one standard deviation change in class sizes (or teacher quality) might be orders of magnitude more costly to achieve than the communication of feedback to students. Hence the provision of feedback appears to be a relatively inexpensive way to raise students' learning and achievement.

The final two columns provide alternative estimates to assess the magnitude of the effect. In particular we show the provision of feedback – (i) significantly increases overall GPA scores (Column 6); (ii) significantly increases the likelihood of obtaining an A-grade over time periods (Column 7). In the latter case the estimates imply a student is 3.8% more likely to obtain an A-grade on her long essay with feedback, relative to a baseline probability of 17% of obtaining an A-grade on the essay in the absence of feedback.[18]

4.3 Mechanisms

To assess the relevance of any signaling effects of feedback we identify subsets of students for whom the value of feedback differs *a priori*. To do so, we use the intuition that students who have previously been undergraduates at the same institution for three years, are more familiar with the institution's style of teaching, grading, examinations, and essay writing. Hence the value of feedback for such individuals, in terms of what they learn about the true returns to their effort, should be lower than for individuals new to the institution, all else equal.

Columns 1 and 2 of Table 4 separately estimate (4.1) for individuals who were previously undergraduates at the same institution, and for those who were not. The results show the provision of feedback only affects those students who are new to the institution, the parameter of interest β is indeed close to zero and precisely estimated in the sample of students with previous experience at

[17] To keep consistency with the theory, we take the mean of the students' exams grades and thus have only one observation for period one. Since the data is at the student-course level, we can also estimate a specification at the student-course level. Results are qualitatively similar and available from the authors upon request.

[18] One concern with the linear specification in (4.1) is that it may be inappropriate to model a dependent variable that is bounded above, and that it does not recognize that a five point increase in test score from 70 is not equivalent to such an increase from 50 in terms of student's effort. To address this we re-scale test scores as $\widetilde{g}_{idt} = -\ln(1-(g_{idt}/100))$ so that we estimate the effect of feedback on the proportionate change in test scores. In this specification, $\widehat{\beta} = .025$ and is significantly different from zero.

the same institution. This suggests one reason *why* feedback matters is because it helps individuals learn the returns to their effort. If, for example, feedback only mattered because it allowed individuals to tailor their second period effort due to complementarities or substitutabilities of test scores in their utility function, then feedback should allow such tailoring of effort irrespective of whether the individual has prior experience of the institution or not.

In addition, this finding allays the concern that our baseline findings are driven by unobservable department characteristics that make period two test scores spuriously higher in the feedback regime. If this were the case, we should find an effect on all students regardless of their undergraduate background, given that exams and essays are marked anonymously. The finding also casts doubt on the alternative explanation that students with better essay-writing skills sort into departments that provide feedback. Since information on feedback policies is not publicly available, students who took their undergraduate degree at the same institution are more likely to be able to find out informally, e.g. by asking other M.Sc. students before choosing which department to apply for. Hence, if the results were driven by students sorting on the basis of the departments' feedback policy, we would expect the effect of feedback to be stronger for those students who could find out about these policies. The pattern of results in Table 4 is the opposite. Finally, the results also help rule out the hypothesis that students respond to the mere fact that their department cares enough to provide feedback, rather than to the informational content of feedback. Again, if this were the case, such effects of feedback should be independent of the value of information.[19]

Motivated by the defining features of the empirical setting, the analysis so far has focused on the role of feedback as a signal for the marginal return to effort in period two. Economic theory however indicates that feedback on past performance can affect current performance directly if past and current performances are substitutes or complements in the agent's utility function. To assess the relevance of any non-separabilities in individuals' utility function that lead to direct effects of feedback the Appendix reports two tests. First, we test for a specific form of non-separability, namely that students care only about their final M.Sc. degree classification – be it a distinction, merit, and so on – rather than their continuous test score in each period (Table A2). Second, we test a general implication of all non-separable utility functions on the anticipation of feedback on behavior in period one (Table A3). Both tests support the assumption of separability, thus indicating that feedback does not affect performance through a direct preference effect.

[19]Two other points are of note. First, Vollmeyer and Rheinberg [2005] present evidence from the laboratory that feedback allows subjects to validate or change previous strategies. Indeed, they show those that unexpectedly receive feedback adopt better strategies. In a similar spirit, we explored whether the dispersion of period one test scores affected the difference-in-difference effect of feedback. On the one hand more dispersed test scores may help a student to specialize more easily in the topic of her long essay. On the other hand, a greater dispersion of marks may be less informative of true return to effort. Overall however, we found no robust evidence that the dispersion of period one test scores significantly changes the effect of feedback. Second, to check for whether students behave as if they care about their absolute or relative performance, we checked whether the effect of feedback varied with the number of students enrolled on the same degree program, or the number enrolled in the department as a whole. We did not find any robust evidence of differential feedback effects in such cohorts of different size.

4.4 Heterogeneity

The theoretical framework makes clear that the documented positive average effect of feedback on period two performance is likely to mask heterogeneous responses across students. To explore this, we estimate the effect of feedback at each quantile $\theta \in [0, 1]$ of the conditional distribution of test scores:

$$Quant_\theta(g_{idt}|.) = \alpha_\theta X_i + \beta_\theta[F_d \times T_t] + \gamma_\theta T_t + \sum_{d'} \mu_{d'\theta} TD_{id'}, \tag{4.2}$$

where instead of the student fixed effects we control for the set of student characteristics X_i.[20]

Figure 3 plots the $\widehat{\beta}_\theta$ coefficients from the quantile regressions in (4.2) at each quantile θ, the associated 95% confidence interval, and the corresponding OLS estimate from (4.1) as a point of comparison. In line with the descriptive evidence, Figure 3 shows that the provision of feedback has little effect on the performance of students in the lowest third of quantiles of the conditional test score distribution, and has positive effects on the upper two thirds. The magnitude of the effect becomes more pronounced between the 33rd and 80th quantiles. At the highest quantiles the effect of feedback declines slightly but remains significantly larger than zero. This is in line with there being increasing marginal costs of effort in generating test scores so there exists some upper bound on how much test scores can improve by for the highest return to effort students when they are provided feedback.

The quantile analysis also allays the identification concerns stemming from unobservable characteristics of departments and students that are correlated both with the feedback policy and the difference in scores between periods one and two. The fact that the effect of feedback is positive only for students at quantiles .33 and above rules out unobservables that would produce a constant difference between period one and two test scores, such as departments with feedback marking more generously.

In summary, the quantile analysis indicates that our baseline result is actually an average of zero and positive effects across all students. There is no evidence the provision of feedback on period one test significantly scores *reduces* the period two performance of any student. In terms of the theoretical framework summarized in Table 1 this rules out several parameter combinations. In particular, we can rule out both combinations of parameters in the off-diagonal cells as these predict a negative effect of feedback. The findings are reconcilable with the theory developed if individual preferences and beliefs are correlated so as to lie on the leading diagonal of Table 1. In other words, the motivation effect dominates for individuals who tend to underestimate their return to effort in the absence of feedback and the slacker effect dominates for individuals who tend to overestimate it, so that there exists a precise interplay between the prior beliefs of individuals

[20] As in previous specifications, the student's characteristics controlled for are her gender, whether she attended the same university as an undergraduate, whether she is a registered for UK, EU, or outside EU fee status, and the academic year of study.

and the nature of their individual preferences. To the best of our knowledge, such interplays have not been previously documented.

4.5 A Placebo Test and Further Robustness Checks

To further allay the concern that our estimated feedback effect captures the effect of department unobservables that are correlated with difference in test scores across periods, this section develops a placebo test based on a subsample of students for whom the feedback policy of their department should have no effect, if the effect is genuine. The test relies on the fact that the essay is always graded by the department the student is enrolled in, yet because students can take courses in different departments, the feedback they receive depends on the feedback policy of the department where courses are taught. In other words, students enrolled in feedback departments will not receive feedback on the courses they take in no-feedback departments. This allows us to disentangle the effect of actual feedback from the effect of having the essay marked by a department that provides feedback on its courses.

Of the 7,738 students in our sample, 5,942 (77%) students take all their examined courses within the same feedback regime and 1,796 (23%) students take at least one examined course in both feedback regimes. This placebo test is predicated on the fact that students do not systematically sort into courses that provide feedback. For example, if those students that expect to benefit the most from feedback seek out such courses, then we likely overestimate the true effect of feedback. To begin with we note that 22% of students from no-feedback departments and 25% of students from feedback departments exhibit variation in the feedback regime they are actually exposed to across their examined courses. Hence at first sight we do not observe proportionally more students from the no-feedback regime choosing courses in the feedback regime. In the Appendix and Table A4 we present more detailed evidence that, on the margin, students do not sort into courses on the basis of whether they will be provided feedback in that course or not.[21]

We then re-estimate (4.1) for these two subsets of students. Column 1 of Table 5 shows that for students who take *all* their examined courses within the same feedback regime, the effect of feedback is $\widehat{\beta} = 1.03$ which is almost one third larger than the baseline estimate in Column 5 of Table 3. We next estimate (4.1) for the subsample of students that take courses in both feedback regimes. Column 2 in Table 5 shows that the feedback policy of the department the student is enrolled into has no significant effect on the difference in test scores across periods, with the point estimate on the parameter of interest falling by over two thirds from that in Column 1.

More importantly, we next estimate (4.1) for the subsample of students that take courses in both feedback regimes and additionally control for an interaction between the time period, T_t, and a dummy variable equal to one if student i actually obtains feedback on at least 75% of

[21] We find no evidence that students that take courses outside of their own regime also take more courses per se, say because they choose more half-unit courses. On average, students that take all their courses in the same feedback regime take 5.51 courses in total, and those that move across regime take 5.52 courses.

her examined courses, and zero otherwise. The result in Column 3 shows there is a significant difference-in-difference in test scores between students that actually receive feedback on 75% of their courses versus those that do not, allowing for a differential effect of the feedback policy of the department in which they are formally enrolled.[22] This placebo test suggests that students enrolled into feedback regime departments do not naturally have higher second period test scores than students enrolled in no-feedback departments. The results therefore provide reassurance that our estimates of the effect of feedback are not contaminated by unobservable departmental characteristics that are correlated to the feedback regime and the difference between test scores in the two periods.

Finally, Columns 4 and 5 include a rich set of interactions between the period two dummy and students and departmental characteristics respectively. The departmental characteristics are the admissions ratio, the number of teaching faculty, the number of enrolled students, and students overall satisfaction with courses. The student characteristics are gender, age, whether the student is British, whether she was an undergraduate at the same institution and race. If feedback is correlated to characteristics that create a difference between period one and two grades, we should observe a reduction in our coefficient of interest β . Columns 4 and 5 show that the main finding is unchanged in this case: $\widehat{\beta}$ remains positive, precisely estimated and of similar magnitude as in the baseline specification when we allow the difference between period one and two scores to differ by students and departments characteristics.

While we cannot, and do not, claim that the decision to provide feedback is random, the heterogeneous effects uncovered in Sections 4.3 and 4.4, together with the placebo test and robustness checks reported in this section suggest that it is orthogonal to various classes of unobservables that could create a positive difference between period one and two test scores.

5 Discussion

This paper exploits a natural experiment in a leading UK university to quantify the effect of feedback on students' academic performance. Our findings suggest that the average student performs better after receiving feedback on their earlier performance. In line with the assumption that feedback conveys information on the marginal return to effort, we find that the effect is stronger for students who are *ex ante* more likely to benefit from feedback.

Two features of our empirical setting have implications for the external validity of our results. First, the tasks individuals perform are not identical across periods: the production of test scores for examined modules and essays does not require precisely the same skill set. On the one hand the production technologies are sufficiently similar for feedback on period one tasks to be informative

[22]The fact that students require feedback on the majority of their courses – not just on one course – before this feedback has a significant effect on their second period performance suggests, as is reasonable, that there is some uncertainty or stochastic component in the production function for grades.

of the returns to period two effort. On the other hand, the results are likely to underestimate the effects of feedback in settings where identically the same task is repeated over time.

Second, it is important to bear in mind that the leading UK university we study selects from the most able students. This may have important implications for the documented heterogeneous effects of feedback across the return to effort distribution. For example, the results suggest the effect of feedback is more pronounced on more able students and that there are no significantly negative effects of feedback on any student. In short, no student becomes discouraged by the feedback they receive.

All these results may be reversed in another setting. For example, the positive effect of feedback may be larger for high return to effort students because of returns to continuous test scores in the UK labor market for graduates from this university. This is because students that graduate with the highest test scores can signal to future employers that they are the most able individuals from a university that is highly competitive to enter to begin with. Such superstar effects of being the highest achieving student from this university reduce the incentives of high return to effort students to slack in response to feedback. At the other end of the return to effort distribution, students may still have strong incentives generated by the labor market return to successfully graduating from this university, albeit with the lowest degree class of pass, rather than becoming discouraged and failing altogether. It would therefore be worthwhile in future research to identify the effects of feedback in a university where students are selected from a different part of the ability distribution.

The result that, in this setting, feedback has no negative effects on the test scores of any student, implies that the optimal policy is to always provide feedback, assuming a department's objective is to maximize the academic achievement of students. There are however, a number of explanations why departments may optimally choose not to provide feedback. First, departments may have incorrect beliefs about the preferences or priors of their students. As the model makes clear, feedback will have negative effects on *all* students' effort if *all* students preferences are such that the motivation effect dominates and they are all optimistic, or if the slacker effect dominates for *all* students and they are all pessimistic. This is because in either of these scenarios the effect of feedback is negative, as highlighted in the off diagonal elements of Table 1. The notion that departments are not well informed about the preferences or priors of their students is supported by two facts – (i) the same departments in another university in close proximity to the one we study have different feedback policies, suggests that even within the same subject, departments' priors over their students behavioral response to feedback is likely to differ; (ii) departments in this university have not in the recent past changed their feedback policies and so are not able to update their beliefs over the behavior of their students in counterfactual feedback policy environments [Levitt 2006].

Second, there may be costs of providing feedback that we do not document. For example, departments may anticipate students will engage in influence activities and other inefficient be-

haviors if feedback on earlier test scores is provided. Alternatively, there may be other margins of behavior – such as cooperation between students – that are crowded out by the provision of feedback. It would be interesting for future studies to try and measure the existence and magnitude of such effects of feedback.

6 Appendix

6.1 Preference Effects of Feedback

To assess the relevance of any non-separabilities in individuals' utility function that lead to direct preference effects of feedback we proceed in two steps. First, we test for a specific form of non-separability, namely that students care only about their final M.Sc. degree classification – be it a distinction, merit, and so on – rather than their continuous test score in each period. Second, we test a general implication of all non-separable utility functions on the anticipation of feedback on behavior in period one.[23]

6.1.1 Non-Separable Preferences

The primary reason why students might care about their final degree classification is that UK based employers usually make conditional offers of employment to students on the basis of this classification rather than on the basis of continuous test scores. Entry requirements into Ph.D. courses or professional qualifications are typically also based on this classification. Letters of recommendation from faculty written during the academic year also stress the overall degree classification the student is expected to obtain, rather than their predicted test scores.

There are four possible degree classes – distinction, merit, pass, or fail. The algorithm by which individual test scores across all courses translate into this overall degree class is not straightforward, but it is well known to students. Our first test exploits this algorithm to identify those students whose exam test scores are such that effort devoted to the essay is unlikely to affect their final classification. We test whether students who, being in the feedback regime, know they have low returns to period two effort, obtain a lower period two test score compared to students who are not given feedback. The intuition behind the test is that these are the students for whom we should find the strongest negative effect of feedback, if feedback affects period two performance directly because test scores are substitutes across periods.

To pin down the returns to second period effort for each student we first measure the difference between her average period one test scores and the grade she requires on the essay to obtain the highest class of degree available to her. We then define a dummy variable which varies at the student level, L_i, which is equal to one if student i lies in the top 25% of students in terms of

[23]Such anticipatory effects have been documented in laboratory settings . For example, Vollmeyer and Rhenberg [2005] show that subjects that anticipate feedback in the future adopt better strategies to begin with.

this difference, and is zero otherwise. Hence the marginal utility of the essay grade for students for whom $L_i = 1$ is low in the sense that they can reach the highest available class of degree even if they get a much lower score in their essay than they previously did in their exams. We then estimate the following specification for student i enrolled on degree program p in department d using data only from the essays in period two,

$$g_{ipd} = \alpha_p + \beta_0[F_d \times L_i] + \gamma_0 L_i + \rho X_i + \varepsilon_{ipd}, \tag{6.1}$$

where ε_{ipd} is clustered by program-academic year, to capture common shocks to all students writing a long essay on the same degree program in the same academic year, such as the availability of computing resources or grading styles.

In Column 1 of Table A2 we estimate (6.1) without including program fixed effects and so we can also estimate the direct effect of feedback. The results show that long essay grades are significantly higher for students with low returns to period two effort ($\widehat{\gamma}_0 > 0$). This is as expected because students are more likely to experience low returns to period two effort if they have performed better on their period one examined courses, everything else equal. Importantly, we find no evidence that students with low returns to period two effort who are actually aware of this because of the feedback they receive, obtain differential period two grades than students in the no-feedback regime ($\widehat{\beta}_0 = 0$). Hence there is little evidence of such students reducing effort because of the feedback they receive. Column 2 shows this result to be robust when we condition out unobserved heterogeneity across degree programs and estimate (6.1) in full.

We next additionally control for a series of interactions between the feedback dummy and student characteristics ($F_d \times \mathbf{X}_i$) to address the concern that low returns to effort for student i may reflect some other individual characteristic other than her returns to effort. The result in Column 3 continues to suggest the behavior of students who are aware that they have low returns to effort do not reduce effort as a result, conditioning on other forms of heterogeneous responses.

One concern with these results is that students' behavior might differ depending on which is the actual highest class of degree they can obtain. More precisely, the algorithm by which individual test scores convert into degree classifications implies there are two types of student that face low returns to their period two effort.

First, there are some students that because of their combination of poor marks on examined courses, can obtain a degree classification no better than a pass irrespective of their performance on the long essay. Such students may be expected to be demotivated by feedback that makes clear to them they are in this scenario. To check for such discouragement effects of feedback that arise from non-separable preferences, Column 4a estimates (6.1) for the subset of students who can at best obtain the degree class of pass – the lowest possible degree class a student can graduate with. The result shows that even among this subset of students we cannot reject that $\widehat{\beta}_0 = 0$.[24]

[24]Recall that the long essay is not accorded more weight than examined courses in the students overall degree

Second, there will be other students that because of their very good marks on examined courses, can obtain a degree classification of distinction even if their performance on the long essay is far below what they have previously achieved on their examined courses. Such students may decide to slack if their preferences are such that they are classmark targeters, namely, they seek to minimize their effort subject to obtaining some degree class. In terms of the model, for such individuals the slacker effect far outweighs the motivation effect. The provision of feedback to pessimistic classmark targeters should then reduce their effort.[25]

To check for this we restrict the sample to students that potentially obtain the highest degree class of distinction. Among such students, those with low returns to second period effort can obtain a distinction even with a low essay mark. The result in Column 4b again show that students do not behave as if they are classmark targeters. This may be because nothing prevents students from eventually revealing their continuous exam marks to their future employer. If there is a wage premium associated with higher continuous test scores – say because of superstar effects [Rosen 1981] – then such students may not have incentives to slack.

6.1.2 Anticipation Effects

The theoretical framework makes clear that all non-separable utility functions share the prediction that the anticipation of feedback should affect behavior in period one. The next test sheds light on the empirical relevance of this prediction. Using the subsample of 1,796 students who take examined courses in both feedback and no-feedback regimes, we test whether the *same* student performs differently in period one on courses for which individual feedback will be provided relative to her courses for which no feedback is provided. We therefore estimate the following panel data specification for student i enrolled in department d on examined course c,

$$g_{idc} = \alpha_i + \beta F_c + \gamma Z_c + \delta X_c + \varepsilon_{idc}, \tag{6.2}$$

where F_c is equal to one if examined course c provides feedback, and zero otherwise, Z_c is equal to one if examined course c is offered by the student's own department d, and zero otherwise, and X_c are other course characteristics.[26] The disturbance term ε_{idc} is clustered by course-year to capture common shocks to the academic performance of all enrolled students such as the difficulty of the final exam or quality of teaching faculty. The null hypothesis is $\beta = 0$ so the anticipation

classification. It is for example, possible to be awarded a degree even if the student fails the essay, if their examined course marks are sufficiently high. The fact that we observe no discouragement effects of feedback throughout is consistent with recent evidence from the laboratory [Eriksson *et al.* 2009].

[25]Our notion of classmark targeting behavior is therefore similar to income targeting in the labor supply literature for which there is mixed evidence [Camerer *et al.* 1997, Oettinger 1999, Farber 2005, Fehr and Goette 2007].

[26]These characteristics of the course in the academic year are the share of women, the mean age of students, the standard deviation in age of students, the racial fragmentation among students, the fragmentation of students by department, the share of students who completed their undergraduate studies at the same institution, and the share of British students.

of feedback has no effect on period one behavior, this implies $U_{g_1 g_2} = 0$ and/or that the students do not know whether feedback will be provided when choosing effort in period one.

Table A3 presents the results. Column 1 shows that unconditionally, the correlation between the examined course test score and whether feedback is subsequently provided is negative but not significantly different from zero. Column 2 shows that conditional on course characteristics X_c, the point estimate of β moves towards zero and is still insignificant. This remains the case when we additionally control for whether the course is offered by the department in which the student is enrolled (Column 3). Column 4 further shows there is no differential anticipatory effect of feedback in a course offered by the department in which the student is actually enrolled. This may have been the case if such feedback was more valuable for the writing of the essay for example, which must always be based on the primary subject matter of the degree program in which the student is actually enrolled.[27] The remaining Columns show that there are no anticipatory effects of feedback controlling for the difficulty of the course in various ways (Columns 5 and 6) and controlling for the assignment of teaching faculty to courses (Column 7).[28]

Overall, the data suggests, consistent with the descriptive evidence in Table 2 and Figure 2, that – (i) there is little effect of the anticipation of feedback on period one behavior; (ii) the baseline difference-in-difference effects of feedback reported in Table 3 relate to the impact of a signaling effect of feedback on period two behavior. This helps rule out any model in which students are aware of the feedback regime and there being complementarities or substitutabilities between g_1 and g_2 in the utility function so that $U_{g_1 g_2} \neq 0$, such as students being risk averse over test scores, or individual preferences being defined over the number of successes as in Lizzeri *et al.* [2002].[29]

6.2 Course Selection

To check whether students appear to purposefully sort into courses on the basis of whether feedback is provided or not, we estimate the following specification for student i enrolled in a degree program

[27] These last two specifications confirm exam performances are higher on examined courses offered by the same department as that in which the student is registered, as documented in Bandiera *et al.* [2010].

[28] In Column 5 the course difficulty is measured by the average test score of students who are not in this sample – namely those students that take all their examined courses in the same feedback regime. The sample drops in this column because there are some courses in which all the students are enrolled in a department with an alternative feedback policy from the department that offers the course. In Column 6 we measure the difficulty of the course by the share of enrolled students that are re-sitting the course from the previous academic year. In Column 7 we control for the assignment of teaching faculty to the course in a given academic year by controlling for a complete series of dummies for each faculty member j, where this dummy equals one if faculty member j teaches on the course in that academic year, and is zero otherwise.

[29] The results also rule out the type of implicit incentives provided by feedback in Ederer [2010]. These arise because in a tournament setting, agents have incentives to increase period one effort in the anticipation of feedback as it enables them to signal to their rival that they are of high return to effort. The evidence we present is in line with there not being such tournament effects in our setting, presumably because given grading is not on a curve, students care about their absolute and not relative performance on the course.

offered by department d,

$$prob(B_{id} = 1) = \alpha F_{id} + \beta X_i + \gamma X_d + \varepsilon_{id}, \tag{6.3}$$

where B_{id} is equal to one if student i from department d takes examined courses in both feedback regimes, and is equal to zero otherwise, F_{id} is equal to one if student i is enrolled on a degree program in department d in the feedback regime, and is equal to zero otherwise, X_i and X_d are characteristics of the student and department respectively, and ε_{id} is a disturbance term that is clustered by program-academic year.[30]

The results in Table A4 show that unconditionally, students from no-feedback departments are no more likely to select courses across both feedback regimes than students that are enrolled in the feedback regime to begin with so that $\widehat{\alpha}$ is not significantly different from zero (Column 1). This result is robust to conditioning on student and departmental characteristics (Column 2) and to estimating (6.3) using a probit regression model (Column 3).

An important class of factors that may determine the choice of courses are course specific characteristics. To control for these we repeat the analysis at the student-course (ic) level and estimate the following specification,

$$prob(B_{idc} = 1) = \alpha F_{id} + \beta X_i + \gamma X_d + \delta_c + \varepsilon_{idc}, \tag{6.4}$$

where B_{idc} is equal to one if student i from department d enrolls onto course c that is in a different feedback regime to the department the student is enrolled onto, and is zero otherwise, F_{id}, X_i, X_d are as previously defined, δ_c is a course fixed effect, and we continue to cluster the error term by program-academic year. The course fixed effect captures all characteristics of the course that are common within the academic year in which student i is enrolled, such as the class size, quality of teaching faculty and so on. The result in Column 4 shows that conditional on course characteristics, there remains no evidence that students enrolled in no-feedback regime departments are more likely to enroll onto courses offered by feedback departments.

References

[1] AARONSON.D, L.BARROW, AND W.SANDER (2007) "Teachers and Student Achievement in the Chicago Public High Schools", *Journal of Labor Economics* 24: 95-135

[30] The student characteristics controlled for are gender, whether the student is British, whether they were an undergraduate at the same university, whether they are registered as UK, EU, or non-EU fee status, race (white, black, Asian, Chinese, other, unknown), and the academic year of study. The characteristics of the department the student is enrolled in that are controlled for are the ratio of teaching faculty to enrolled students, and the ratio of enrolled students to applicants.

[2] ANGRIST.J AND V.LAVY (1999) "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics* 114: 533-75.

[3] AOYAGI.M (2007) Information Feedback in a Dynamic Tournament, mimeo, Osaka University.

[4] AZMAT G. AND N. IRIBERRI (2010) "The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment Using High School Students", *Journal of Public Economics*, 94: 435-452.

[5] BANDIERA.O, V.LARCINESE AND I.RASUL (2010) "Heterogeneous Class Size Effects: New Evidence from a Panel of University Students", *Economic Journal* 120: 1365-98.

[6] CAMERER.C, L.BABCOCK, G.LOEWENSTEIN AND R.THALER (1997) "Labor Supply of New York City Cabdrivers: One Day At A Time", *Quarterly Journal of Economics* 112: 407-41.

[7] EDERER.F (2010) "Feedback and Motivation in Dynamic Tournaments", *Journal of Economics and Management Strategy* 19: 733-69.

[8] ERIKSSON.T, A.POULSEN AND M-C.VILLEVAL (2009) "Feedback and Incentives: Experimental Evidence", *Labour Economics* 16: 679-88.

[9] ERTAC.S. (2006) "Social Comparisons and Optimal Revelation: Theory and Experiments", mimeo UCLA.

[10] FARBER.H.S (2005) "Is Tomorrow Another Day? The Labor Supply of New York City Cab Drivers", *Journal of Political Economy* 113: 46-82.

[11] FEHR.E AND L.GOETTE (2007) "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment", *American Economic Review* 97: 298-317.

[12] FISCHHOFF.B AND R.BEYTH (1975) "I Knew it Would Happen: Remembered Probabilities of Once-future Things", *Organizational Behavior and Human Performance* 13: 1-16.

[13] GERSHKOV.A AND M.PERRY (2009) "Tournaments With Midterm Reviews", *Games and Economic Behavior* 66: 162-90.

[14] HANNAH.R.L., KRISHNAN.R. AND D.NEWMAN (2008) "The Effects of Disseminating Relative Performance Feedback in Tournament Versus Individual performance Compensation Plans", The Accounting Review, 83: 893-913.

[15] HOFFMAN.F AND P.OREOPOULOS (2009) "Professor Qualities and Student Achievement", *Review of Economics and Statistics* 91: 83-92.

[16] KRUEGER.A (1999) "Experimental Estimates of Education Production Functions", *Quarterly Journal of Economics* 114: 497-532.

[17] LEVITT.S.D (2006) An Economist Sells Bagels: A Case Study in Profit Maximization, NBER Working Paper 12152.

[18] LIZZERI.A, M.MEYER AND N.PERSICO (2002) The Incentive Effects of Interim Performance Evaluations, CARESS Working Paper 02-09.

[19] OETTINGER.G.S (1999) "An Empirical Analysis of the Daily Labor Supply of Stadium Vendors", *Journal of Political Economy* 107: 360-92.

[20] RIVKIN.S.G, E.A.HANUSHEK, AND J.F.KAIN (2005) "Teachers, Schools, and Academic Achievement", *Econometrica* 73: 417-59.

[21] ROCKOFF.J.E (2004) "The Impact of Individual Teachers on Student Achievement: Evidence From Panel Data", *American Economic Review* 94: 247-52.

[22] ROSEN.S (1981) "The Economics of Superstars", *American Economic Review* 71: 845-58.

[23] VOLLMEYER.R AND F.RHEINBERG (2005) "A Surprising Effect of Feedback on Learning", *Learning and Instruction* 15: 589-602.

Table 1: The Effect of Feedback

	Pessimistic	Optimistic
	$a>\hat{a}$	$a<\hat{a}$
Motivation effect dominates	$e^F>e^{NF}$	$e^F<e^{NF}$
Slacker effect dominates	$e^F<e^{NF}$	$e^F>e^{NF}$

Table 2: Student Performance in Exams and Long Essays, by Feedback Regime

A. Mean and Standard Deviation in Test Scores, by Period and Feedback Regime

Period 1 test score (average exam)	No Feedback	Feedback	Test of equality [p-value]
Mean	62.2	62.4	.195
Standard deviation	4.49	4.60	.371
Period 2 test score (essay)			
Mean	62.9	63.8	.000
Standard deviation	6.23	6.79	.000

B. Difference in Difference in Performance by Period and Feedback Regime

	No Feedback	Feedback	Difference
Period 1 test score (average exam)	62.2	62.4	.138
	(.098)	(.194)	(.217)
Period 2 test score (essay)	62.9	63.8	.898***
	(.154)	(.251)	(.293)
Difference	.697***	1.46***	.760***
	(.124)	(.199)	(.234)

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. There are two observations per student - one for the period 1 (examined) component and one for the period 2 (essay) component. The test score for the examined component is the average test score over such modules. There are 7,738 students, 4,847 (2,891) of whom are enrolled in departments in the no feedback (feedback) regime. In Table 2A, we report the p-value on a test of equality of means based on a two sided t-test where the variance of test scores are not assumed to be equal. We also report the p-value on a test of equality of variances based on Levene's statistic. In Table 2B there are again two observations per student, and we report the average test score by period and feedback regime. We report means, and standard errors in parentheses. The standard errors on the differences, and difference-in-difference, are estimated from running the corresponding least squares regression, allowing the standard errors to be clustered by program-year. There are 369 such clusters.

Table 3: Baseline Estimates of the Effect of Feedback

Dependent Variable (Columns 1 to 5) = Student's module test score in period t
Dependent Variable (Column 6) = Student's GPA in period t
Dependent Variable (Columns 7) = 1 if Student obtains an A-grade in period t, 0 otherwise
Standard errors reported in parentheses, clustered at the program-year level

	(1) Controls	(2) Teaching Department FE	(3) Enrolment Department FE	(4) Enrolment Programme FE	(5) Student FE	(6) GPA	(7) A-Grade
Period 2 x Feedback	.882***	.926***	.742***	.755***	.797***	.108***	.045***
	(.267)	(.250)	(.245)	(.245)	(.247)	(.026)	(.014)
Period 2	1.28***	.650***	.696***	.636***	.448**	.034	.038***
	(.163)	(.179)	(.180)	(.184)	(.192)	(.021)	(.011)
Feedback	.041	-.165					
	(.188)	(.269)					
Fixed effects	None	Teaching dept	Teaching dept Enrolment dept	Teaching dept Enrolment prog	Teaching dept Student	Teaching dept Student	Teaching dept Student
Adjusted R-squared	.037	.054	.057	.076	.711	.697	.644
Number of observations (clusters)	15476 (369)	15476 (369)	15476 (369)	15476 (369)	15476 (369)	15476 (369)	15476 (369)

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. In Columns 1 to 5 the dependent variable is the student's test score on the module, where there are two observations per student - one for the period 1 (examined) component and one for the period 2 (essay) component. The test score for the examined component is the average test score over such modules. Standard errors are clustered at the program-year level throughout. In Column 6 the dependent variable is the average GPA in modules. The GPA on any given module is equal to 4 for an A-grade, 3 for a B-grade, 2 for a C-grade, and 1 for a pass. The dependent variable in these columns is then the GPA averaged across all examined modules and the GPA for the dissertation module. In Column 7, an A-grade corresponds to an average mark of 70 and above. In Columns 1 to 4 we control for the following student characteristics - gender, whether they were an undergraduate at the same university, whether they are registered as a student from the UK, EU, or outside the EU, and the academic year. The teaching department dummies are equal to one for the student if she takes at least one module in the department. The enrolment department and enrolment program fixed effects refer to the department or program the student herself is enrolled into.

Table 4: Mechanisms

Dependent Variable = Student's module test score in period t

Standard errors reported in parentheses, clustered at program-year level

	(1) Undergraduate at Same University	(2) Undergraduate at Different University
Period 2 x Feedback	.072	.925***
	(.421)	(.265)
Period 2	.970***	.338
	(.371)	(.210)
Fixed effects	Teaching dept	Teaching dept
	Student	Student
Adjusted R-squared	.734	.707
Number of observations (clusters)	2500 (279)	12976 (365)

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Standard errors are clustered at the program-year level throughout. In Columns 1 and 2 the dependent variable is the student's test score on the course, where there are two observations per student - one for the period 1 (examined) component and one for the period 2 (essay) component. The test score for the examined component is the average test score over such modules.

Table 5: Placebo Tests and Robustness Checks

Dependent Variable = Student's module test score in period t

Standard errors reported in parentheses, clustered at the program-year level

	(1) All Courses in Same Feedback Regime	(2) Not All Courses in Same Feedback Regime	(3) Not All Courses in Same Feedback Regime	(4) Student Interactions	(5) Departmental Interactions
Period 2 x Feedback policy of department enrolled in	1.03***	.397	-.130	.877***	.671**
	(.273)	(.890)	(.878)	(.256)	(.274)
Period 2	.349*	.081	.067	-1.89***	2.10
	(.198)	(.961)	(.950)	(.528)	(1.81)
Period 2 x Feedback actually received in 75% of examined courses			1.83***		
			(.634)		
Fixed effects	Teaching dept Student	Teaching dept Student	Teaching dept Student	Teaching dept Student	Teaching dept Student
Student characteristics x period 2 interactions	No	No	No	Yes	No
Departmental characteristics x period 2 interactions	No	No	No	No	Yes
Adjusted R-squared	.716	.701	.703	.714	.714
Number of observations (clusters)	11884 (336)	3592 (220)	3592 (220)	15476 (369)	13204 (301)

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. The dependent variable is the student's test score on the course, where there are two observations per student - one for the period 1 (examined) component and one for the period 2 (essay) component. The test score for the examined component is the average test score over such modules. Standard errors are clustered at the program-year level. In Column 1 the sample is restricted to those students who take all their examined modules in departments with the same feedback regime. Columns 2 and 3 restrict the sample to those students who take examined modules across both feedback regimes. The teaching department dummies are equal to one for the student if she takes at least one course in the department. Column 4 includes the interaction between the period 2 dummy and all student characteristics listed above. Column 5 includes the interaction between the period 2 dummy and the following department characteristics: the admissions ratio, the number of teaching faculty, the number of enrolled students, and students overall satisfaction with courses.

Table A1: Department Characteristics, by Feedback Regime

	No Feedback [14 departments]		Feedback [8 departments]		p-value on Mann Whitney test of equality
	Mean	SE	Mean	SE	
Student and Faculty Characteristics					
Intake	117	23.9	102	26.1	.633
Applications	635	121	868	322	.946
Admissions rate (intake/applications)	.198	.018	.163	.023	.219
Number of teaching faculty	17.6	3.35	10.6	1.37	.183
Share of teaching faculty that are professors	.302	.019	.261	.048	.191
Teaching Evaluations					
Module satisfaction [1-4]	3.19	.033	3.14	.069	.682
Library satisfaction [1-4]	2.87	.040	2.97	.054	.191
Reading list [1-4]	3.17	.020	3.18	.039	.823
Lectures integrated with classes [1-4]	3.32	.039	3.28	.053	.585
Lectures were stimulating [1-5]	3.98	.048	3.88	.117	.585
Lecturer content was understandable [1-5]	4.22	.051	4.15	.125	.785
Lecturer's delivery was clear [1-5]	4.29	.051	4.19	.121	.371
Class teacher encouraged participation [1-5]	4.26	.040	4.22	.045	.682
Class teacher was well prepared [1-5]	4.42	.038	4.28	.070	.088
Class teacher gave good guidance [1-5]	3.94	.041	3.88	.068	.339
Class teacher returned work within two weeks [1-5]	4.09	.093	4.34	.078	.086
Class teacher made constructive comments [1-5]	4.13	.088	4.26	.091	.456
Class teacher was available outside office hours [1-5]	4.36	.052	4.36	.045	.891
Module Characteristics					
Class size in examined modules	41.0	5.07	54.7	9.83	.117
Credits for examined modules	.762	.039	.671	.062	.152
Share of final mark that is based on the final exam in examined module	.806	.039	.716	.065	.172
Essay length [words]	10898	435	10833	833	.542

Notes: All statistics are averaged over the five academic years 1999/0 to 2003/4. A faculty member is referred to as professor if they are a full professor (not an associate or assistant professor). On the departmental teaching evaluations data, students were asked, "In general, how satisfied have you been with this module", "How satisfied have you been with the library provision for this module?", "How satisfied have you been with the indication of high priority items on your reading list?", and, "How well have lectures been integrated with seminars/classes". For each of these questions students could reply with one answer varying from very satisfied (4) to very unsatisfied (1). For the questions relating to the behavior of lecturers and class teachers, students could respond with one answer varying from agree strongly (5) to disagree strongly (1). In the module characteristics, the class size refers to the average number of students enrolled on each examined module. The final column reports the p-value from a Mann-Whitney two-sample test. The null hypothesis is that the two independent samples are from populations with the same distribution.

Table A2: Preference Effects of Feedback

Dependent Variable = Student's test score in period two

Standard errors reported in parentheses, clustered at program-year level

	(1) Unconditional	(2) Enrolment Program	(3) Interactions	(4a) Pass	(4b) Distinction
Low return to effort	1.70***	1.56***	2.11***	-.091	2.92***
	(.271)	(.279)	(.423)	(1.07)	(.830)
Feedback x Low return to effort	.495	.487	.418	-.041	.414
	(.390)	(.392)	(.404)	(1.43)	(1.39)
Feedback	.757**				
	(.344)				
Student level controls	No	Yes	Yes	Yes	Yes
Fixed effects	None	Enrolment prog	Enrolment prog	Enrolment prog	Enrolment prog
Feedback-student characteristic interactions	No	No	Yes	No	No
Adjusted R-squared	.020	.090	.091	.127	.218
Number of observations (clusters)	7738 (175)	7738 (175)	7738 (175)	2377 (167)	1560 (161)

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. Standard errors are clustered at the programme-year level throughout. The low return to effort variable is equal to one if the student requires a grade on her essay that is far lower than her average grade on examined modules, and still remain within the same degree classmark. We control for the following student characteristics - gender, whether the student is British, whether they were an undergraduate at the same university, whether they are registered as a student from the UK, EU, or outside the EU, the number of credits each module is worth, and the academic year. The enrolment department and enrolment programme fixed effects refer to the department or programme the student herself is enrolled into. In Column 3 we control for a series of interactions between whether the student has low returns to effort on her dissertation and the following student level characteristics - gender, whether the student is British, and whether she was an undergraduate at the same institution. In Columns 4a and 4b we consider the subset of students that can at best, obtain an overall degree classification of pass or distinction, respectively.

Table A3: Anticipation Effects of Feedback

Dependent Variable = Student's test score in period one

Standard errors reported in parentheses, clustered at course-year level

	(1) Unconditional	(2) Controls	(3) Own Department	(4) Interaction	(5) Difficulty	(6) Re-sits	(7) Faculty Assignment
Feedback [module]	-.302	-.091	-.031	-.492	.125	-.073	.048
	(.264)	(.225)	(.217)	(.482)	(.201)	(.214)	(.703)
Module offered by enrollment department [=1 if yes]			.934***	1.28***	.833***	.900***	.363
			(.224)	(.370)	(.214)	(.221)	(.224)
Feedback [module] x Module of enrollment department				.794			
				(.788)			
Module difficulty					.421***		
					(.035)		
Share of students re-sitting the module						-13.1**	
						(5.27)	
Student fixed effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Faculty dummies	No	No	No	No	No	No	Yes
Adjusted R-squared	.001	.538	.540	.540	.565	.541	.595
Number of observations (clusters)	7248 (1058)	7248 (1058)	7248 (1058)	7248 (1058)	6817 (993)	7248 (1058)	7248 (1058)

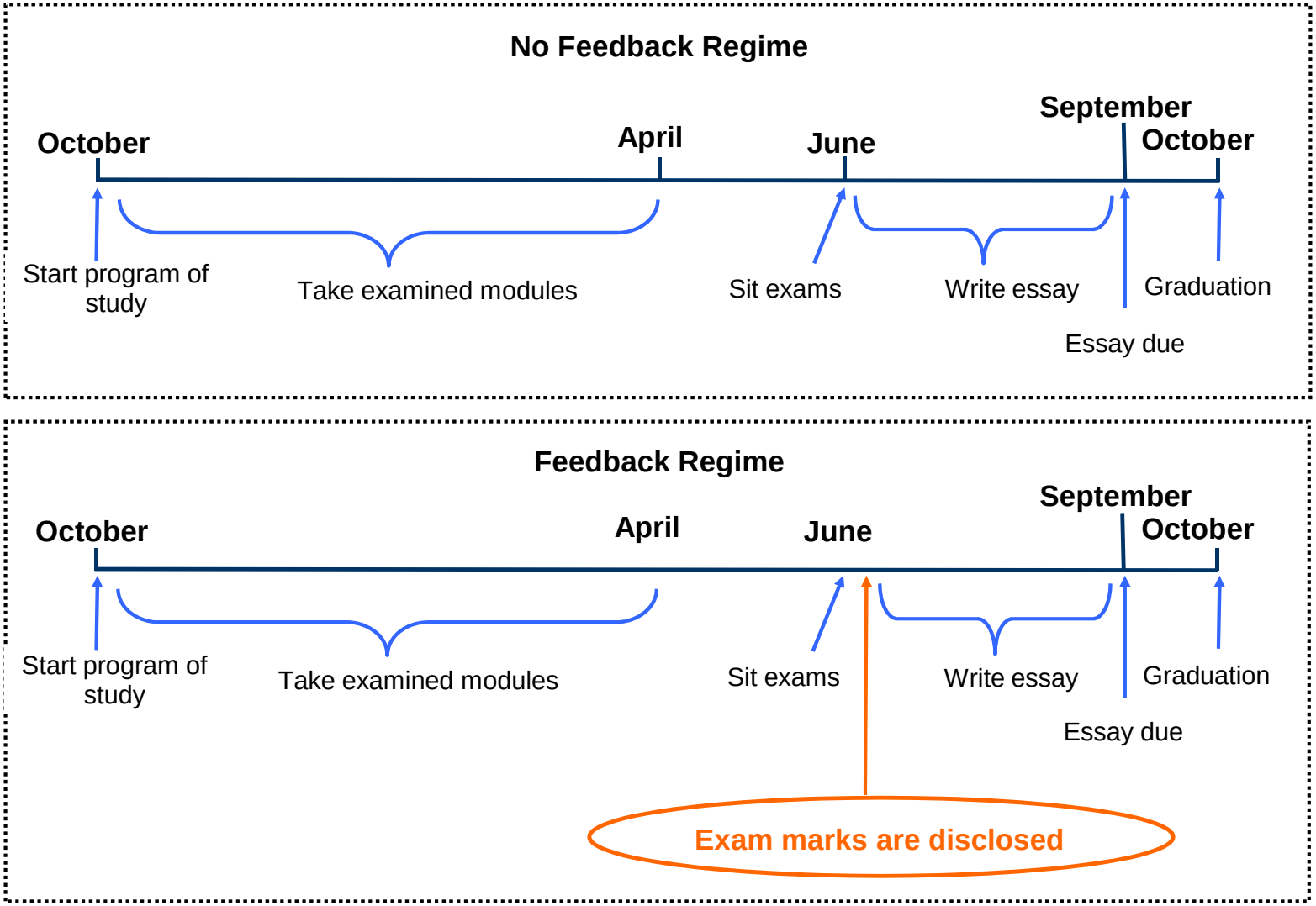
Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. The sample is restricted to students who take examined modules across both feedback regimes. The dependent variable is the student's grade on the period 1 (examined) module. Standard errors are clustered at the programme-year level throughout. We define the feedback [module] dummy to be equal to one if the module is offered by a department that provides feedback on its examined modules, and zero otherwise. In Columns 2 onwards we control for the following module-year characteristics - the share of women, the mean age of students, the standard deviation in age of students, the racial fragmentation among students, the fragmentation of students by department, the share of students who completed their undergraduate studies at the same institution, and the share of British students. In Column 5 the course difficulty is measured by the average mark of students who are not in this sample -- namely those students that take all their examined courses in the same feedback regime. The sample drops in this column because there are some courses in which all the students are enrolled in a department with an alternative feedback policy from the department that offers the course. In Column 6 we measure the difficulty of the course by the share of enrolled students that are re-sitting the course from the previous academic year. In Column 7 we control for the assignment of teaching faculty to the course in a given academic year by controlling for a complete series of dummies for each faculty member j, where this dummy equals one if faculty member j teaches on the course in that academic year, and is zero otherwise.

Table A4: Selection Into Courses and Feedback Regimes

Dependent Variable:	Dummy = 1 if student takes examined courses in both feedback regimes, 0 otherwise			Dummy = 1 if examined course taken is in a different feedback regime, 0 otherwise
	(1) Unconditional	(2) Controls	(3) Probit	(4) Fixed Effects
Enrolled in feedback department [yes =1]	.040	.023	.017	-.027
	(.046)	(.043)	(.045)	(.065)
Student controls	No	Yes	Yes	Yes
Department controls	No	Yes	Yes	Yes
Course-academic year fixed effects	No	No	No	Yes
Adjusted R-squared	.002	.016	.026	.619
Number of observations (clusters)	7483 (361)	7483 (361)	7483 (361)	30813 (361)

Notes: *** denotes significance at 1%, ** at 5%, and * at 10%. In Columns 1 to 3, observations are at the student level. The dependent variable is a dummy equal to one if the student takes at least one examined course in a department that has a different feedback policy to the policy of the department the student is actually enrolled in, and zero otherwise. In Column 4, observations are at the student-course level. The dependent variable is a dummy equal to one if the course is in a department that has a different feedback policy to the policy of the department the student is actually enrolled in, and zero otherwise. Standard errors are clustered at the program-year level in all columns. In Columns 2 to 4 the student characteristics controlled for are - gender, whether the student is British, whether they were an undergraduate at the same university, whether they are registered as a student from the UK, EU, or outside the EU, and their racial group (white, black, Asian, Chinese, other, unknown), and the academic year of study. In Columns 3 and 4 the characteristics of the department the student is enrolled in that are controlled for are - the ratio of teaching faculty to enrolled students, and the ratio of enrolled students to applicants.

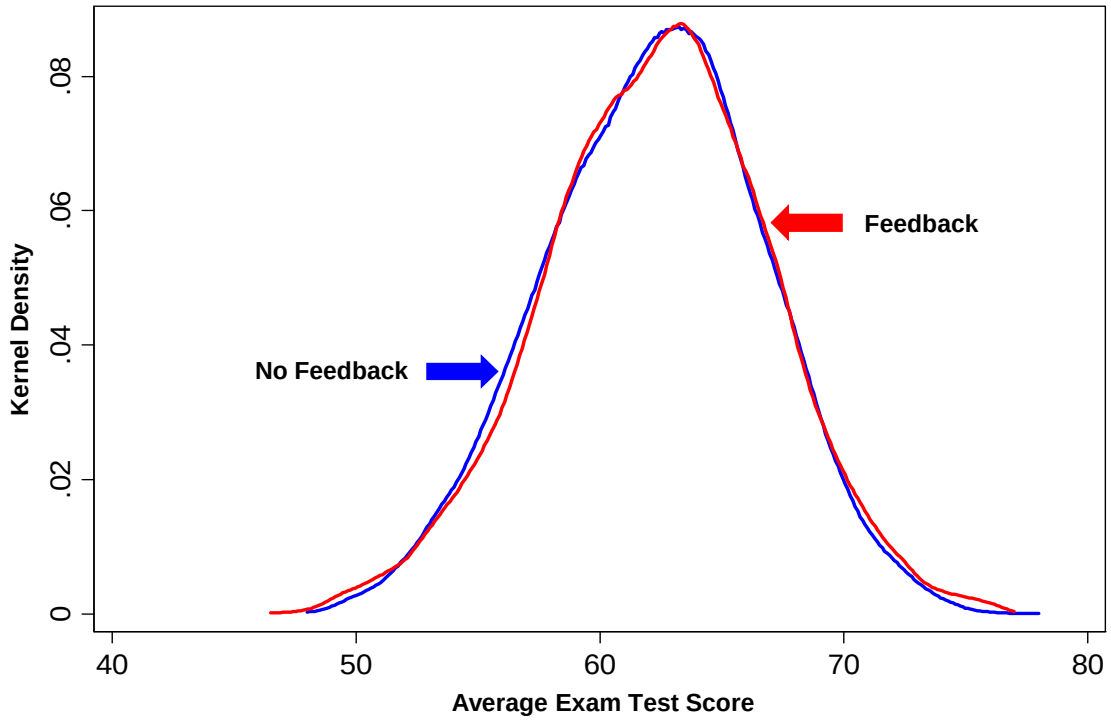
Figure 1: Timing and Feedback in One Year M.Sc. Courses



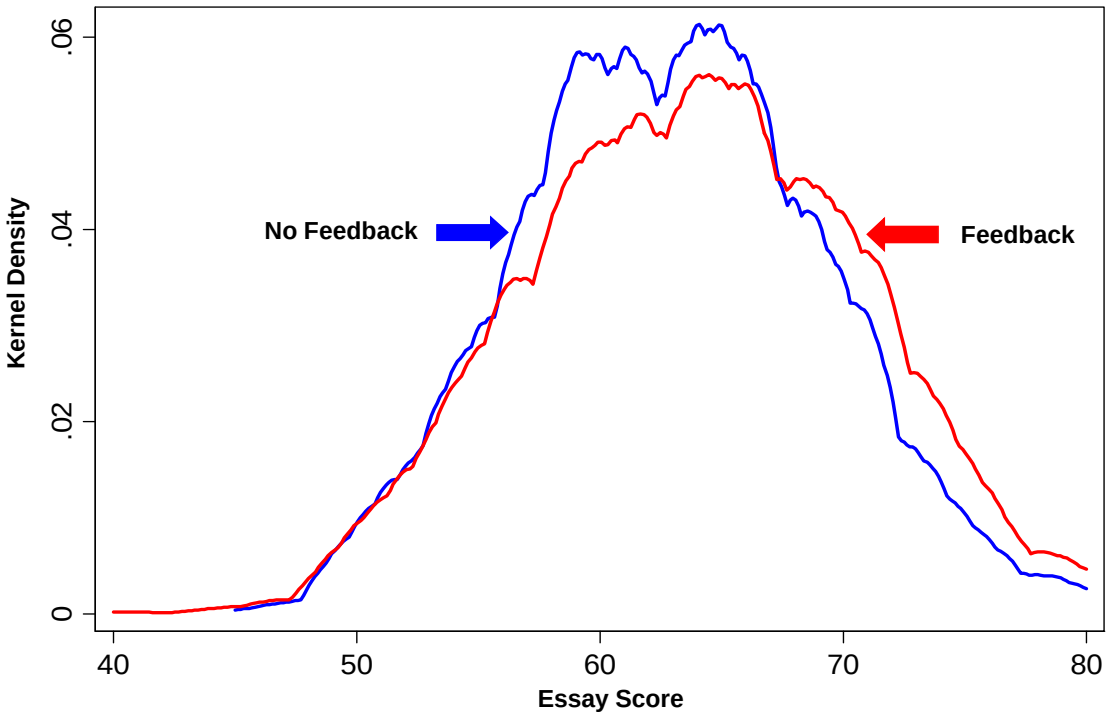
Notes: The classification of departments into the feedback and no feedback regimes is based on our own survey of heads of department.

Figure 2: Kernel Density Estimates of Student Performance in Exams and Dissertation, by Feedback Regime

A. Average Exam Test Score (Period 1)

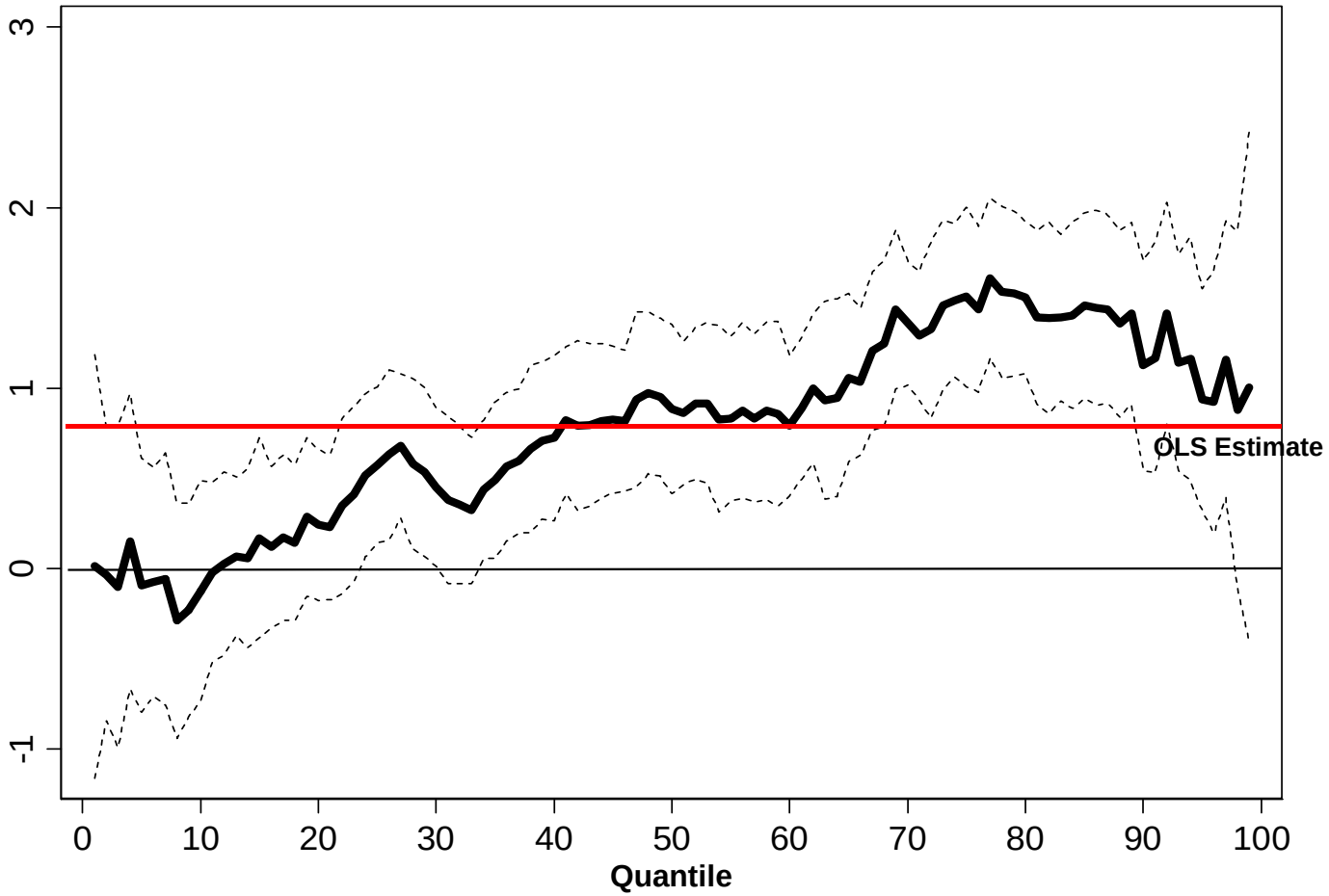


B. Essay Score (Period 2)



Notes: Each figure plots a kernel density estimate based on an Epanechnikov kernel function with an optimally chosen bandwidth. In Figure 3A the average exam test score refers to the average final test score for the student across all their examined modules. In Figure 3B the essay score refers to the final test score on the essay the student writes. There are 7,738 students, 4,847 (2,891) of whom are enrolled in departments in the no feedback (feedback) regime.

Figure 3: Quantile Regression Estimate of the Effect of Feedback



Notes: We graph the estimated effect of feedback on test score at each quantile of the conditional distribution of student test scores, and the associated 95% confidence interval. The distribution of test scores is conditioned on the following student characteristics - gender, whether the student is British, whether they were an undergraduate at the same university, whether they are registered as a student from the UK, EU, or outside the EU, the academic year, the time period, an interaction between the feedback regime and the time period, and teaching department dummies. These are equal to one for the student if she takes at least one module in the department.