

Grades and Incentives: Assessing Competing GPA Measures and Post-Graduate Outcomes

Michael A. Bailey, Jeffrey S. Rosenthal, and Albert Yoon¹

Last revised: May 28, 2014

Abstract

In many educational settings, students may have an incentive to take courses where high grades are easier to achieve, potentially corroding student learning, evaluation of student achievement, and the fairness and efficiency of post-graduation labor outcomes. A grading system that takes into account heterogeneity of teacher standards and student ability could mitigate these problems. Using unique data from a major Canadian research university, we calculate student grade point averages net of course difficulty and find evidence that raw grade point averages systematically distort student achievement across majors. We then link undergraduate performance and law school data. We find that adjusted GPAs better predict Law School Admissions Test scores, while the raw GPAs better predict admission to law school and grades in law school. These results suggest nuanced relationship between grades, incentives and subsequent academic outcomes. We conclude by discussing implications of our results for university leaders.

JEL classification: I2, I21, I23, J24, J44

Keywords: grading; universities; labor markets; grade point averages; admissions

¹Department of Government and Public Policy Institute, Georgetown University; Department of Statistics, University of Toronto; Faculty of Law, University of Toronto. This paper has benefitted from comments from Ben Alarie, Tracey George, Helen Levy, Jide Nzelibe and anonymous reviewers. We are very grateful to Glenn Loney and Sinisa Markovic of the University of Toronto for providing us with student grade data. We are also grateful to the University of Toronto Registrar and the Faculty of Law for making their data available. All remaining errors are our own.

Grades matter greatly in students' lives. Grades almost certainly affect students' subjective feelings of academic accomplishment (Boatright-Horowitz and Arruda 2013). More concretely, grades affect students' futures. Grades influence whether students can go on to graduate work and, in many cases, whether they can be interviewed and ultimately hired in the private sector (NACE, 2011).

Grades are typically summarized with a grade point average (GPA), the average of the student's grades across all of his or her coursework. This measure is ubiquitous and simple. It is also flawed. GPAs produce valid comparisons across students only if the course demands and teacher standards are either constant or randomly distributed across courses, conditions that are unrealistic in most academic contexts.

Flaws in the GPA could undermine academic standards, and perhaps already have. Students in pursuit of high grades may gravitate to easy courses, lowering the academic rigor of their studies and potentially initiating a race to the bottom among academic programs that may lower standards to attract students.

In this paper we make use of a novel data set to investigate the relationship

between grades in college, course difficulty and subsequent outcomes. We focus on the predictive capacity of actual GPAs and of GPAs that we adjust based on an item response theory model that accounts for variations in course difficulty and student ability across courses. Our data comes from complete transcripts for virtually all students at a major Canadian university from 2000 to 2010. We also have post-graduate outcomes for the students who applied for and, in some cases, attended law school at the same university. This allows us to assess whether unadjusted or adjusted GPAs better predict standardized test scores, admission to law school and grades in law school.

Several of our results highlight weaknesses of unadjusted GPAs. There is clear evidence that course difficulties vary by major, lowering GPAs for students in hard majors. We also find that the GPAs adjusted for course difficulty (net of variation in student ability across courses) better predict standardized test scores than do unadjusted GPAs, suggesting the adjustments track academic ability in a meaningful way. The adjusted GPAs matter not at all in the law school admissions process, however, and unadjusted GPAs matter a lot. This suggests students are, if anything, punished for taking more difficult courses.

Somewhat surprisingly, we also find weak evidence that, controlling for law school aptitude test (LSAT) scores, raw GPAs better predict law school grades than adjusted GPAs. This raises the possibility that raw grades reflect a set of useful skills (such as task orientation and perseverance) that may be unrelated to course difficulty but nonetheless matter in future academic achievement.

The paper proceeds as follows. In Part I we discuss grades and their connection to incentives and academic standards. We describe our statistical approach to adjusting GPAs for course difficulty and our data in Part II. In Part III, we assess the predictive power of the raw and adjusted GPAs for LSAT scores, law school grades and law school admissions. We conclude in Part IV by discussing the implications of our findings for higher education and labor markets.

I Grades, incentives and standards

Concerns about the GPA as a measure of achievement would largely disappear if grading standards were the same across classes or if students were randomly assigned to courses. For most students at most universities, how-

ever, neither condition holds.

As any undergraduate student quickly learns, courses vary in difficulty. These differences stem either from differences in content (e.g., some material is more challenging than others) or grading practices of teachers (e.g., some teachers are more apt to give high grades). Grading standards appear to vary across academic disciplines (Eaton and Eswaran 2014; Brown and Van Niel 2012). For example, Achen and Courant (2009) report that math professors consistently give lower average grades than other professors.²

And what courses students take is hardly random. Students openly discuss course selection based on course difficulty (Johnson 2003, 169). In a Duke survey, 40 percent of students said that the grading policies had some effect on their course selection (Johnson 2003, 186; see also Artz and Welsch 2013; Oettinger 2002).³ Sabot and Wakeman-Linn (1991) estimated that students at Williams taking more than one course in Economics would increase by 12 percent if Economics graded like English; students taking more than one

² It is possible, of course, that low grades in math arise because the students are weaker. The model we use to adjust grades allows us to directly assess such a question as it measures course difficulty and student ability simultaneously.

³ Another 21 percent answered “Not applicable” to the question even though it included a full range of responses from “no effect” to “very significant effect.” Presumably, selecting courses on basis of expected grade is socially undesirable and underreported in surveys.

English course would decrease by 47 percent if English graded like Math. Recent scholarship shows that students, and especially lower ability students, at a competitive university generally gravitate towards more “leniently graded courses” (Bar et al, 2009).

Such behavior can harm students. Students pursuing high grades may take courses, or even majors, that fall outside their stronger interests, and in some cases, aptitude (Sabot and Wakeman-Linn 1991). For example, a solid - but not stellar- student interested in the hard sciences may abandon science because other majors offer the promise of significantly higher grades. Recent scholarship shows that college students disproportionately move away from STEM (science, technology, engineering, or mathematics) majors, in part because of their more difficult grading (Stinebrickner and Stinebrickner 2014). More broadly, students gravitate in large numbers to majors with reputations for lack of rigor (Glenn 2012). To the extent that students move away from technical majors, they may be sacrificing tens of thousands of dollars in initial salaries and even more in mid-career salaries (Payscale 2014).

Pathologies of the GPA may also corrode academic standards. Individual instructors often face professional pressure to receive high teaching evalua-

tions; they may be tempted to give high grades in order to obtain higher evaluations from students (Zangenehzadeh 1998). Academic departments often face pressure to enroll many students in their courses; they may be tempted to lure students to their classes with high grades or not to scare away students with low grades (Anglin and Meng 2000). If other departments respond in kind, university standards unravel. This could be one factor behind the apparently low expectations at many universities. Arum and Roksa (2011, 36, 106), for example, found that nearly half of the university students they studied displayed “no statistically significant gains in critical thinking, complex reasoning, and writing skills” with students in some majors performing markedly worse than others.

Empirical work can help us better understand the role of grades in university education. First, it is useful to assess the problem. Are GPAs in fact distorted by differences in difficulty across courses? Are these distortions particularly bad for technical courses? Even as this question has been researched already, it is useful to build a body of research across various institutions. In addition, it is useful to extend the inquiry to class level analysis as differences across majors are not necessary for raw GPAs to be problematic. Even if all

majors had the same average difficulty, courses within majors could differ in standards and induce some of the problems discussed so far.

Second, does adjusting GPAs for course difficulty produce more useful measures of achievement? Is it the case, for example, that GPAs adjusted for course difficulties better predict subsequent outcomes such as scores on standardized tests or grade in graduate school? Students who fare better on adjusted than unadjusted grades may perform better on subsequent tasks for two reasons. One, they may simply be smarter. Such students may be less likely to be intimidated by difficult courses and may therefore take a course load that yields positive adjustments when we account for course difficulty. Two, students who take more difficult courses may learn more than students who take easier courses.

Third, what incentives do GPAs impart? Are students rewarded or punished for taking difficult courses when applying to graduate school? The GPAs available to admissions committees are not adjusted for course difficulty. While we might hope that admissions committees figure out a way to account for course difficulty, this may not be what happens in practice. Swift, Moore, Sharek and Gino (2013) present experimental results that admissions

officers focus on raw GPAs even when given information that GPAs from some applicants came from schools where GPAs were markedly higher. The same study also presents observational evidence that students coming from universities with higher overall GPAs are more likely to be admitted to business school, consistent with concerns that course difficulty goes unaccounted for in admissions processes.

Student pursuit of easy grades may then be a perfectly rational response to incentives. If admission committees do not account for course difficulty when evaluating transcripts, then students have an incentive to take easier courses as high grades in these courses would boost unadjusted GPAs more than adjusted GPAs. This raises the possibility that students, in choosing their courses, may pursue a strategy that is weakly related to, or even orthogonal to, intellectual development. Grades, rather than learning, could become the tail that wags the proverbial dog of higher learning.

II Methods and data

This paper addresses these questions by analyzing a unique data set that links post-graduate results to complete undergraduate transcripts for virtu-

ally all students at a major Canadian university from 2000 to 2010.

The first task for us is to adjust GPAs for course difficulty. Some scholars and universities use normalizations to grades that compare performance to other students, perhaps within and beyond majors (Felton and Koper 2005; Brown and Van Niel 2012). These are promising in several respects, but have some weak spots. For example, consider an approach that normalizes grades relative to the mean of the course. Under such a normalization an A will be worth less in a course that gives very high grades than in a course that gives very few high grades. The problem is that a course with a large number of excellent students (e.g. an “honors” class or advanced physics class) may be unfairly penalized as most students may get high grades not because the standards are low, but because they are excellent students. Likewise, under such a normalization a mediocre student who enrolls in a course where the other students are mostly inept would be unfairly advantaged.⁴

Our approach is to use an item response theory (IRT) model to estimate

⁴Swift, Moore, Sharek and Gino (2013) use GPAs relative to the high school or university average as an adjusted measure of achievement. Such a measure has difficulty distinguishing between grades earned at an excellent school where all high grades are earned in challenging courses versus grades earned at a weak school with low standards. Within schools, such an adjustment does not distinguish between a student with a high GPA who took hard classes and a student with a similar GPA who took easy classes.

student achievement while accounting for differences in course difficulty in a way that accounts for differences in student abilities (Johnson 1997, 2003; Young 1990).

We begin with a very simple model in which student achievement is an unobserved latent variable, θ_i . The observed grade in course t for student i , y_{it} , is a function of this latent student ability and a course specific error term:

$$y_{it} = \theta_i + \epsilon_{it}$$

If we treat the error term as random noise uncorrelated with latent ability, we can estimate student achievement with the average grade, a measure we call the raw (or unadjusted) GPA:

$$\hat{\theta}_i = \text{Raw GPA}_i = \frac{\sum_t y_{it}}{N_i}$$

where N_i is the number of grades received by student i .

The problem with this very simple model is that the error term ϵ_{it} contains many factors: not just course difficulty, but extraneous factors such as whether or not a student was distracted by a part-time job or a personal is-

sue. We focus on course difficulty as this could exert a major effect on grades and, for the reasons discussed above, could be correlated with achievement and thereby bias raw GPA.

Our IRT fixes this flaw in raw GPA by model controlling for course difficulty in a way that takes into account the ability of students in a course when estimating the course's difficulty. The model is

$$y_{it} = \theta_i + \kappa_t + \epsilon_{it}$$

where κ_t is a course easiness parameter: the higher it is, the easier the course. This means that for high κ_t s, students do better conditional on ability. Courses with negative κ_t s are difficult; students with similar ability can be expected to do worse in these courses than in courses with higher κ_t s.

This model allows us to account for the ability of students when determining course easiness. A course in which students receive high grades might initially reflect easy grading, but if the students in the course were high ability then the high grades may be explained by ability and the easiness parameter could be small or even negative. This approach is analogous to sports rank-

ings (e.g., RPI rankings for basketball) which are based on wins and losses, but account for the quality of teams against whom the wins and losses occur.

The approach exploits the overlap of courses that students have with one another. Suppose, for example, all freshmen take a single survey course and then go on to major in different subjects. If students from one major systematically performed higher than students of another major in this introductory course, we learn something about differential standards across majors.

Grade data in the data set we describe below is on a 0 to 100 scale, unlike the letter grade systems used widely at other universities. This simplifies estimation by not requiring a polytomous ordinal model that would typically be necessary for grading models in the U.S. or anywhere else with letter grades. Patz and Junker (1999) and Bailey, Stezhnev and Voeten (2014) present IRT models for ordinal data that are appropriate for letter grade systems. The underlying logic is similar, even as the ordinal nature of the grades introduces several complexities absent here.

There are a number of approaches to estimating such a model. A dummy variable approach is not practical given the large number of students (99,397) and unique courses (31,191) in the data set we describe below. We therefore

use a Markov Chain Monte Carlo simulation approach implemented in the C programming language⁵. In this approach, we begin by using the raw GPAs as starting value estimates for the θ s. We calculate the distribution of the κ parameters (reflecting course easiness) conditional on the values of the θ and we draw values from that distribution. Then, conditional on the value of our κ values, we simulate from the conditional distribution of the θ parameters. After a “burn-in” period, the random draws will come from the correct posterior distribution, allowing us to characterize the posterior distribution for each θ_i and κ_t . These models have been widely used in educational testing literature (Johnson and Albert 1999; Patz and Junker 1999) and other measurement contexts (e.g. Clinton, Jackman and Rivers 2004).

We define individual courses by unique combinations of term, instructors and course numbers. This allows for the possibility that different instructors have different standards when teaching the same course or that an instructor grades differently across terms or course numbers. We normalize our estimates of θ_i to have the same mean as the raw GPAs in our data set in order

⁵Available at: <http://probability.ca/jeff/ftplib/grademcmc.c>

to facilitate easy comparison.

We use data provided by the University of Toronto. The university registrar provided individual-level student grades for undergraduate students enrolled at the Faculty of Arts and Science at the University of Toronto – St. George campus between 2000 and 2010. This data contained all grades given for every undergraduate course, providing unique identifiers for student, course number and instructor for each grade.⁶

The University of Toronto Faculty of Law provided admissions data and grades during the first year of law school. Law grade information spans 2005 to 2010. Our admissions data is for applicants from the University of Toronto for 2008, 2009 and 2010. When combined with the undergraduate data, we are able to observe University of Toronto undergraduates who applied to its law school and whether they were accepted. For those who then attended University of Toronto Law School, we observe their first year law school grades.

Table 1 displays the summary statistics for the main variables of interest,

⁶We omitted courses with fewer than three students and students with fewer than three courses. We repeated grades for two semester, full-year courses as these count twice as much as single semester courses in raw GPAs.

Table 1: Summary statistics

	Mean	Standard Deviation	Min.	Max.	N
All students					
Raw undergraduate GPA	69.53	12.09	0.00	98.2	99,397
Adjusted undergraduate GPA	69.53	12.69	-13.56	105.65	99,397
LSAT and admissions sample					
Raw undergraduate GPA	78.24	6.33	40.0	92.2	503
Adjusted undergraduate GPA	77.27	7.10	39.6	96.6	503
LSAT	158.34	8.65	131.0	177.0	503
Admission	0.26	0.42	0.00	1.00	503
Law grades sample					
Raw undergraduate GPA	82.41	5.11	28.6	93.4	206
Adjusted undergraduate GPA	83.30	5.61	30.0	98.2	206
Law grades	74.79	2.29	67.2	80.8	206

the two GPA measures and the three dependent variables.

III Results

Differences in course difficulties There is little doubt that courses differ in difficulty. Figure 1 plots a histogram of estimated course difficulty parameters. High positive values indicate “easy” courses in which students did better than their achievement. Negative values indicate “hard” courses in

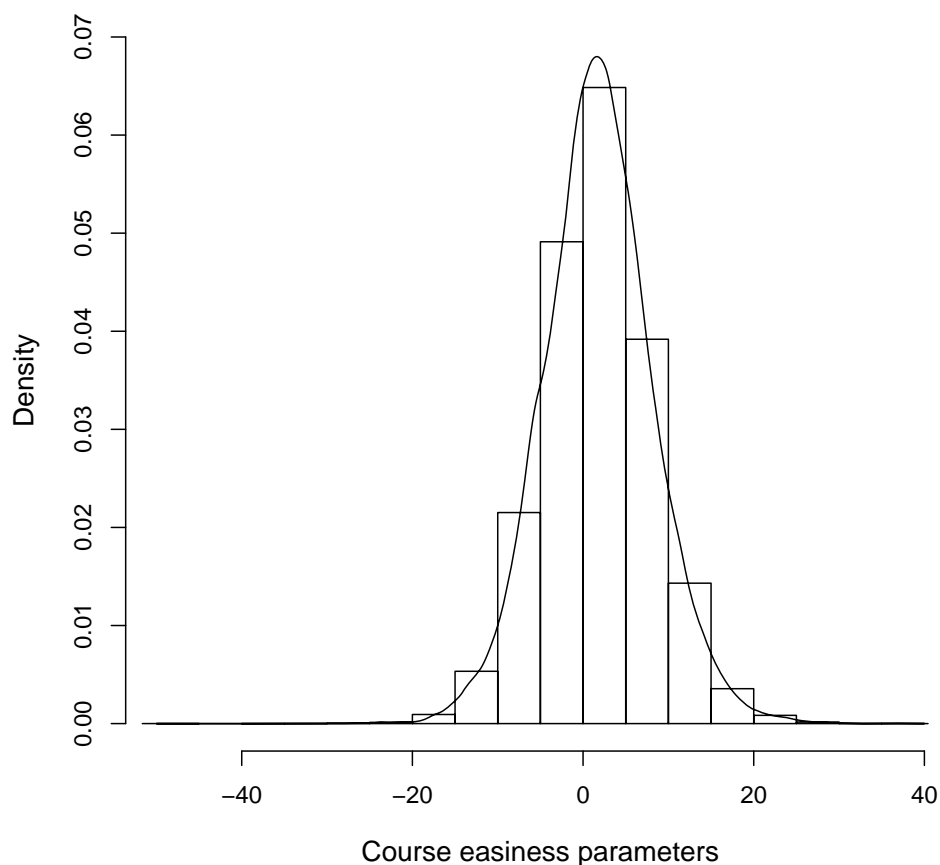


Figure 1: HISTOGRAM OF COURSE ESTIMATED EASINESS PARAMETERS

which students did worse than their achievement. A student in a course with a $\kappa=10$ could expect to get a grade 10 points higher than he or she would in a course with a $\kappa = 0$. A student in a course with a $\kappa = -10$ could expect to get a grade 10 points lower than he or she would in a course with a $\kappa = 0$. Almost 13 percent of courses have an absolute value of κ over 10.

Do these differences in course difficulties translate into chronic differences in raw and adjusted GPAs? In general, the answer is no. The correlation

of the two GPA measures is 0.96. However, there are considerable range of differences across the two measures. Figure 2 shows the histogram and kernel density of differences between adjusted and raw GPAs. The percent of students with a difference between raw and adjusted GPAs of more than 5 points is 13.6. In addition, these differences are associated with field of study.⁷ Figure 3 plots the average adjusted and raw GPAs for each major. A point to the left of the 45 degree line indicates the students in the major had higher adjusted than raw GPAs; students in these majors were punished by the raw GPA. A data point on the right of the 45 degree line indicates a major in which students had higher raw than adjusted GPAs; students in these majors benefited from the raw GPA.

The differences are not random: students in majors with higher achievement indices were hurt more by raw GPA. Majors with higher adjusted than raw GPAs include many in the sciences such as math, chemistry, physics and statistics. This is similar to findings by others (Johnson 2003, 210).

⁷We do not have data on the declared major of each student. Because we have students' entire transcripts we are able, however, to identify the field in which a student took the most classes. We refer to the department in which a student took at least three classes and more classes than any other department as his or her "major"; we believe we do not misclassify many students and for those that we do, using the modal discipline may be a better measure of assessing the influence on GPA measures than declared major.

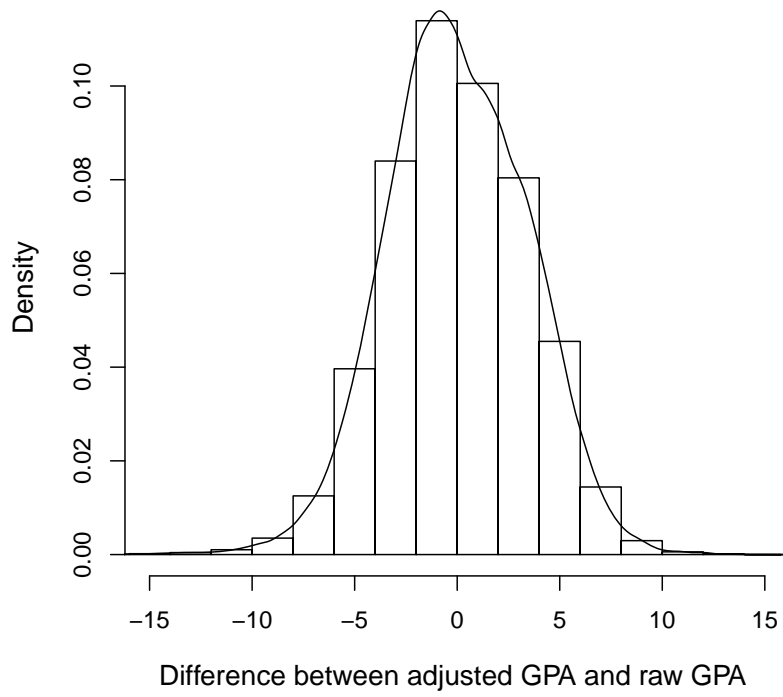


Figure 2: HISTOGRAM OF DIFFERENCES IN ADJUSTED AND RAW GPAs

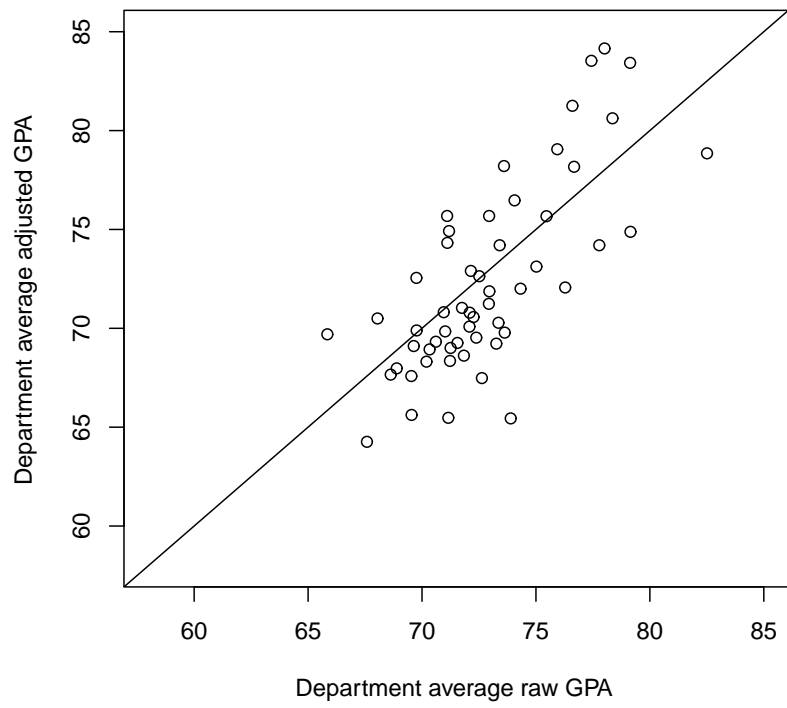


Figure 3: ADJUSTED AND RAW GPAs BY MAJOR

Performance on the LSAT One way to compare the validity of the competing GPA measures is to compare their abilities to predict performance on an external academic task such as a standardized test score. Table 2 displays how well the two measures predict LSAT scores using law school application data from 2008 to 2010. In the first column we include only raw GPA and in the second column we include only adjusted GPA. When included individually, each GPA measure is highly significant.

We are most interested in which GPA measure matters more. Therefore in the third column we include both measures and in the fourth column we add other control variables available to us: a dummy variable indicating the student majored in a STEM (science, technology, engineering and math) field; separate dummy variables indicating whether the student was Black, Asian or female and dummy variables for years.

When both grade measures are included in the model, the adjusted GPA measure is highly statistically significant and the raw GPA is statistically insignificant (and negative). A F-test for the null hypothesis that the coefficients on the two GPA measures are equal is rejected at $p=0.001$ for the specification in column (c) and at $p=0.008$ for the specification in column

Table 2: Predicting LSAT Scores

	(a)	(b)	(c)	(d)
Raw GPA	0.67*	-	-0.12	-0.10
	(0.05)		(0.15)	(0.16)
Adjusted GPA	-	0.65*	0.75*	0.70*
		(0.05)	(0.13)	(0.15)
STEM major	-	-	-	0.30
				(1.00)
Black	-	-	-	-5.39*
				(1.60)
Asian	-	-	-	-1.82
				(1.05)
Female	-	-	-	-1.99*
				(0.65)
Year 2009	-	-	-	2.09*
				(0.80)
Year 2010	-	-	-	2.67*
				(0.81)
(Intercept)	105.79*	107.76*	109.4*	111.59*
	(4.18)	(3.57)	(4.09)	(4.02)
N	503	503	503	503
R^2	0.24	0.29	0.29	0.34
$H_0 : \beta_{Raw\ GPA} = \beta_{Adjusted\ GPA}$	-	-	Reject	Reject
			(p=0.001)	(p=0.008)

Standard errors in parentheses. * $p \leq 0.05$. The dependent variable is the LSAT score of students who attended the University of Toronto as undergraduates and applied to the University of Toronto Law School.

(d). A J-test that the model with only raw GPA is better than the model with only the adjusted GPA is rejected at $p=0.000$. The results support the idea that the adjusted GPA more effectively predicts ability as measured via the standardized LSAT test.⁸

This evidence is consistent with results that show students who score high-

⁸We explored nonlinear specifications as well and did not find meaningful non-linear effects.

est on the LSAT tend to come from majors that are known for more difficult grading, majors such as math, physics and economics (Nieswiadomy 2010, 2014).

Law School Admissions We are also able to assess how well the two grade measures predict University of Toronto Law School admissions. Table 3 displays results from a probit model predicting admission to law school. The sample consists of students who went to the University of Toronto as undergraduates and then applied to the University of Toronto Law School. Column (a) includes only raw GPA and LSAT. Column (b) includes only adjusted GPA and LSAT. Column (c) includes both GPA measures and LSAT and column (d) adds demographic covariates and year dummies.

Raw GPA is clearly more predictive than the adjusted GPA when both are included. A likelihood ratio test of the hypothesis that the measures have the same effect is rejected at the $p=0.048$ level for the specification reported in column (c); the hypothesis is rejected at the $p=0.022$ when more controls are included as in column (d).

Figure 4 displays the effect of raw GPA on law school admission by plot-

Table 3: Predicting Law School Admissions, 2008 - 2010

	(a)	(b)	(c)	(d)
Raw GPA	0.171*	-	0.153*	0.201*
	(0.02)		(0.04)	(0.05)
Adjusted GPA	-	0.122*	0.016	0.002
		(0.02)	(0.03)	(0.04)
LSAT	0.121*	0.114*	0.120*	0.145*
	(0.02)	(0.02)	(0.02)	(0.02)
STEM major	-	-	-	0.161
				(0.25)
Black	-	-	-	1.882*
				(0.46)
Asian	-	-	-	-0.171
				(0.30)
Female	-	-	-	-0.031
				(0.18)
Year 2009	-	-	-	-0.276
				(0.22)
Year 2010	-	-	-	-0.566
				(0.22)
(Intercept)	-34.15*	-28.96*	-33.78*	-40.53*
	(3.21)	(2.76)	(3.25)	(4.03)
N	503	503	503	503
$H_0 : \beta_{Raw\ GPA} = \beta_{Adjusted\ GPA}$	-	-	Reject	Reject
	-	-	(p=0.048)	(p=0.022)

Standard errors from probit model are in parentheses. * $p \leq 0.05$. The dependent variable is admission into the University of Toronto Law School for students who attended the University of Toronto as undergraduates and applied to the University of Toronto Law School.

ting predicted probabilities of admission as a function of LSAT scores for two hypothetical students. The first student is a student of average ability who took courses at the 25th percentile of course difficulty; this student took “easy” courses. The probability of admission for this student is the higher, dashed line. For high LSAT scores, he or she is virtually guaranteed admission. The other student is a student of average ability who took courses at the 75th percentile of course difficulty; this student took “hard” courses. The average course difficulty parameters for this student were 8 points more negative than for the first student, resulting in a raw GPA that is 8 points lower. For this student, the predicted probabilities of admission are lower for every LSAT score, often much lower. In fact, even with a perfect LSAT score, the student who took hard courses would have only a 84 percent probability of admission.

Performance in First Year of Law School We are also able to assess the predictive capacity of the two measures for performance as first year law students. In their first year of law school, students at the University of Toronto Faculty of Law, like students at most law schools, take courses in administrative law,

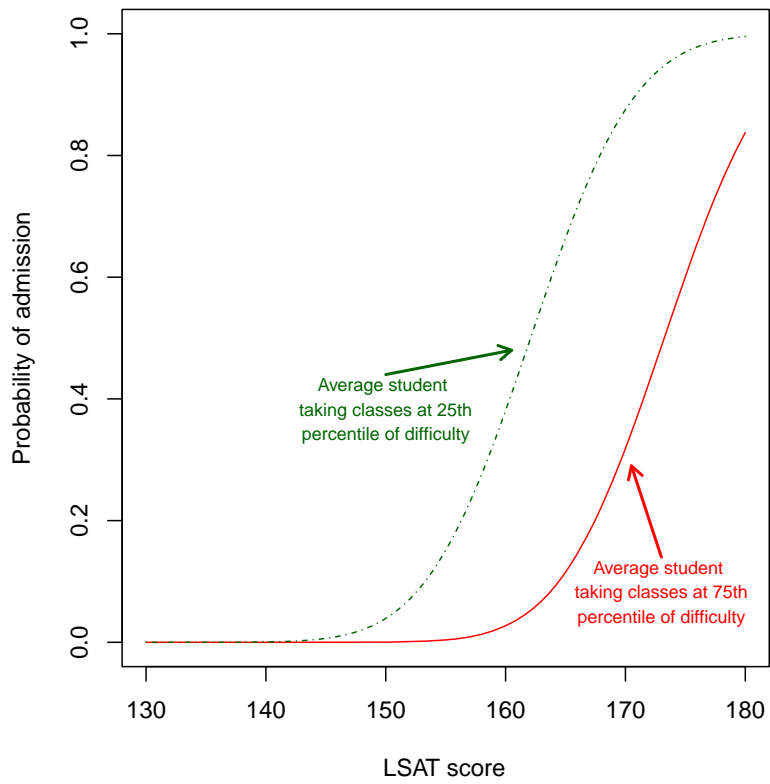


Figure 4: PROBABILITY OF LAW SCHOOL ADMISSION AS A FUNCTION OF LSAT AND COURSE DIFFICULTY

constitutional law, contracts, civil procedure, criminal law, property, torts, as well as legal research and writing. Most of these classes have enrollment of 60 to 80 students. Crucially, for our purposes, students are randomly assigned to their professors. Other features of these grades make them useful measures of performance. They are taken very seriously by students as first year grades are typically regarded as the most important for summer employment, clerkships, and other law school honors. The grades are also given in fairly standardized conditions as first year courses are anonymously graded to the same mean based final exam performance.⁹

Table 4 displays how well the two GPA measures predict law school grades. The first three columns do not control for LSAT scores; this helps us assess whether the measures differ in ability to directly predict performance. The last three columns control for LSAT to allow us to condition on the other information that law school admission committees have before them.

In the specification without LSAT there is weak evidence that raw GPA better predicts law grades. A J-test that model (a) (with only raw GPA) is better than model (b) (with only the adjusted GPA) is not rejected ($p= 0.57$)

⁹An exception is that each student takes one small (15 person) section course which is graded based on paper assignments and an anonymized final exam.

Table 4: Law School Grades

	(a)	(b)	(c)	(d)	(e)	(f)
Raw average grade	0.14*	-	0.26*	0.26*	-	0.29*
	(0.03)		(0.11)	(0.06)		(0.11)
Adjusted grade	-	0.12*	-0.05	-	0.17*	-0.07
		(0.03)	(0.09)		(0.05)	(0.09)
LSAT	-	-	-	0.13*	0.12*	0.11*
				(0.04)	(0.04)	(0.04)
STEM major	-	-	0.32	-	-	0.35
			(0.55)			(0.53)
Black	-	-	-2.46*	-	-	-1.45
			(1.03)			(1.06)
Asian	-	-	-0.32	-	-	-0.26
			(0.52)			(0.50)
Female	-	-	-0.69	-	-	-0.48
			(0.38)			(0.38)
Year 2006	-	-	0.35	-	-	0.29
			(0.59)			(0.57)
Year 2007	-	-	-0.69	-	-	-0.90
			(0.62)			(0.62)
Year 2008	-	-	-0.89	-	-	-1.04
			(0.59)			(0.58)
Year 2009	-	-	0.02	-	-	-0.15
			(0.61)			(0.59)
(Intercept)	63.59*	65.04*	57.76*	31.05*	40.24*	39.82*
	(2.47)	(2.28)	(5.31)	(7.42)	(6.66)	(8.29)
N	206	206	140	137	137	135
R^2	0.09	0.08	0.21	0.21	0.18	0.27
$H_0 : \beta_{Raw\ GPA} = \beta_{Adjusted\ GPA}$	-	-	Fail to reject (p=0.11)	-	-	Reject at 0.10 level (p=0.059)

Standard errors in parentheses. * $p \leq 0.05$. The dependent variable is the average first year grade at the University of Toronto Law School for students who attended the University of Toronto as undergraduates.

while a J-test that model (b) is better is almost rejected (with $p=0.12$). In model (c) the coefficient on raw GPA is positive and statistically significant while the coefficient on adjusted GPA is negative and statistically insignificant. We do not reject the null hypothesis that the two GPA coefficients are equal, although the p-value is close to marginal significance (at $=0.11$) (and keep in mind that the sample size is small and the two measures are highly correlated, thereby reducing the power of statistical tests).

When LSAT is included in models (d) through (f), the raw GPA measure is more predictive than the adjusted GPA. A J-test that model (d) (with raw GPA and LSAT) is better than model (e) (with adjusted GPA and LSAT) is not rejected ($p= 0.51$) while a J-test that model (e) is better is rejected ($p=0.02$). These p-values are 0.43 and 0.008, respectively, if we include the control covariates from model (f). The coefficient on raw GPA in the model with covariates is significant and larger and the F-test of the null hypothesis that the coefficients on the two GPA measures are equal is rejected at the 0.10 level (with $p= 0.059$).

While these results are not particularly precise, they do suggest that the law school admissions process is not irrational. Controlling for LSAT, raw

GPA is more predictive of performance in law school than adjusted GPA. One possible explanation for this unexpected result is that raw GPA may reflect a subset of student skills such as perseverance and perhaps goal-orientation (e.g. finding ways to get a GPA that is high enough to get into law school). Once we control for academic ability via LSAT scores, these skills may translate into better performance in law school.

IV Discussion and conclusion

Our analysis of grading of University of Toronto undergraduates show that even as adjusted GPAs correlate highly with raw GPAs, real differences between the two measures exist, especially for certain majors. The two GPA measures also predict outcomes differently. The adjusted GPA better predicts LSAT results, suggesting that the course difficulty adjustment either accounts for higher underlying ability or additional learning among students who take harder courses.

Unadjusted GPA better predicts admission into law school, suggesting that a student who enrolls in easier classes and therefore gets a higher raw GPA improves his or her chance of being admitted to law school.

Results for first year law school grades suggest raw GPA does better, especially when controlling for LSAT scores, suggesting that raw grades do nonetheless account for a dimension of student personality or achievement that can be relevant in future outcomes. Perhaps non-cognitive skills that economists are increasingly emphasizing as determinants of future success (see, e.g., Almlund, Duckworth, Heckman and Kautz 2011) are measured by raw grades in some meaningful way.

As always, one needs to be cautious in generalizing from results for a single institution. Perhaps the results are overstated because the University of Toronto is anomalous, but we suspect not as there are a number of commonalities with other studies especially with regard to grade differentials across majors and the influence of unadjusted grades on admissions. It seems equally possible to us that the findings here could understate the effects of using unadjusted GPAs. Our intuition is that inter-disciplinary differentials in grading standards could be more severe in an institution with high grade inflation (Angling and Meng 2000). The University of Toronto does not have a reputation for grade inflation, unlike other major universities such Prince-

ton (Foderaro 2010), Duke (Johnson 2003) and Harvard (Rothstein 2001).¹⁰

The fact that Toronto grades on a 100 point scale may also matter; we could imagine that pressure to bump borderline students up into a higher grade is higher with discrete letter grades than with numeric grading.

How one interprets these findings depends on one's goals. If the goal is simply to put together an incoming class of law students who will perform best in law school, then the raw GPA performs well, especially when LSAT scores are controlled for. Our results suggest that students who get good grades will, conditional on LSAT scores, continue to get good grades even though some of their good grades as undergraduates may have been in relatively easy courses.

If the goal is to encourage students to take challenging courses, however, our findings highlight flaws in raw GPA measures. The adjusted GPAs are higher, on average, than raw GPAs for students in the sciences and does a better job in predicting performance on an objective test, the LSAT. The adjusted GPA measures do not, however, better predict admission into law school, suggesting a disincentive for students to take challenging courses.

¹⁰We found a small rise in raw grades of around 1 point on the 100 point scale from 2000 to 2010. Some schools, such as Princeton, have taken measures to reign in grade inflation (Foderaro 2010).

What should university administrators do? One route is to report GPAs adjusted for course difficulty. Doing so would encourage students to take challenging classes (or, at least, not discourage them from taking these courses) for two reasons. The first is subjective. Absent an adjustment for course difficulty, students may feel that they get no “credit” for taking difficult courses and shy away from them. The second is objective. If law school admissions used adjusted GPAs in their admissions decision, students would literally get credit for taking difficult courses. In the absence of these distortions on grades across courses or majors, students could pursue courses based on their true interests, without fear of being subsequently punished by graduate schools or employers for taking challenging courses.

This approach has several attractive properties for university leaders. First, it could boost enrollment in the hard sciences by reducing grade-induced student flight from technical fields in the first year of college (Chang, Cerna, Han, Senz, 2008). Second, adjusted grades are not deflated grades. One of the big problems in efforts to fight grade inflation is that students feel (with some legitimacy) that they will be hurt on the job and graduate school markets if their school has lower grades than their peer institutions (Foderaro

2010; Swift, Moore, Sharek and Gino 2013). Grades that are adjusted for course difficulty can be calculated in a way that maintains whatever overall average administrators feel is appropriate for their students. Our adjusted GPA, for example, had the same average as the raw GPAs; the difference is that our adjusted GPA accounted for differential course difficulties, unlike the raw GPA.

Some university leaders may believe that their institutions are not ready to report adjusted grades on transcripts. Our findings suggest that these leaders should still consider using adjusted grades for diagnostic purposes. GPA adjustments and course-difficulty estimates can be fairly easily estimated with an IRT model each year. These results can then be used to assess whether certain classes or majors exhibit substantial differences in course difficulty relative to others. Academic leaders could then take steps to address the disparities, thereby promoting fairness in grading and helping the institution maintain high academic standards.

Bibliography

- Achen, Alexandra C. and Paul N. Courant. 2009. What are Grades Made Of? *Journal of Economic Perspectives*. 23(3): 77-92.
- Almlund, Mathilde, Angela L. Duckworth, James J. Heckman and Tim Kautz. 2011. Personality Psychology and Economics. *Handbook of the Economics of Education*, Eric A. Hanushek, S. Machin and L. Wmann, eds. Amsterdam: Elsevier.
- Angling, Paul M. and Meng, Ronald. 2000. Evidence on Grades and Grade Inflation at Ontario's Universities. *Canadian Public Policy*. 263: 361-368.
- Artz, Benjamin and David M. Welsch. 2013. The Effect of Student Evaluations on Academic Success. *Journal of Education Finance and Policy*. 100(1): 100-19.
- Arum, Richard and Josipa Roksa. 2011. *Academically Adrift: Limited Learning on College Campuses*. Chicago: University of Chicago Press.
- Bailey, Michael A., Anton Strezhnev and Erik Voeten. 2014. Estimating Dynamic State Preferences from United Nations Voting Data. Manuscript, Georgetown University.
- Bar, Talia, Kadiyali Vrinda and Vrinda Zussman. 2009. Grade Information and Grade Inflation: The Cornell Experiment. *Journal of Economic Perspectives*. 23: 93-108.
- Betts, Julian R. and Jeff Grogger. 2003. The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry Level Earnings. *Economics of Education Review*. 22: 343-354.
- Boatright-Horowitz, Su L. and Chris Arruda. 2013. College Students' Categorical Perceptions of Grades: It's Simply 'Good' vs. 'Bad.' *Assessment & Evaluation in Higher Education*. 38(3): 253-259.
- Brown, Philip H. and Nicholas Van Niel. 2012. Alternative Class Ranks Using Z-Scores. *Assessment & Evaluation in Higher Education*. 37(7): 889-905.
- Chang, Mitchell J., Oscar Cerna, June Han, and Victor Saenz. 2008. The Contradictory Roles of Institutional Status in Retaining Underrepresented Minorities in Biomedical and Behavioral Science. *Journal of Higher Education*. 31: 433-464.
- Clinton, Josh, Simon Jackman and Doug Rivers. 2004. The Statistical Anal-

- ysis of Legislative Roll Call Data. *American Political Science Review*. 98(2): 355-370.
- Eaton, B. Curtin and Mukesh Eswaran. 2008. Differential Grading Standards and Student Incentives. *Canadian Public Policy*. 34(2): 215-236.
- Felton, James and Peter T. Koper. 2005. Nominal GPA and Real GPA: A Simple Adjustment that Compensates for Grade Inflation. *Assessment & Evaluation in Higher Education*. 30(6): 561-69.
- Foderaro, Lisa W. 2010. Type-A-Plus Students Chafe at Grade Deflation. *New York Times*. (January 29).
- Glenn, David. 2011. Default Major: Skating Through B-School. *New York Times*. (April 14).
- Johnson, Valen. 1997. An Alternative to Traditional GPA for Evaluating Student Performance. *Statistical Science*. 12: 257-278.
- Johnson, Valen and James Albert. 1999. *Ordinal Data Modeling*. New York: Springer-Verlag.
- Johnson, Valen. 2003. *Grade Inflation: A Crisis in College Education*. New York: Springer-Verlag.
- Nace Research. 2010. *Job Outlook 2011*. National Association of Colleges and Employers.
- Nieswiadomy, Michael. 2014. LSAT Scores of Economics Majors: The 2012-13 Class Update. *Journal of Economic Education*. 45(1): 71-73.
- Nieswiadomy, Michael. 2010. LSAT Scores of Economics Majors: The 2008-9 Class Update. *Journal of Economic Education*. 41(3): 331-333.
- Oettinger, Gerald S. 2002. The Effect of Nonlinear Incentives on Performance: Evidence from 'Econ 101.' *Review of Economics and Statistics*. 84: 509-517.
- Patz, Richard J. and Brian W. Junker. 1999. A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models. *Journal of Educational and Behavioral Statistics*. 24(2): 146-178.
- Payscale. 2013. 2013-2014 Payscale College Salary Report. Accessed May 14, 2014 at <http://www.payscale.com/college-salary-report-2014/majors-that-pay-you-back>.
- Rothstein, Richard. 2001. Doubling of A's at Harvard: Grade Inflation or Brains? *New York Times*. (December 5).
- Sabot, Richard. and John Wakeman-Linn. 1991. Grade Inflation and Course

- Choice. *Journal of Economic Perspectives*. 5: 159-170.
- Stinebrickner, Ralph and Todd R. Stinebrickner. 2014. A Major in Science: Initial Beliefs and Final Outcomes for College Major and Dropout. NBER working paper 19165.
- Swift, Samuel A., Don A. Moore, Zachariah S. Sharek and Francesca Gino. 2013. Inflated Applicants: Attribution Errors in Performance Evaluation by Professionals. *PLoS One*. 8(7): doi:10.1371/journal.pone.0069258.
- Young, John W. 1990. Adjusting the Cumulative GPA Using Item Response Theory. *Journal of Educational Measurement*. 27: 175-186.
- Zangenehzadeh, Hamid. 1998. Grade Inflation: A Way Out. *Journal of Economic Education*. 19(3): 216-26.