The Impact of Initial Grades on Students' Further Academic Achievement

Liudmila Galiullina[*]

April 22, 2016

Abstract

I study the effect of grading on students' further academic achievement. I introduce an anonymous university data set and use it do discover grading patterns, in particular to find out whether higher grades can lead to better academic outcomes later on. I build a model that illustrates how effort-augmenting grading generosity can improve student effort through higher reference points and higher returns to effort.

JEL Codes: C81, I20, I21, I23.

Keywords: grading, university data set, learning outcomes, student motivation.

1 Introduction

As education economists, we are looking for optimal combinations of limited resources that would deliver desired educational outcomes. One of such limited resources is student time. As human beings, students are known to have all sorts of cognitive biases (time inconsistency, short-sightedness, self-control problems, etc.) that can prevent them from putting in enough study effort. Thus, if we want students to use their study time with maximum efficiency, we may try, in the spirit of libertarian paternalism, to construct mechanisms that would nudge students to choose optimal study behavior.

One of such behavioral economics no-cost tools that can help students overcome self-control problems and exert more study effort is grading. A major form of summative assessment, grades are supposed to provide students with direct feedback on how well they did in the course. However, grading choice is at the discretion of the grader (who can pursue their own interest); grade alternation can cost nothing and can even go unnoticed, but the impact can be significant as grades might influence students' self-beliefs and perceptions of their academic environment, change their expectations, and alter students' decisions about their future study targets and choices of effort.

[*] Department of Economics (AE2), School of Business and Economics, Maastricht University. Tongersestraat 53, 6211 LM Maastricht, The Netherlands. Tel.: +31-433-883-841, e-mail: l.galiullina@maastrichtuniversity.nl. I would like to thank Lex Borghans, Thomas Dohmen, Gábor Kézdi and Adam Szeidl for numerous helpful discussions, comments and suggestions. I am also grateful to the administration and staff of the anonymous university for providing me with the data and helping me understand them. This project was approved by the Ethical Research Committee of the Central European University, Hungary (Ref. No.: 2014-2015/1/RD). I am solely responsible for the contents of this paper.

Grading provides extrinsic motivation that can compensate students' costs of study effort. However, as an extrinsic motivator, it can also crowd out students' intrinsic motivation. Thus, grading effects can be ambiguous. Grading methods and systems — absolute, relative, blended, etc. — also vary a lot in different countries, institutions, levels of education, departments and courses. They can also vary within a classroom, when the teacher assigns different grades to different students based on factors irrelevant to learning assessment per se (cf. favoritism, prejudice, gender bias). Variety of grading systems (chosen ad hoc, without experimental testing) and lack of clarity in grading effects suggest that current grading practices may not induce optimal student effort, and there is room for research and possible evidence-based policy improvement. In this paper, I investigate what impact exposure to higher or lower grades has on students' further academic achievement.

The paper is organized as follows: first, we introduce our data set, discuss the construction of the sample and its limitations. Then we look for the grading patterns in the data, in particular for motivational effects of higher grading, such as passing important grading cutoffs. Having obtained some weak support for this conjecture, we build a simple behavioral economics model of student choice of effort in the second study period conditional on generosity of grading in the first study period. Our model shows that under effort-augmenting grading, higher grades result in higher effort in the second study period, which in particular means that passing an important threshold should lead to increase in effort. We conclude that we need more data to get enough power to test this hypothesis.

2 Data

2.1 Background

I introduce a data set of academic records and basic demographics of Masters students from an anonymous European university. The data cover eleven years at the very beginning of the 21st century. [1] Each year, around 500-600 freshmen, domestic and international, are enrolled in the university Masters programs.

For each course, in which a student enrolls, there is a grade recorded (except for dropped courses). The course record contains the grade point [2] , teacher code [3], student credit, effective year and study period, and indicators of whether the course is mandatory, or elective, or a language course. Data also distinguish graduation results: whether and when the student graduated, thesis grades, credit, and final GPA. GPAs—both preliminary and final—are always reported in the university information system as rounded to two digits. Students have full electronic access to their own records throughout their study programs.

At graduation, students receive a diploma of the quality that depends on their final GPA:

- a GPA of 3.67 or above results in a diploma with Distinction;

[1] For confidentiality reasons, I cannot disclose additional information about the university or give third parties access to the data set.

[2] The positive grades are 4 (the highest), 3.67, 3.33, 3, 2.67, 2.33, 2, 1.67, and 1 (the lowest); 0 means a fail.

[3] For teachers, original data contained only names, and I straightforwardly encoded them. There is a slight possibility that some two teachers had a common name and thus got the same code.

- a GPA of 3.33-3.66 results in a diploma with Merit;
- in some programs, a GPA of 3.00-3.32 leads to a diploma with Pass; in other programs—with a Satisfactory quality statement;
- in some programs, a GPA of 2.66-2.99 leads to a diploma with Pass, in others—results in no degree at all;
- students with a GPA below 2.66 receive no degree.

Also, a GPA below 2.66 in the first year of a two-year study program normally results in the student enrollment termination.

A major limitation of the data set is that it does not provide the variables needed to directly calculate the GPA. Using data on student credit as weights and not knowing exactly which elective courses count towards GPA, we can confidently calculate GPA in the first period only for students who took only mandatory courses (and might also have taken language courses, as they don't count towards GPA). This limits the sample drastically. [4] We work on obtaining more data from the source.

2.2 Construction of the sample

We analyze population of 5,788 Masters students spread across eleven academic years. For a small fraction of classes, students were not uniquely identified due to differences in departmental semesters encoding (we exclude them from our data). We also exclude 191 students who took courses in overlapping study periods (from different departments), so that we cannot tell their first and second study period. We don't use data on seven students with corrected student credit. [5] We also drop 4 students who took courses that counted towards a couple of their degrees. We lose students with unknown GPA in the first period (in one of the years, the system doesn't contain data on student credit that normally serves as grade weight). Finally, we restrict the sample to regular (non-visiting) students, who took only mandatory and possibly also language courses. [6] As mandatory courses with recorded grade points always count towards GPA, this ensures we correctly calculate first period GPA for our sample. The sample size shrinks to 606 students though, which results in 3,015 second period course records.

3 Patterns

Passing a grade threshold at the end of the first year and of the entire program must be an important issue for Masters students as this can influence their scholarship status and future academic prospects.

[4] Another difficulty is determination of the time sequence of study periods. I lose a few observations, for which time sequence cannot be inferred from the data. I also drop observations on a few students that complete courses that count towards two degrees simultaneously.

[5] Their student credit was corrected in order to adjust to departmental policy. E.g., if a student took a 5 credit course while according to his degree regulations only 4 credits could count towards his degree, then his credit was changed to 4.

[6] Language courses don't count towards GPA.

There is also evidence showing that the students care about their current nominal grades. [7] Before looking at the data, I expect to find discontinuities in grading effects near the meaningful cutoffs (2.66, 3.00, 3.33, 3.67 and 4.00). [8]

In order to see how second period educational outcomes depend on the nominal values of the first period grades, we can exploit discontinuities and compare means of learning outcomes near the grading thresholds. However, we have to be careful with the grades: they are ordinal by nature, so it makes little sense to compare their means straightforwardly and we need to come up with alternative outcome measures. [9]

We consider ”for grade” courses only. For each student we calculate the shares of second period mandatory courses that he completed with a grade point of 4.00, 3.67+, 3.33+, 3.00+, 2.67+, 2.33+ and a positive grade. [10] We group students in narrow intervals of their first period GPA (GPA1 hereafter), and calculate average abovementioned shares in the intervals. [11] Fig. 1 depicts these calculations.

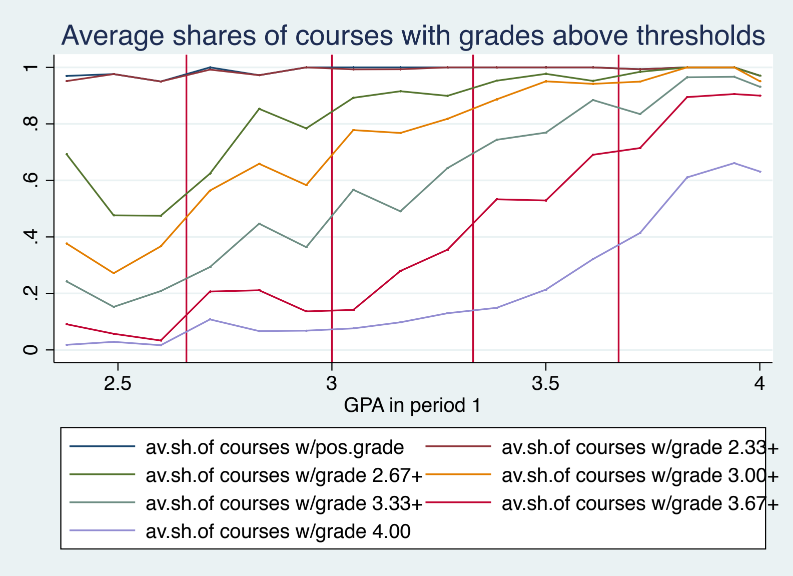

Fig. 1. Average shares of period 2 mandatory courses with grades above certain thresholds. *Note:* we don’t consider observations with GPA1 below 2.33, as they consist of highly heterogeneous

[7] I made a pilot survey in a group of Masters students who received quite heterogeneous results for their first exam, ranging from 4.00 to 2.33. The results showed that only 2 out of 13 respondents stated that they neither compare their grades with grades of others, nor assign much meaning to the academic ranking in their program - the remaining 85 percent of the respondents reported that they care either about grades or about ranking, or both, and thus we can conclude that their learning efforts can be influenced by the grading statistics. Also, 69 percent of the students reported that good or bad luck was one of the main determinants of their results, 62 and 54 percent attributed their result to grading method and teaching style respectively, and only 31 percent of the respondents stated that their hard work was one of the main determinants of their grade.

[8] I generated 10,000 first period random student records, assuming that each student takes five two-credit courses, and for each course has equal probabilities of 0.16 of getting a grade of 4.00, 3.67, 3.33, 3.00, 2.67, or 2.33, and a probability of 0.04 of getting a grade of 0 ("fail"). In my calculations, 36 percent of students received their grades on the border, i.e., if for one exam they received a marginally higher or lower grade (e.g., 3.33 instead of 3.00), then their GPA would fall into a different class category (e.g., "GPA $\geq$ 3.33 ("with merit")" instead of "GPA $\geq$ 3.00").

[9] Moreover, average grades (including GPA) are highly non-robust to single failures — they drop dramatically when any of the positive grade points switches to zero.

[10] Positive grades are all non-zero grades, i.e. 1.00, 1.67, 2.00, 2.33, 2.67, 3.00, 3.33, 3.67, and 4.00.

[11] Intervals are constructed in such a way that students in adjacent intervals would practically have the same GPA in period 1, only slightly differing in its nominal value.

failures.

To decipher the message of Fig. 1, we run a thought experiment and make a few assumptions along the way. Imagine that first period grades have no influence on second period ones (like in a situation when students don't learn their first period grades until the end of the second period). Then we should find no effect of the nominal value of GPA1 on second period outcomes, other things equal. To properly control for other things, we assume that GPA positively depends both on ability and effort, and that effort in the first period is orthogonal to ability. [12] GPA can also be influenced by random factors (grading strictness, quality of peers in case of grading-on-the-curve, external shocks to health, etc.), and we introduce a random component in the GPA that we call grading "luck" and assume to again be orthogonal to ability and effort. [13] This kind of luck cannot make as strong an impact on GPA as ability and effort though. We can imagine an average student switching from zero to full effort, resulting in shifting his GPA, say, from 1.5 to 3.5, or consider a difference between full-effort students with low ability earning a 3.0 and high ability earning a 4.0, but luck is not expected to be that influential: even if you receive preferential treatment in any course meaning that your grades are one unit (0.33 points) overstated, your GPA will still be only 0.33 points higher than without this luck.

Assuming independence and mutual substitutability of ability and effort, we expect that average ability and effort of students increases from left to right in Fig. 1. Assuming that intervals are small enough to make differences in ability and effort between immediate neighbors negligible, we can attribute differences in GPA1 between them to pure grading luck. [14] On average, having almost the same ability and exerting almost the same amount of effort in period 1, these immediate neighbors, if they had no information about their GPA1, should have indistinguishable second period outcomes. However 2-step neighbors (as well as 3-step, 4-step, etc., neighbors) should already differ in both average ability and effort (students with higher GPA1 should be more able and exert more effort in period 1, on average); these differences should be visible in second period outcomes unless second period effort is not altered by grading effecs. To discover these patterns, we conduct t-tests of the equality of means of neighbors and present the results in Table 1.

[12] We can also relax this assumption, accepting instead that first period effort is non-negatively correlated with ability.

[13] Shocks that influence student's grades through effort, such as, e.g., demotivating behaviors of peers or teachers, are excluded from the definition of this orthogonal grading "luck"; they represent another source of influence of student's environment on his outcomes through grades that we do not investigate here.

[14] Though, we need to make sure that these differences are not due to systematic differences between the neighbors (such as enrollment in particular programs and years, and taking particular numbers of courses).

Table 1. Dependence of period 2 grades on period 1 GPA

Interval	GPA in period 1	Number of students	Average shares of period 2 mandatory courses with the grades				
			2.67+	**3.00+**	**3.33+**	**3.67+**	**4.00**
1	[2.33…2.43]	22	0.692	0.377	0.242	0.091	0.018
2	[2.44…2.54]	14	0.476	0.271	0.152	0.057	0.029
3	[2.55…2.65]	20	0.475^{2}	0.367^{2}	0.208^{2}	0.033^{2}	0.017^{2}
4	[2.66…2.77]	25	0.624^{23}	0.564	0.293^{3}	0.207*	0.108
5	[2.78…2.88]	21	0.853*	0.659	0.447	0.211	0.067^{3}
6	[2.89…2.99]	22	0.784	0.583^{2}	0.364^{2}	0.136^{24}	0.068^{234}
7	[3.00…3.10]	28	0.892^{2}	0.778	0.567^{2}	0.142^{23}	0.076^{23}
8	[3.11…3.21]	30	0.916^{3}	0.768^{3}	0.491^{3}	0.279^{34}	0.098^{234}
9	[3.22…3.32]	43	0.899^{24}	0.818^{2}	0.643^{2}	0.355	0.130^{23}
10	[3.33…3.44]	53	0.953^{2}	0.887	0.744	0.533	0.149^{2}
11	[3.45…3.55]	61	0.977	0.951	0.769	0.529	0.213
12	[3.56…3.66]	32	0.952^{24}	0.942^{2}	0.884	0.691	0.321
13	[3.67…3.77]	37	0.984^{2}	0.950^{2}	0.835^{2}	0.714	0.414
14	[3.78…3.88]	19	1.000	1.000	0.965*	0.895	0.611
15	[3.89…3.99]	15	1.000	1.000	0.967	0.906	0.661
16	4	17	0.971^{234}	0.951^{234}	0.931^{24}	0.900^{2}	0.631^{2}
Total:		**459**					

Notes: * 1% significant positive difference (t-test of equality of means) from the previous interval (motivational effect of higher grades).

^{2} 10% insignificant positive difference (t-test of equality of means) from the 2-step previous interval (demotivational effect of grades).

^{3} 10% insignificant positive difference (t-test of equality of means) from the 3-step previous interval (demotivational effect of grades).

^{4} 10% insignificant positive difference (t-test of equality of means) from the 4-step previous interval (demotivational effect of grades).

Denote average effort of students in interval i as $E_i (i = 1, ..., 16)$. Table 1 (also Fig. 1) suggests the following patterns:

- $E_3 < E_1$, because these 2-step neighbors have insignificantly different results, in fact all the five educational outcomes in interval 3 are even (insignificantly) lower than those in interval 1. Also, $E_4 > E_3$, as interval 4 students do significantly better on the 3.67+ indicator and insignificantly better on all the others as well. Thus E_3 is a systematically small value, showing a demotivational grading effect of being in that interval. Interval 4, on the other hand, suggests a motivational effect of passing that crucial termination cutoff;

- $E_5 > E_4$, as it is significantly higher on the 2.67+ indicator (and insignificantly different on others); it is also insignificantly differenent (though a bit weaker) from period 8, supporting from the other side the idea that it is a local maximum as well;

- intervals 5-13 each don't show significant differences from their immediate neighbors, suggesting that second period effort wasn't significantly affected by particular GPA1 values in the range between 2.89 and 3.77;

- effort is significantly high in interval 14: $E_{14} > E_{13}$ and $E_{14} > E_{16}$;

- overall, we find some weak support for the conjecture that passing a threshold improves student's further effort: results are insignificant, though positive, near the 3.00, 3.33 and 3.67 cutoffs, and significantly positive only near 2.66.

4 Theoretical modeling

There can be several sources of randomness in grades (e.g., health, accidents, peers, prejudice, favoritism). We focus on one random component: grading strictness. Grades represent a partial order in the class, and thus give the teacher some discretion over which grade points to assign. For example, if there are four positive grades: A, B, C and D, and the class consists of three students, who do significantly differently from each other though all pass, then there are four possible grade assignments: ABC (the most generous), ABD, ACD, and BCD (the strictest). Manipulating grading generosity, teachers generate exogenous shocks to students' grades that can have impact on students' motivation further on.

We use reference-dependent preferences (Kőszegi and Rabin, 2006 [15]) to model learning behavior of a representative student whose program of study consists of two periods, in each of which he receives a separate period grade (e.g., a GPA): g_1 and g_2. The student has an innate ability a (which will be considered constant throughout the program), and in period i he can choose to apply learning effort e_i ($i = 1, 2$). The effort has a marginal cost c (and no fixed cost), and when choosing the effort level at the start of each period, the student takes into account his expected grade and his gain-loss utility (if he has set a reference point for his future grade). The student is behavioral and naive (in the O'Donoghue and Rabin, 1999 [16] , sense): he is concerned with his grades, costs of effort and pains of falling behind his expectations (if he has got any), but he is short-sighted – he optimizes his behavior only in the current period, not taking into account that his first period accomplishment will become his reference point to which he will be comparing his achievements later on. Consider a representative student who decides at the beginning of the first period what amount of effort he wishes to invest in learning. The effort has opportunity cost, and it is rewarded with a higher period grade (at least, in expected value). We normalize grades, abilities, efforts and luck to lie in between 0 and 1. We model knowledge k as a convex combination of ability a and effort e, and grade g as an imperfect measure of knowledge, such that the teacher translates only a fraction t of it into the grade g:

$k = \alpha e + (1 - \alpha)a$, $g = tk$, where $0 \leq \alpha, a, e, k, t, g \leq 1$.

Without initial expectations and naively believing that teachers will grade his work straightforwardly ($t = 1$), in the first period, the student solves his optimization problem:

$$\max_{e_1 \in [0,1]} g_1 - ce_1 = \max_{e_1 \in [0,1]} (\alpha - c)e_1 + (1 - \alpha)a.$$

[15] Kőszegi, Botond, and Matthew Rabin. "A model of reference-dependent preferences." The Quarterly Journal of Economics (2006): 1133-1165.

[16] O'Donoghue, Ted, and Matthew Rabin. "Doing it now or later." American Economic Review (1999): 103-124.

Consequently,

- if $c < \alpha$, then $e_1^* = 1, g_1 = \alpha + (1-\alpha)a$;
- if $c > \alpha$, then $e_1^* = 0, g_1 = (1-\alpha)a$.

Obviously, if too much emphasis is placed on ability (e.g., the student is told that he either ”has it” or not), the effort shrinks to zero as long as it's costly.

In the second period, a reference point emerges in the form of the first period grade, and the student also learns about the teacher's grading pattern t (which he believes to be constant), so the student modifies his optimization problem:

$$\max_{e_2 \in [0,1]} \begin{cases} g_2 - ce_2 + \eta(g_2 - g_1), g_2 > g_1; \\ g_2 - ce_2 + \eta\lambda(g_2 - g_1), g_2 \leq g_1. \end{cases} =$$

$$\max_{e_2 \in [0,1]} \begin{cases} t(\alpha e_2 + (1-\alpha)a) - ce_2 + \eta\left(t(\alpha e_2 + (1-\alpha)a) - g_1\right), t(\alpha e_2 + (1-\alpha)a) > g_1; \\ t(\alpha e_2 + (1-\alpha)a) - ce_2 + \eta\lambda\left(t(\alpha e_2 + (1-\alpha)a) - g_1\right), t(\alpha e_2 + (1-\alpha)a) \leq g_1. \end{cases}$$

The results are the following:

- if $c < t\alpha(1+\eta)$, then $e_2^* = 1$;
- if $t\alpha(1+\eta) < c < t\alpha(1+\lambda\eta)$, then $e_2^* = e_1^*$;
- if $c > t\alpha(1+\lambda\eta)$, then $e_2^* = 0$.

Combining with the results from the first period, the student's choice of effort falls into the following three categories depending on the grading style t of the first period teacher:

- if $t < \frac{1}{1+\lambda\eta}$ (”cruel grading” regime), some low-cost students get discouraged in the second period and switch to zero effort;
- if $\frac{1}{1+\lambda\eta} < t < \frac{1}{1+\eta}$ (”strict grading” regime), the student just sticks to his first period effort level (status quo) — low-cost students exert full effort, high-cost students exert no effort — as the motivation that comes from aversion of falling behind his first period result is just enough to keep his effort on the same level (and thus expect the same grade) when he faces not-too-high appreciation of his efforts;
- if $t > \frac{1}{1+\eta}$ (”generous grading” regime), some not-too-high-cost students get encouraged in the second period and switch to full effort.

The results also hold for Cobb-Douglas, Leonieff and generalized CES-function model. They are, however, different in a model with high grades coming as ”free lunch”. Having more data, we'll discover more reliable patterns and build our model around them, with testable implications.