Combinatorics, mean convergence, and grade-point averages

Glen R. Waddell and Robert McDonough *

October 2022

Preliminary.

Abstract

Grade-point averaging is fundamentally a combinatorics problem, which challenges inference that relies on the comparison of students with similar GPAs. In the context of a regression discontinuity, we show that it is at smaller bandwidths and with fewer classes contributing to GPA that researchers are most exposed to this sensitivity. While larger bandwidths therefore shield the estimator from this challenge, we show that this accommodation relies on having sufficient overlap of student types at a given GPA. In the end, the ability to overcome the challenges associated with combinatorics therefore depends on how much noise there is in the distribution of GPAs.

Keywords: gpa, grades, program evaluation, random forest, regression discontinuity

JEL: I21, I26, C21

* Waddell (waddell@uoregon.edu) is a Professor of Economics at the University of Oregon and Research Fellow at IZA Bonn. McDonough (rmcdono2@uoregon.edu) is a Ph.D. student at the University of Oregon.

1 Introduction

We deconstruct the common practice of grade-point averaging to illustrate how variation in a set of grade-point averages (GPAs) can challenge intuition about local comparability. In so doing, we identify how empirical tests that rely on the comparability of students with similar GPAs can often be misleading.

Notably, there is little if any guidance in the literature regarding the use of GPA in the implementation of policy or allocation of resources. Yet, important decisions are often made around small differences in GPA. Without a full appreciation for how students end up on one side of a required GPA or the other, for example, admission decisions, probation decisions, and scholarship qualifications are often implemented with GPA cutoffs or minimum-GPA mandates. What majors are available to students, or what opportunities will be made available to them can likewise be determined by relatively local considerations of GPA—and often quite early in a college student's career, which will matter. Upon graduation, students also ask potential employers to interpret GPA, and those employers are presumably induced into rank ordering the relative merits of each applicant's GPA compared to others. These too may be marginal decisions of the sort that are benefited by better understanding what it means to have a "slightly higher" or "slightly lower" GPA. When is it meaningful, and in what ways?

We will demonstrate that while large differences in GPA reliably signal differences in average student ability, due to combinatoric sorting, average ability regularly decreases in GPA—the process is highly nonlinear, with discontinuities especially evident when the number of contributing classes is small. This has immediate implications for policy that relies on GPA—it suggests that there are potential welfare benefits associated with reimagining the adoption of minimum-GPA cutoffs, for example.[1] At the same time, researchers use GPA-based rules in the evaluation of education policy, making the variation in GPA (both across students and as it evolves with additional classes being contributed) important to understand. Regression-discontinuity designs, for example, fashion empirical tests out of environments where students have been treated differently on either side of a given GPA. Without a coincident discontinuity in the type of student on either side of a GPA threshold, the intuition follows that the difference in outcomes across that threshold is reasonably interpretable as induced by that difference in treatment. Yet, if the very construction of GPA supports the *expectation* that there are systematic discontinuities in student type around given GPAs, then smoothness in type

[1] For example, this suggests that the average ability of applicants with 2.99 GPAs can be *higher* than that at 3.00.

through the treatment threshold is in jeopardy.

Thus, acknowledging that large differences in GPA still signal the expected differences in the types of student generating those GPAs, we aim to demonstrate that local comparisons of GPAs do not reliably do the same. In doing so, we draw on two characteristics of grade-point averaging that are fundamental to decoding variation in GPA.

First, we consider the underlying *combinatorics* problem within grade-point averaging—the process by which discrete grades feasibly combine into aggregate GPAs. In particular, the challenges posed by combinatorics materialize insofar as grades are informative of a student's type, but the aggregating of those grades into GPAs ends up aligning students with different sets of course grades into close (or even equivalent) GPAs. For example, a student with grades of C+, B+, B+ (i.e., 2.3, 3.3, and 3.3, in the traditional grade-point rubric) in her first term may be meaningfully different than a student with grades of C+, C+, A+ (2.3, 2.3, and 4.3). Yet, these two students have the same first-term GPA of 2.96. It is also combinatorics that determines that some GPAs are entirely unavailable to students, and that others only become available after having taken a specific number of classes. At the same time, some GPAs remain rare and are only arrived at in very particular (and unlikely) combinations of grade. If students are selecting into individual grades randomly then this is uninteresting. However, assuming that individual course grades are informative of student type (e.g., ability, command of subject), we demonstrate that the underlying combinatoric process is fundamental to understanding variation in GPA.

Second, we consider *convergence in mean* whereby a student's GPA better represents his or her ability as they engage in additional classes. In a way, this is not unlike any stochastic process that informs better with a larger number of draws. However, with important decisions often being made early in a student's education, the variation in GPA across the number of classes contributing to a student's GPA has consequences. While this process will somewhat secondary in importance, it will play an important role in the ability of researchers to accommodate combinatorics in GPA.

In what follows, we will consider variation in GPA using traditional methods as well as with modern machine-learning approaches. Machine methods will prove to be particularly well suited to unlocking the complexity of GPA. Thus, while we will argue for an extra measure of care when interpreting variation in GPA, we will also highlight a productive best response to the underlying complexity in GPA data. In particular, while combinatorics challenges linear estimators, it will directly advantage machine classifiers—combinatorics are systematic, so learnable with enough data. Thus, while we tell a cautionary tale, we will also exemplify recent developments in machine learning and how those tools perform.

Random forests (RF), for example, can flexibly incorporate a student's entire transcript of grades in a prediction exercise. Rather than making comparisons between students at adjacent GPAs, an RF learner can distinguish different types of student even when they have the same GPA. That is to say, the paths to a given GPA are learnable, and to the extent that outcomes are associated with *how* students arrive at their GPAs, the flexibility offered by machine methods will leave us with a better understanding of student heterogeneity. The intuition is not always straightforward, but it's refined somewhat by considering the combinatorics at play. For example, consider the problem of classifying two students who have the same GPA despite only one of them having an F on his transcript. All else equal, this F would typically be associated with lower ability. Yet, conditional on having the same GPA, having an F a student's transcript increases the likelihood that that student is *higher* ability. (With a failing grade included, in order to have achieved the same GPA as others one must have received relatively better grades in other classes.) Ultimately, having even the exact same GPA is not sufficient to satisfy the "all else equal" comparison we desire, as it ignores potentially meaningful difference in the grades that aggregated to that GPA.

In Section 2 we discuss mean convergence and combinatorics in the context of grade-point averaging. In so doing, we illustrate these sources of concern for estimators that rely on localness. In Section 3 we then simulate course-taking behavior, which allows us to observe the properties we highlight within a known data-generating process (i.e., the distribution of a student's *potential* grades, in particular). There, we also demonstrate several implications for the use and interpretation of GPA. In Section 4 we then consider how machine learning can be useful when there is incomplete information about student ability and evaluators must rely on GPA to distinguish students. We first demonstrate the advantage of machine-learned approaches with a random forest predicting outcomes as a function of student performance. Random forests are known to be good predictors so that it does well is expected, and learns that there can be students of different ability at the same GPA. We then extend this framework to consider the casual forest of Wager and Athey (2018), which further demonstrates the advantage of machine-learned methods when GPA is central to a research design. We draw conclusions in Section 5.

2 Two characteristic components of GPA

2.1 Mean convergence

Consider a simple data-generating process where high-ability students draw grades from distributions that dominate those of low-ability students (in a first-order stochastic sense), but that they can still receive low grades with some probability. Specifically, we initially imagine that there are "C" students who receive course grades uniformly from the set {D+, C-, C, C+, B-}, and "B" students who receive course grades uniformly from the set {C+, B-, B, B+, A-}. With the associated grade points of {1.3, 1.7, 2.0, 2.3, 2.7} and {2.3, 2.7, 3.0, 3.3, 3.7}, respectively, a "C" student's expected grade is 2.0, and a "B" student's expected grade is 3.0. Notably, the overlap in potential grades allows for "C" students to outperform "B" students in a given class—it's with additional classes that "B" students are increasingly likely to be observed with higher grade-point *averages*. This is how aggregating course grades to GPAs is informative. However, as GPA becomes increasingly informative of student type we are likewise increasingly confident that students on either side of a given GPA are different, on average. As the number of contributing classes increases, there's simply less noise in GPA as a signal of type and students have separated more across GPA.

This raises an interesting risk in employing regression discontinuities, for example, where the identifying assumption is that potential outcomes are smooth at the discontinuity. To have comparability on both sides of a threshold, similar types of student must appear on both sides of a treatment threshold, and in similar proportion. However, mean convergence implies that this overlap is not guaranteed. In particular, the degree to which different student types "overlap" with each other in the distribution of GPAs depends on the number of classes students have taken.

To demonstrate this intuition we allow for four types of student, each drawing grades uniformly from distributions that center on grade-points of 1, 2, 3, or 4. Similar to the types defined above, we allow grade points that are within ± 1 of their central tendency. We also assume that students of different ability levels are level-different in expected outcomes, with higher-ability students experiencing better outcomes on average. Specifically, we assume that "D" students have outcomes described by $\sim N(10, 5)$, "C" students have outcomes described by $\sim N(20, 5)$, "B" students have outcomes described by $\sim N(30, 5)$, and "A" students have outcomes described by $\sim N(40, 5)$.

We depict this data-generating process in Figure 1, where we simulate the course taking of 125 students of each type. (While we distinguish student types with color, student heterogeneity is presumably unobservable *ex ante*.) Despite there being no treatment occurring anywhere in the data-generating process, we proceed to estimate discontinuities in each panel of Figure 1, as though one was inquiring into evidence of treatment at some GPA with a discontinuity estimator. In Panel A, for example, students have drawn four classes and we consider whether one could establish evidence of "treatment" having fallen on students with GPAs at or above 2.50. Even on visual inspection, we see that there are "true 2s" on the right of 2.50—this is beneficial to the estimator, as an abrupt change in the makeup of students at 2.50 would be troubling. Despite level differences in outcomes across student type, fitting $y_i = f(\text{GPA}_i)$ on either side of a 2.50 GPA threshold yields a confidence interval within which the true $\beta = 0$ is contained.

In Panel B we consider the same students after they have taken eight classes. As a general rule, we expect that students will begin to separate in GPA as they engage with additional classes—this is clear in Panel B, as there are fewer "true 2s" above 2.50 in Panel B than in Panel A. More to the point, though, if we fit $y_i = f(\text{GPA}_i)$ to the same students just one-semester later we see the beginning of mistaking what is unobserved heterogeneity in type (i.e., what we know are just level differences in outcomes in this example) for something that looks like a discontinuity in y_i. If we tightened up around the 2.50 threshold, as in Panel C, we would clearly identify a significant discontinuity in outcomes despite there being no treatment at all.[2]

Despite the density of students itself being smooth across GPA, there is no similar assurance that there is smoothness in student *type* at the threshold.[3] What drives the estimated discontinuities in panels A through C is the 2.50 (placebo) threshold relative to the central tendencies of students who are in the vicinity of 2.50. Some are converging to the left of 2.50 while others are converging to the right.

In Panel D of Figure 1 we estimate a discontinuity at a threshold that is safely in the middle of

[2] These are exemplary of the systematic variation in GPA and not meant to be prescriptive of how one models regression discontinuity estimators. For a more flexible environment in which to explore the variation in RD estimates in simulated GPA data, see `https://glenwaddell.shinyapps.io/RD-in-GPA-data/`.

[3] In all panels of Figure 1 we fail to reject that the density is continuous (using the test provided in McCrary (2008)). As noted in Frandsen (2017), the McCrary test can over- or under-reject the assumption of smoothness when the running variable is discrete. However, the test proposed for use with a discrete running variable is also inappropriate in GPA data, as it relies on the assumption that the support of the running variable has equally spaced intervals. Grade points themselves are unequally spaced, and combinatorics yields unequal spacing. See Lee and Card (2008) and Kolesár and Rothe (2018) for discussions of standard-error estimation and the inference problems associated with research designs in which treatment is determined by a discrete covariate. This relates to our discussion as GPA should arguably be considered discrete data—especially in small numbers of classes.

one type of student. Moreover, we choose a small-enough bandwidth that there is only rare weight on other student types when fitting $y_i = f(\text{GPA}_i)$. The implications of mean convergence should be less evident here—as one might expect, the confidence intervals overlap and the traditional RD analysis cannot reject that $\hat{\beta} = 0$.[4] As there is an absence of treatment in this data-generating process, there is a sensitivity in $\hat{\beta}$ across the panels of Figure 1 that is generally disconcerting—and more so, given that it is systematic with the number of classes students have taken and the alignment of placebo thresholds with respect to the central tendencies of student types.

We demonstrate the variability in treatment estimates more generally in Figure 2, where we consider placebo tests at every GPA in increments of 0.01, in each case adopting the optimally chosen bandwidth (Imbens and Kalyanaraman, 2012). With no treatment anywhere within the data-generating process, confidence intervals should generally include zero. Yet, as students draw additional classes and their GPAs converge to their central tendencies, the risks associated with mean convergence become apparent—point estimates are deviating from true β, and with increasing frequency at larger numbers of classes. In this simulated environment (where we actually know the central tendencies of student types) we see that the biases are largest at the placebo thresholds that tend to separate students, leaving the average type of student quite different on the left and right sides of the cutoff.

2.2 Combinatorics

GPAs are contributed to by various combinations of course-level grades. In order to consider the process by which this evolves, we begin by deriving discrete probability density functions (PDFs) for any number of classes and for varieties of grading curves. We then demonstrate two particularly relevant determinants of this sorting of students into GPAs—the grading rules students can experience, and the number of classes students have experienced. In all cases, we will restrict our attention to GPAs measured at a 0.01 level of precision.[5]

[4] In particular, in Panel D we simulate students between their first and second years of classes, around a GPA of 3.00, where "B" types are equally likely to be on the left and right sides of the discontinuity we estimate (and smoothness *in type* is reconstituted).

[5] While reporting GPAs at 0.01 precision is most common, some institutions do report GPA to a precision of 0.001. Adding precision in this way can exaggerate the combinatoric complexity, but does not change the concerning implications of combinatorics as we describe here.

Setup

Consider the discretized "grade points" that exist as contributions to a GPA. The traditional letter grades of $\{\mathrm{F}, \mathrm{D}, \mathrm{C}, \mathrm{B}, \mathrm{A}\}$, for example, are often notated in grade points as

$$\Gamma = \{0, 1.0, 2.0, 3.0, 4.0\}, \tag{1}$$

with student performance mapping into this scale according to a rule, or "curve." For our purpose, a curve amounts to a rule by which a distribution of student performance maps into a distribution of letter grades. However, the realization of a letter grade by a student can be thought of as a stochastic process, with probabilities assigned to the likelihood that a student will draw each grade-point in Γ. In Figure 3 we produce several PDFs that define various class-level curves. In Panel A, for example, we plot the probability distribution of potential grade-points for an individual student who faces the letter grades in Γ with probability weights

$$\gamma = \{.05, .10, .30, .30, .25\}. \tag{2}$$

In this example, a student earns an A with 25 percent probability, a B with 30 percent probability, and so on. In the remaining panels of Figure 3 we offer a menu of other potential PDFs over grade outcomes.

In Figure 4 we then consider how combinatorics influences the distribution of potential grade-point averages as a student draws *multiple* grades from Γ according to the probability weights γ. Across panels in Figure 4 we plot PDFs over GPAs at the end of two classes, at the end of the first and second semesters (4 and 8 classes), at the end of two years (16 classes), and at the end of four years (32 classes). What is made clear in Panel B, for example, is that there is no possible combination of two draws from Γ in Equation (1) that yields anything other than a GPA from the set

$$\Gamma_2 = \{0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}, \tag{3}$$

notating the number of draws (i.e., the number of classes taken) in the subscript. Given probability weights γ (from Equation 2), we can attach weights to each of these GPAs in Γ_2 quite easily. (Panel B represents *two* draws with replacement from Γ with probabilities γ.) As such, we account for all paths by which one can arrive at the same grade-point average. For example, receiving an "A" in the first class and a "C" in the second combine for a grade-point average of 3.00, as would receiving a "B" in both classes, or a "C" in the first class and an "A" in the second, etc.[6] In the remaining panels of Figure 4 we fill in the potential grade-point averages that can occur with repeated draws. By the end of four years of classes (Panel F), we have captured the full discrete probability density function implied by the 376,992 potential grade combinations from 32 draws.[7]

As with many combinatorics problems, seeing through the combinations quickly becomes intractable—this is true even in this simple example, where we do not allow for any variation in Γ or in γ, and do not consider unequal increments in grade point as will occur when we introduce plus/minus modifiers.[8] Though seeing through the problem is challenging, what is important is to recognize that in the complexity of it all there is still something *systematic* about the evolution of GPA for individual students over time, and in the resulting variation it produces in GPAs across students.

Non-random sorting of students into GPA

Given the role of combinatorics on the evolution of a given student's potential GPAs, we now explore how students sort into GPAs as they accumulate classes. We will consider two natural sources of variation in the accumulation of grades: (i) variation in the grading curves students are exposed to, and (ii) variation in the number of classes students have taken. With those mechanisms established, if there is *any* non-random selection of students in either way (i.e., into classes/curves, or number of classes) we will arrive at particular examples that give empirical context to the sorting problem. For example, if students of varying ability levels sort into classes with different curves, or if students of varying ability levels sort into taking additional classes, comparisons across similar GPAs will be

[6] Below, we consider variation in the underlying PDFs. However, we abstract away from the potential that courses have unequal credit-hour weights, though that would likewise introduce a source of potential heterogeneity if different types of student engage in different course loads.

[7] Specifically, if there are k potential grade points in Γ, there are k^n grade permutations that a student can receive over a sequence of n classes. As an unordered sample with replacement, the set of possible grade-point combinations is then

$$\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!\,(n-1)!} \,.$$

The set resulting from this operation is referred to in combinatorics as a multiset. See Brualdi (2009) for more on multisets and their properties.

[8] Were we to add the traditional plus/minus modifiers to Γ over 32 classes, the resulting combinatorics problem yields 51,915,526,432 unique grade combinations—the distribution of potential GPAs also fills in faster than in Figure 4, with positive weight on 430 of the 431 GPAs from 0.00 to 4.30 by the end of 32 classes. In the simpler example, with only five distinct letter grades, at two-decimal places of precision there is positive weight on 145 of the 431 two-digit GPAs between 0.00 to 4.30.

confounded by student type.

Variation in grading curves

In Figure 3 we produced the discrete probability density functions that define several class-level curves. In Figure 5 we reconsider these in various pairwise comparisons, as though there are two types of student in the data-generating process and we're interested in how the probability of one or the other type being at a given GPA changes across GPA. In all cases, we compare densities at the end of one year of (eight) classes for two types of student, across GPAs of 0.00 to 4.30 (in 0.01 increments). For example, in Panel A of Figure 5 we consider the density of student types when one type experiences plus/minus grading and the other faces a similar distribution but without plus/minus grading (i.e., comparing panels A and B of Figure 3). In Panel B we consider student types across GPA when both types experience "triangle" distributions but with different modal grades (i.e., comparing those in panels C and D of Figure 3). That the probability that a student is one type or the other changes across GPA, generally, is to be expected. However, as each underlying PDF dictates its own combinatorics, what is noteworthy across panels in Figure 5 is that this probability does not change smoothly across GPA. As *smooth* changes in the composition of students are more easily accommodated, this non-monotonicity will be troubling. Through the combinatorics of GPA, variation in the curves experienced can trigger complex and irregular variation in the student types that occupy neighboring GPAs—to assume that two students with "similar" GPAs are actually similar in type is at odds with combinatorics.

Variation in the number of classes

In Figure 6 we consider one of the comparisons of Figure 5 (i.e., Panel C) while varying the number of classes explicitly. As the number of classes increases the pattern of non-monotonicity clearly changes—while there are still non-monotonicities evident after two years of classes (i.e., 16 classes), by the end of three years of classes (24), student types is changing monotonically across almost all GPAs.[9] What becomes apparent, generally, is that the non-monotonicities induced by combinatorics are a "small n" problem, of sorts, but where n is the number of classes. In education environments, however, where important decisions are often made well before 24 classes, this is not easily ignored. Major choice, for

[9] Small non-monotonicities actually occur 7 times in Panel D (out of a possible 430 GPAs), but only at extremely high (e.g., 4.24) or low (e.g., 0.14) GPAs. Further, at these non-monotonicities the fractional change in type is vanishingly small (i.e., on the order of 10^{-22}).

example, will typically occur well before students have taken enough classes to not worry about this "small n" problem. Moreover, schools and departments might admit students who earn a GPA of 2.8 or higher across very few introductory courses. For example, Bleemer and Mehta (2020) exploits UC Santa Cruz's 2.80 GPA threshold over three introductory economics classes to evaluate the return to an economics major. The number of classes required to smooth out the distributions of student type will depend on the underlying grading distributions, of course. However, as a general rule, the less overlapping are the grade distributions of student type (i.e., the less overlap there is in the potential grades that an H or L type draw from) the sooner we see the relationship between GPA and average ability become smooth.

In Figure 7 we consider variation in the number of classes from a different perspective—comparing students who have accumulated a slightly different number of classes at several benchmarks in their academic careers. In so doing, we assume that both students face the same PDFs governing class-level grades and differ only in the number of classes they have taken.[10] Across panels, we mimic what we anticipate researchers or practitioners doing—assuming that two students are comparable at various points in time during their tenures, without regard for the number of classes they may have taken. For example, in Panel B we consider two students who are both at the end of one semester of coursework with one of them having taken four classes while the other having taken five. Likewise, in Panel C we consider two students who are both at the end of one year of coursework, but with eight or nine classes contributing to their GPAs. At each GPA between 0.00 and 4.30 (in 0.01 increments) we plot the probability that a student with that GPA has taken nine classes. If this probability was smooth through the range of GPA, or even locally smooth in places, we would be less concerned—again, we are used to accommodating smooth changes in the fraction of students who are of a particular type by simply controlling for GPA (e.g., as the running variable in an RD). However, we again see that the combinatorics of GPA yields a significant amount of troubling variation in student type across the domain space of GPA given this DGP. Across two otherwise-identical students, even taking one additional class can fundamentally change which GPAs are even possible.[11] Given the resulting non-monotonicity, if there is any degree of non-random selection into taking an extra class here or there, then it is reasonable to anticipate that these students are selecting into distinct sets of potential

[10] In Figure 7 we use the discrete probability density functions from the 13-point grade distribution in Panel D of Figure 3. However, the results are representative of the comparison that could be made for any class-level PDF.

[11] While we do not show this, a similar problem arises from students taking different numbers of credit hours.

GPAs—this requires attention beyond our typical approach to policy evaluation.

What if different student types select differentially on these margins?

In figures 3 through 7 we offer examples of a more-general problem associated with interpreting GPAs. In the end, as the sorting of students into GPA is governed by combinatorics, if there is any meaningful heterogeneity across students that correlates with either the grading regimes they experience or the number of classes they've taken by a point in time, then we should expect that inference that relies on comparisons (or similarity) of GPA will be challengeable by combinatorics-induced sorting.

While the above figures make pairwise comparisons, and in that way capture the roles of various contributors as comparative-static exercises, it's more likely that students select into a variety of classes, with different curves (i.e., different PDFs). In Figure 8 we plot the proportional breakdown of many students and types. In Panel A there are six types of student, for example, with each having experienced classes like those we imagined earlier Figure 3. Again, non-monotonicity is evident. In Panel B we go further, imagining that there are 13 student types, each having a different modal grade but drawing from a similarly shaped PDF (i.e., triangle distributions akin to those in Panels C and D of Figure 3). Even as we increase the number of student types populating the GPA domain, discrete changes in student composition across local changes in GPA persist.

3 Evaluating GPA-determined treatment

3.1 Setup

Given the construction of GPA, and the non-random sorting into local GPAs in particular, identifying unbiased estimates of treatment in GPA data is non-trivial—this is especially true the more local is the identifying variation. As demonstrated in Section 2, a clear violation exists in designs that rely on smoothness around a treatment threshold, for example, and the non-monotonicity in student type across GPAs should give pause as we consider experiments with GPA as a running variable. (Moreover, with interest in collapsing on *smaller* bandwidths where power allows, we should be particularly concerned that the implications of combinatorics around treatment thresholds could lead to questionable inference from well-powered RD designs.)

In the sections below we consider two thought experiments. First, we consider the scenario we feel is more relevant in practice—unobservable student heterogeneity, where their type is only learned over time as they draw from better and worse grade distributions. Second, we consider the scenario where type is observable—here we will assume that students draw from *the same* grade distribution, but will have some students take more classes. This will both demonstrate the importance of considering the component parts of GPA and solidify some intuition around the ways in which identifying treatment in such an environment might be salvageable.

In both scenarios we envision a data-generating process in which there are two types of students who differ in their average outcomes—we keep with the idea that in the population of students there are L types and H types. In particular, suppose that each student realizes some outcome—here, we will simulate weekly wages, w_i—according to the simple process,

$$w_i = \alpha + \delta \mathbb{1}(\mathrm{H}_i = 1) + e_i \ . \tag{4}$$

The parameterization of (4) will be immaterial, so we roughly mimic the 25^{th}, 50^{th}, and 75^{th} percentiles of weekly incomes among college graduates in the United States in 2020, assuming that H types experience δ-higher average wages.[12] Other than from the level differences in outcomes associated with being an H or L type, then, wages vary randomly. To be clear, there is no treatment-induced variation in outcomes, so we are in an environment in which well-identified models should fail to reject the null hypothesis that there is no systematic discontinuity in outcomes. Nonetheless, adopting traditional methods in such an environment exposes researchers to the risk of identifying a discontinuity in w_i, as combinatorics facilitates a source of non-random selection into GPA by type of student.

3.2 When student heterogeneity is only partially observable

Here, we assume that H types draw grades from a distribution that first-order stochastically dominates that of L types, with modal (mean) grades at the individual class level of 2.3 and 3.7 (2.23 and 2.68), respectively.[13] In what follows, we produce regression discontinuity estimates from 1000 simulations

[12] Assuming $\alpha = \$1,133$, $\delta = \$566$, and $e_i \sim N(0, 300)$ centers our DGP on the weekly incomes of college graduates, approximating the first quartile (\$977), median (\$1,416), and third quartile (\$2,110) of weekly income of college graduates, according to the Usual Weekly Earnings of Wage and Salary Workers section of the Current Population Survey. (See Bureau of Labor Statistics (2020) for details.)

[13] Specifically, L types draw grades from the 13-point plus/minus letters as in Panel C of Figure 3, and H types draw "better" grades, on average, as in Panel D of Figure 3. In other words, L and H types draw from $\Gamma = \{0, 0.7, 1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 4.3\}$ with probabilities $\gamma_L = \{0.024, 0.058, 0.073, 0.088, 0.107, 0.122, 0.136, 0.113, 0.095, 0.077, 0.053, 0.036, 0.018\}$ and $\gamma_H = \{0.016, 0.038, 0.048, 0.057, 0.07, 0.08, 0.089, 0.102, 0.111, 0.121, 0.134, 0.089, 0.045\}$, respectively. That the mean GPA from a grade distribution with mode of 3.7 is 2.68 reflects top-coding, as discussed in Appendix **??**.

of 30,000-student panels, inquiring into whether there is evidence of a systematic discontinuity in outcomes for those with GPAs of 2.50 or above. While we consider a GPA cutoff of 2.50, the issues we illustrate generalize to any GPA where students of different type are present. We choose 2.50 as it falls between the mean grades of L and H types, where we can anticipate that there might be interest in a rule that separates students. As we initially allow students to have taken four classes, in that sense it's fitting to have in mind the sort of decisions that are made around the middle of the first year of college (e.g., major choice) or policies that tend to be most binding in the first year of college (e.g., the initiation of probationary status, admittance into greek affiliations, enrollment into professional schools).

In Panel A of Figure 9 we plot the mean point estimate (across simulations) associated with each bandwidth between 0.01 and 0.50 in increments of 0.01. The absence of any treatment in the DGP should have us anticipate estimated discontinuities of zero. However, consistent with combinatoric sorting, there is significant sensitivity evident in the estimated treatment effects at smaller bandwidths. In fact, the bias tends to be both large and sensitive to changes in bandwidth, and is directly resulting from combinatorics-induced violations of the required smoothness assumption. In an environment where average weekly wages are \$1,416, RD estimates across bandwidths range from \$-105 (-0.25σ) to \$47 (0.11σ).

In Panel B of Figure 9 we see how this sensitivity arises—the combinatoric sorting of different student types around the treatment threshold. In particular, in Panel B we plot the fraction of students who are H types across the same range of bandwidths (0.01 through 0.50) both to the left- and right-hand sides of the 2.50 cutoff. The lumpy introduction of H and L types included in the sample as the bandwidth changes is clearly evident, and around small bandwidths in particular—other than noise, this imbalance is the only factor that drives estimates away from zero in Panel A. We see this from a different perspective in Panel C, where we plot counts of student type across bandwidths. This makes the source of the lumpiness is particularly evident. While there are bandwidth adjustments that do not trigger changes in the number of observations at all—by its nature, combinatoric sorting will leave behind GPAs that are not occupied—when adjusting the bandwidth allows "new" GPAs into the estimator there are discrete changes in the number of H and L types. It is at these same GPAs that mass shifts discretely to one type of student or the other—this abrupt tipping of the balance one way or the other explains fully the change in point estimates across bandwidth.

Note, importantly, that even with combinatorics playing an active role in facilitating the non-random sorting of student types into GPA, the overall density of students can still itself to the appearance of smoothness in the aggregate. In fact, our DGPs routinely pass standard tests for changes in the density of students around the threshold (i.e., McCrary 2008; Frandsen 2017).

In Figure 10 we produce similar plots of bandwidth sensitivity tests as students take additional classes, mimicking their progress through the institution and the potential for similar evaluations being performed at the end of two classes, one semester, one year, and two years. This highlights that the non-random sorting is particularly egregious when students have taken few classes. That said, estimates are sensitive to combinatorics at smaller bandwidths through the end of one year of classes (in Panel C). Notably, it's here where the tell-tale signs of mean convergence appear, albeit to a lesser degree than in Panel D where students have taken even more classes. This reveals a general pattern in the evolution of GPAs—while the combinatorics problem diminishes with additional classes, student populations also converge to their mean performance. Thus, the domain of GPAs over which there is overlap (i.e., in student type) is getting increasingly narrow. While the smallest bandwidths in Panel D evidence "zero" treatment effect, by the time students have taken sixteen classes the degree of mean convergence experienced is such that larger bandwidths now use increasingly different student types to identify the discontinuity.[14] As we should anticipate, this results in biased point estimates.[15]

In Figure 11 we demonstrate how the uneven distribution of student types changes as they take additional classes. Similar to the lumpiness we saw in Figure 9, the problematic influence of combinatorics on treatment evaluation shows up where we anticipate it (i.e., panels A and B, and to some extent C) and disappears as combinatorics present less of a first-order concern (Panel D). At sufficiently large numbers of classes—16 in our data-generating process—it matters less that GPA is the product of a combinatorics problem.

In the end, the roles of combinatorics and mean convergence are clearly in tension as students progress through college. The degree to which this tension is felt should inform bandwidth choice. At smaller numbers of classes, estimates are more sensitive to combinatorics—here, larger bandwidths can act as a mitigating device, because they leave parameter estimates less sensitive to the *types* of student at particular GPAs. However, at larger numbers of classes combinatorics induces less local variation in student type and smaller bandwidths become appropriate—here, larger bandwidths expose researchers

[14] Recall Figure 6, where by the time students had drawn 16 classes there was no smoothness violation.

[15] We saw similar evidence in Figure 1, for example.

to lost comparability. Overall, considering the evaluation of treatment that falls on students around some number of classes, the researcher's choice of bandwidth is implying something of a tolerance for combinatorics-related bias over the bias induced by mean convergence. We summarize this tension as follows, highlight the particularly concerning source of bias.

The first-order source of bias in RD designs with GPA data

With fewer classes?	With more classes?
Smaller bandwidths expose identification to the pitfalls of combinatorics. With sufficient overlap in student type, larger bandwidths shield researchers from bias.	The bias from combinatorics likely attenuates with more classes, but as GPA converges to each student's central tendency we tend to lose overlap in type (across GPA).

By choosing smaller bandwidths we are down-weighting the potential that mean convergence will itself bias $\hat{\beta}$ while at the same time up-weighting the potential bias from combinatorics. In choosing larger bandwidths we are down-weighting the potential bias associated with combinatorics while at the same time up-weighting the potential bias from mean convergence. In considering these tradeoffs, the variation in the number of classes taken by students in the population under study cannot be ignored—employing a small-bandwidth design when students have taken many classes implies a lower overall potential for bias, but the equivalent small-bandwidth design will be much more vulnerable to bias when students have taken few classes.[16]

3.3 Does the problem go away if student heterogeneity is observable?

Before moving on to consider potential solutions to the above problem, there is good intuition in considering how the classification of students into type can move us toward better identification. In this section, we recast the problem as one in which both types draw repeatedly from the same grade distribution, but H types simply draw one extra grade (i.e., they take classes at a faster rate) than L types. In this experiment we shut down entirely on any heterogeneity coming from grades

[16] Another important consideration for researchers navigating this tradeoff is that the bias stemming from mean convergence is likely signable. For example, to the extent we've populated the right-hand side of a discontinuity estimator with H types who attain better average outcomes for reasons not associated directly with treatment, we expect $\hat{\beta}$ to be biased up. However, the non-monotonicity in student type introduced through combinatorics implies that there can be discretely more or less of one type on the left or right of *any* threshold. This leaves the resulting bias unsignable, and any associated inference more uncertain.

themselves.[17] This highlights the problem that can arise simply due to the mechanistic sorting of combinatorics, which allows some students to populate GPAs that other students simply cannot. We will then consider the implications of controlling for the "number of classes" in regression analyses.

Given that we rarely control for the number of classes contributing to a student's GPA, the results of this exercise are quite striking. In Figure 12 we estimate "discontinuities" in outcomes at a variety of GPAs (i.e., 2.30 though 3.10 in increments of 0.10). There should be no discontinuities in outcomes in this environment—here there will be, as they are left behind by the combinatoric sorting of students into GPA simply through the inclinations of H types to take one more class.[18]

Consistent with combinatoric sorting, the bias is unsignable in general and particularly problematic at smaller bandwidths where the density of student "types" can be different either side of a given GPA and can change abruptly as different bandwidths allow different GPAs to populate the estimator. Across the nine thresholds we illustrate (between 2.30 and 3.10), the mean bias in point estimates is positive in four of them and negative in five of them. However, in all but two of the nine placebo thresholds the point estimate itself changes sign across bandwidth.[19] Given the combinatorics of grade-point averaging, not controlling for the number of classes that contribute to GPA is clearly problematic. As the number of classes is observable—while not often used, it often is observable—in Figure 13 we plot the bandwidth sensitivity around one of these placebo treatments (a GPA of 2.50, in this case) with and without controlling for the number of contributing classes. Even though students are drawing from the same distribution of grades, combinatorics allows the number of classes taken to transmit through to the set of GPAs H types are able to occupy. Consistent with the omitted variables bias, controlling for the number of courses also resolves the bias in estimated treatment.

While we've designed a simple experiment here, where student type is perfectly inferable through one observable attribute, the reality is that field data rarely presents such a clean opportunity to merely control for differences. Fortunately, however, while combinatoric sorting will defeat local estimators,

[17] All students draw all grades from a triangle centered at 2.70.

[18] Moreover, our simulated environment suggests that the bias can be large in magnitude. Granted, we've constructed this data environment. However, in so doing we've matched mean weekly wages (in 2020 for US college graduates) and 25th and 75th percentiles while assuming a level difference in wages of \$566 for H types. In this environment, mean wages are \$1,416, and RD estimates across the cutoffs and bandwidths range from \$-1,365 to \$483—the associate impacts range from -96.3 percent to 34.1 percent at the mean. That is, the bias associated with not accounting for the number of classes is roughly five times as large as the largest bias we found from not controlling for the difference between a C- student (drawing grades around 2.3) and an A- student (drawing grades around 3.7) when the number of draws was common.

[19] Only in the estimation of a discontinuity at 2.80 and 2.90 do we find the point estimates never changing—not as a rule, of course, but in this DGP.

it is also ripe for the more-sophisticated remedies made available through machine learned approaches to prediction, even when student heterogeneity is not easily inferable. Thus, having documented the potential pitfalls associated with the underlying combinatoric sorting of students into GPAs, below we turn to considering the ways in which the variation in GPA is actually learnable. In short, while the nature of GPA (and the combinatoric problem, in particular) has been challenging thus far, it also leaves behind evidence that can be learned—in effect, evidence that can be used to salvage better inference.

4 Machine-learned approaches to classifying students

The above analysis suggests that we should clearly not anticipate that student ability is smooth across *local* changes in GPA—the non-monotonicity evident above makes it easy to demonstrate that average student ability can even *decrease* in response to small increases in GPA. Moreover, where estimators rely on local smoothness at some threshold, the combinatorics of GPA can be of first-order importance, especially when we fail to recognize that *local* variation in GPA does not control well for student heterogeneity. However, decisions made at marginal considerations of GPA are fundamentally relying on a classification, of sorts. As problematic as GPA can be as a measure of performance—having more-complex variation than we had recognized—it is still *systematic* in its construction. As such, sophisticated methods, instead of succumbing to the complexity, find it something to be learned.

4.1 Predicting outcomes using GPA and course-level grades

In Figure 14 we compare different approaches to modeling the data-generating process of Section 3—a distribution of wages within which there are two level-different types. In all panels we plot the true average wage and the *predicted* wage for each GPA (again, at 0.01 precision). In panels A through C we consider linear approaches to this exercise, projecting w_i onto GPA_i across a variety of polynomials. With enough flexibility linear models can eventually track the *global* non-linearity in outcomes across GPA fairly well. However, even a ninth-order polynomial does a fairly poor job of capturing *local* changes in GPA.

As an alternative to these methods, in Panel D we plot the predictions from random forest regressions of w_i on GPA_i.[20] To be clear, in Panel D we restrict the model to only the single covariate (i.e.,

[20] See Breiman (2001) for details of the random forest algorithm as it applies to regression problems.

GPA) and, even then, the flexibility offered by the random forest is evident—the estimated relationship tracks the true pattern of average wages so well that it is difficult to distinguish the two lines by visual inspection. (We use a dashed line to represent the prediction to help somewhat with this inspection.) Being fully nonparametric, at every GPA the random forest yields a predicted $\hat{w}_i$ that is independent of the predictions at surrounding GPAs. In this setting, then, this flexibility improves performance markedly and the non-monotonicity in the data is as evident in the predictions of the model as is it in the underlying DGP. (This bodes well for our return to considering causal estimation in this environment.) Here, we know that this non-monotonicity is driven by different student types sorting into GPAs through combinatorics. While field data would not afford the same ability to see the source of heterogeneity, a similar exercise with field data would nonetheless capture variation in outcomes that are predictable through the learned combinatorics of GPA—and to the extent learnable, they would mitigate the problems we identify.[21]

However, there is little reason to limit the RF environment to learning simply through GPA when the individual contributors to GPA were available. The obvious return to the addition of course-level information will be through the RF learning to distinguish systematic heterogeneity in outcomes even from students who have *the same* GPA. Without transcript-level information, RF learning is restricted to making one prediction for each GPA, while variation in course-level grades within given GPA is able to predict *different* outcomes for those who have the same GPA but arrived at it differently. In Figure 15, then, we consider the addition of course-level information and the RF's ability to predict outcomes of H and L types. As we know each student's type—to be clear, the RF does not—we report predicted outcomes separately for H and L types. The RF learner can distinguish heterogeneous outcomes among students who share the same GPA—predicted outcomes are higher for H types than for L types more than 90 percent of the time.[22]

[21] In Appendix A we report quantitative measures of model performance across 1,000 simulations of the above process.

[22] In simulating this process 1,000 time at the end of one semester of classes, the RF with GPA and course grades predicts higher average wages for H types at 78.1 percent of GPAs. After one year of classes, the RF predicts higher outcomes for H types at 90.8 percent of GPAs, and after two years of classes predicts higher outcomes for H types at 92.9 percent of GPAs. (Again, these values are based on only the inner-95 percent of the data according to GPA. When we use all of the data, there is a marginal decline in performance, to 70.9, 85.7, and 91.0 percent, respectively at one semester, one year, and two years of classes. This decline is a consequence of the sparsity of students at extremely high or low GPAs, which limits the RF models' ability to distinguish students.)

4.2 Can we exploit this learning in a causal environment?

Above, we have demonstrated the suitability of machine-learned approaches to disentangling the underlying combinatorics within GPA, inclusive of identifying student heterogeneity even among those who share *the same* GPA—we identify heterogeneity through variation in the combination of grades that got them to that GPA, essentially. Thus, it is natural to consider the performance of companion methods in a causal framework.

Here, we simulate post-graduation wages and consider the performance of causal forests with respect to their ability to retrieve the causal parameter of interest. To do so, we make one addition to the environment we've considered above—a treatment that is experienced in some discontinuous way at a GPA threshold, *but available to all students with some positive probability.* As this "overlap" is required in order to satisfy the identifying properties of the causal forest, we will benchmark the casual forest estimators against fuzzy regression discontinuities.

4.2.1 Causal forests

Introduced in Wager and Athey (2018), the causal forest (CF) is an application of machine learning to causal inference in the presence of randomized treatment. CF procedures build on the strengths of random forest learning, which is known to work well as a classifier (Hastie et al., 2017). In particular, these strengths include flexibility in modeling nonlinear processes, and an ability to handle large numbers of covariates. By leveraging these strengths, it is also notable that the causal forest estimates *individual* treatment effects—heterogeneity in the effect of treatment across individuals is then likewise identifiable. A causal forest, then, is simply an algorithm that leverages a random forest's classification strength to group together those observations that are good counterfactuals for each other—having grouped them, we can then estimate the effect of treatment within those groups. In our working example above, given course-level information the RF procedure produced different predicted outcomes, on average, for the two types of student in the DGP—this was true *even when they had the same GPA*. Likewise, then, the RF component within the CF estimator will exploit that there are multiple (and learnable) paths by which students arrive at given GPAs and thereby better estimate their counterfactual outcomes would have been—their outcomes in the absence of treatment.

With identifying assumptions met, a causal forest estimates a treatment effect at every set of covariates x by estimating the mean difference in the outcomes of treated and control units who have those covariates in common. Given this condition, the difference,

$$\tau(x) = E[Y_i^1 - Y_i^0 \mid X_i = x] \ , \tag{5}$$

is commonly be referred to as the *conditional average treatment effect* (CATE), as it is conditional *on a set of covariates*. As we allude to above, we adjust our DGP somewhat to meet the identifying assumptions—"unconfoundedness" and "overlap." Wager and Athey (2018) defines unconfoundedness as the independence of treatment assignment from outcomes, conditional on covariates. This is similar to the conditional independence assumption discussed at length in Angrist and Pischke (2009). Likewise, the overlap assumption should be familiar from matching-type estimators more generally, as the overlap assumption for the CF requires that $0 < Pr(W_i = 1|X_i) < 1 \ \forall \ X_i \in \mathbf{X}_i$. That is, in order to estimate the effect of treatment in our context there must be treated and control observations *within* each GPA. Of note, then, is that the overlap assumption implies that a causal forest cannot estimate treatment in a classic "sharp-RD" design, where the probability of assignment into treatment is zero on one side of a threshold and one on the other.[23]

Borrowing from the weekly wages we simulated in Section 4.1, here we augment our data-generating process to satisfy unconfoundedness and overlap by introducing a degree of noise in treatment assignment. Specifically, those at GPAs below the threshold now randomly experience treatment with 15-percent probability, while those above the threshold randomly experience treatment with 75-percent probability. In Panel A of Table 1 we evaluate the performance of a regression-discontinuity estimator employing either an optimal bandwidth (Imbens and Kalyanaraman, 2012) or a smaller bandwidth (defined as 10 percent of the optimal bandwidth)—both largely fail to identify that the average treatment effect is zero.[24] This is the benchmark against which it is informative to compare the performance of a causal forest.

[23] Other common situations can also violate the overlap assumption. For example, if students with an F on their transcript are ineligible for treatment then the overlap assumption would be violated in a causal forest that ran on course-level grades. In such cases, researchers could resurrect the internal validity of the estimator by limiting the sample to that for which there is overlap.

[24] At the end of one semester (i.e., in Column 1), the RD estimator with an optimal bandwidth leads to a rejection of the null hypothesis in 31 percent of iterations, with an average increase in weekly wages of \$21.91 in weekly wages ($.05\sigma$) relative to the mean wage of \$1416. At the end of one and two years, this rejection rate increases to 56 percent of iterations (0.09σ) and 61 percent of iterations (0.11σ). That the bias is positive on average is merely an artifact of the data-generating process and chosen treatment threshold—the student types at the threshold we consider just happen to split with more H types on the right-hand side of 2.50. At smaller bandwidths, we again see over-rejection of the null hypothesis and biased point estimates, but the bias interacts with the number of classes differently. Applying smaller-bandwidth designs to the same data we reject the null in 13 percent of iterations at the of one semester (0.13σ), in 8 percent at the end of one year (0.11σ), and in 6 percent at the end of two years (-0.003σ).

4.2.2 The CF model appropriately rejects the $\beta = 0$ null where the RD over-rejected

In Panel B of Table 1 we reconsider the same data as was drawn for the exercise in Panel A but with a CF procedure performing the same task. (As in Panel A of Table 1, then, each simulation includes 30,000 students, with grades drawn from the same distributions as in Section 4.1, with H types are again a level-difference higher than L types.) We do this across two models, including only GPA in the first model and then adding individual course grades in the second.[25] Notably, where there is nothing introduced into the data-generating process other than the naturally occurring combinatorics-induced imbalance of type, RD-type methods falsely identify "treatment." (RD methods reject the null 31 percent of the time after one semester, 56 percent of the time after one year, and 61 percent of the time after two years.) CF procedures, however, retrieve treatment estimates that appropriately reject the null (i.e., reject that $\beta = 0$ only five percent of the time). Whether at the end of one term, one year, or two years, the estimated effect sizes identified in the RD estimates are also orders of magnitude larger than those identified in the CF procedure (which are hovering around zero, and less than $.0005\sigma$, on average).[26]

For a more-direct comparison to the regression discontinuity benchmarks, in Panel C Table 1 we consider CF performance when the samples are restricted to observations within the bandwidths we imposed on the RD estimators (in Panel A). Again, the CF estimators do not over reject the null.

4.2.3 The CF model also informs us of underlying heterogeneity in treatment

A unique strength of using a causal forest is that its estimation of conditional average treatment effects can retrieve underlying heterogeneous in treatment effects. Since treatment effects are identified up to every unique x in the data, a causal forest will explore whether treatment effects vary systematically across any combination of covariates. (Indeed, Athey and Imbens (2016) is motivated by a desire to identify such heterogeneity.) This is particularly useful in the setup we've developed, where the

[25] In the random forests of the previous section, we saw evidence that a machine-learned estimator could find heterogeneity between students using only course grades. However, as GPA determines treatment in our DGP and H types earn higher GPAs, on average, unconfoundedness requires that we condition on GPA. (Since type is correlated with GPA, and therefore with treatment assignment, the unconfoundedness assumption is only satisfied if GPA is included in the explanatory variables provided to the CF. This is similar in spirit to including the running variable in RD estimation—capturing the relationship between the running variable and the outcome of interest that exists independent of treatment.)

[26] These CF estimates translate to an estimates treatment effect of roughly \$0.20 over a standard deviation in observed wages of roughly \$412. The most-conservative RD estimates imply a \$20.00 "increase" in wages.

"types" of student might benefit differentially from treatment.[27] Here, we allow H types to respond to treatment—H types receive a $300 increase in weekly wages coincident with treatment. L types, though exposed randomly to treatment as in our baseline DGP, experience no benefit to treatment. In this way, the true parameters are $\beta_H = 300$ and $\beta_L = 0$.

In Panel A of Figure 16 we plot the distributions of CATE estimates for our two causal forests in the presence of this treatment-effect heterogeneity. Assignment into treatment is unconfounded so long as we condition on GPA, meaning that the CF algorithm can identify collections of individuals with the same underlying propensity for receiving treatment (i.e., good counterfactual groups). Thus, the first CF learner, provided with only GPA, still produces unbiased CATEs, even in the presence of treatment effect heterogeneity. But with only GPA as a covariate, the CATEs from this first CF can only recover the average across H and L types at a given GPA.[28] However, given that expected student type changes across GPA, even this CF demonstrates some ability to detect the heterogeneity in β_H and β_L. In the left of Panel A, many students assigned to low treatment effects are in fact low types, reflecting the fact that most "low" GPAs are associated with L types. Likewise, "high" GPAs are associated with H types, on average.

In contrast, the right side of Panel A shows that providing the CF with individual course grades allows it to identify heterogeneity much more cleanly. When provided with course grades, the students with CATEs near zero are overwhelmingly L types and the students with CATEs near 300 are overwhelmingly H types. As with RF classifiers, the CF has minimized the variance of within-group treatment effects when splitting students into groups—given that there are two types of student in our DGP drawing from different distributions of grades, transcript-level grades make the two types distinguishable.[29] That is, the addition of course grades allows the CF to make distinctions between students with the same propensity for receiving treatment (i.e., students at a certain GPA). When the causal forest ultimately groups these students together to estimate treatment effects, it is effectively

[27] With a CATE for every set of x in the data, testing for heterogeneity in treatment effects is then feasible. The formal test for heterogeneity in estimated treatment effects is an implementation of the best linear projection method for detecting treatment heterogeneity in machine learning estimates, proposed by Chernozhukov et al. (2020).

[28] Wager and Athey (2018) demonstrates that when the estimator's identifying assumptions are met (i.e., overlap and conditional unconfoundedness) each $\hat{\tau}(x)$ is point-wise consistent, asymptotically Gaussian, and centered in the sampling distribution. As such, the CATEs can be aggregated to estimate the average treatment effect with consistency and conduct valid hypothesis testing. (In an implementation of the causal forest estimator, Athey et al. (2019) constructs average treatment effects using augmented inverse probability weighting (AIPW). As noted therein, this method generally leads to estimates of average treatment effects that are more accurate than naive averages, based on results from Chernozhukov et al. (2018). We use the implementation of AIPW in Athey et al. (2019) to produce average treatment effects.

[29] We find that, in general, random covariate sampling of roughly 75 percent of available grades allows the CF algorithm to identify heterogeneity in the student population.

grouping students by type, and thereby identifying the underlying treatment-effect heterogeneity.[30]

In Panel B of Figure 16 we plot the distributions of CATE estimates associates with two causal forests with placebo treatment—again, we allow only for GPA, and then for both GPA and individual course grades. To demonstrate the CFs handling of the heterogeneity in treatment effect by type, we identify type of each student assigned to each treatment effect. In Panel B, in particular, we again see the value added in using individual course grades as the estimated parameter collapses on the true $\beta = 0$ with the additional ability to learn through course-level variation in grades.

In scaled simulations of the above experiment, we also formally test for heterogeneity in estimated treatment effects. The test is an implementation of the best linear projection method for detecting treatment heterogeneity in machine learning estimates, proposed by Chernozhukov et al. (2020). In 100 percent of simulations across both causal forest models, the heterogeneity identified is statistically significant.

5 Conclusion

We demonstrate the complex ways in which grade-point averages can challenge causal identification. We model the process through which students arrive at GPAs, highlighting the roles of mean convergence and combinatorics. In the context of treatment evaluation, we then show how these two mechanisms tradeoff as students engage in additional classes and thereby interfere with the interpretation of GPA variation in rather complex ways.

In general, the set of feasible GPAs is governed by the combinatorics of GPA, which determines the paths by which a student can arrive at a given grade-point average. As this is a systematic process, the non-random selection of students into GPAs exposes estimators to non-monotonicities in student type across GPA. Even though collapsing on students with more-similar GPA sounds very much in the spirit of constructing "all-else-equal" conditions, combinatorics is at active in local comparisons, especially where the number of classes is small. In the context of treatment evaluation, we demonstrate that this induces a form of selection through which students on either side of a given GPA can end up being different. This calls into question the validity of identification strategies that rely on local

[30] The fact that both L and H type students have a chance of receiving each letter grade explains why we see treatment effects massed near, but not on, the true β_L and β_H. Since each letter grade has positive probability for both types, so too does each transcript of grades. For instance, an H type student can occasionally draw an unlucky set of grades that would be more common for an L type. This means that every CATE is still an average over both L types and H types.

comparisons. We argue further, that the failure of linear GPA-based estimators is fundamentally due to this type of classification problem. In short, the combinatorics of GPA requires a flexibility in modeling for which linear methods are not well equipped. Machine-learned methods, on the other hand, often excel at modeling such spaces.

This sort of local non-comparability is ameliorated with additional classes, as the artifacts associated with combinatorics tend to subside with repeated draws from a distribution. However, with additional classes, GPAs also converge to students' central tendencies (i.e., to their grade types, in a way). While this increases the informativeness of GPA—as a signal-to-noise ratio, the informativeness of GPA is increasing with additional classes—this can also expose estimators to having different student types on either side of a given GPA. In the limit, even though the density of students across GPA might appear continuous, there is the possibility of losing overlap in student type entirely, especially with many classes contributing to grade-point averages. This highlights a trading off in the informativeness of comparing GPAs across the number of contributing classes, with implications for how local comparisons should be interpreted at various stages of students' academic progress. Moreover, it suggests that larger bandwidths are not a universal fix to combinatorics-induced bias, as this accommodation relies on having sufficient overlap of student types at a given GPA. That said, it is important to note that it is where combinatorics bias is likely to be most egregious (when few classes contribute to GPA) that we can expect the most overlap in student types across GPA.

In the end, we find benefits associated with supplementing traditional methods with machine-learned methods that are capable of exploiting what is learnable through *how* students arrive at given GPAs. The inclusion of transcript-level data (which is easily incorporated into these methods) allows researchers to distinguish heterogeneity in outcomes even among students at the very same GPA, for example. In terms of treatment evaluation, causal forests (Athey et al., 2019) can likewise learn through the combinatorics of GPA to identify not only average treatment effects, but also the heterogeneity in individualized treatment effects.

References

Angrist, Joshua and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, 2009.

Athey, Susan and Guido Imbens, "Recursive Partitioning for Heterogeneous Causal Effects," *PNAS*, 2016, *113* (27), 7353–7360.

_, Julie Tibshirani, and Stefan Wager, "Generalized Random Forests," *Annals of Statistics*, 2019, *5.*

Barreca, Alan I., Jason M. Lindo, and Glen R. Waddell, "Heaping-Induced Bias in Regression-Discontinuity Designs," *Economic Inquiry*, 2016, *54* (1), 268–293.

_, Melanie Guldi, Jason M. Lindo, and Glen R. Waddell, "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification," *The Quarterly Journal of Economics*, 2011, *126*, 2117–2123.

Bleemer, Zachary and Aashish Mehta, "Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major," *American Economic Journal: Applied Economics*, 2020.

Breiman, Leo, "Random Forests," *Machine Learning*, 2001, *45*, 5–32.

Brualdi, Richard A, *Introductory Combinatorics*, Pearson, 2009.

Bureau of Labor Statistics, "Median Weekly Earnings by Education Second Quarter 2020," Technical Report, U.S. Department of Labor 2020.

Butcher, Kristin, Patrick McEwan, and Akila Weerapana, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 2008, *142* (2).

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 2018, *21*, C1–C68.

_, Mert Demirer, Esther Duflo, and Iván Fernández-Val, "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments," Working Paper 24678, National Bureau of Economic Research 2020.

Frandsen, Brigham, "Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design When the Running Variable is Discrete," in Matias Cattaneo, ed., *Regression Discontinuity Designs: Theory and Applications*, Vol. 38, Emerald Publishing Ltd., 2017, pp. 281–315.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 ed., Springer, 2017.

Imbens, Guido and Karthik Kalyanaraman, "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of Economic Studies*, 2012, *79*, 933–959.

Kolesár, Michal and Christoph Rothe, "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review*, August 2018, *108* (8), 2277–2304.

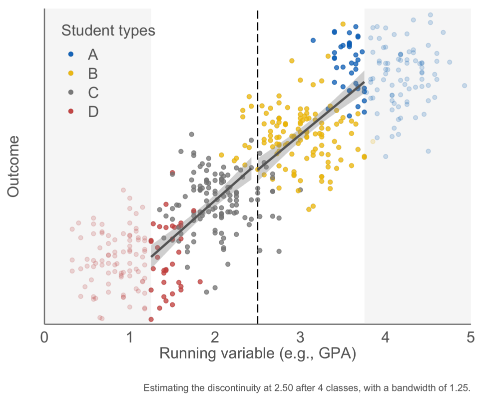
Lee, David S. and David Card, "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics*, 2008, *142*, 655–674.

McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 2008, *142* (2), 698–714.

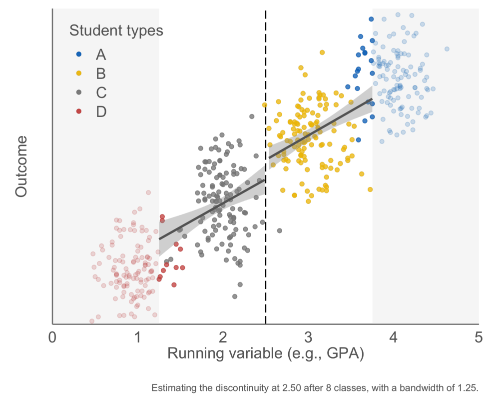
Wager, Stefan and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 2018, *113*, 1228–1242.

Figure 1: The effect of mean convergence on RD estimates in the absence of treatment

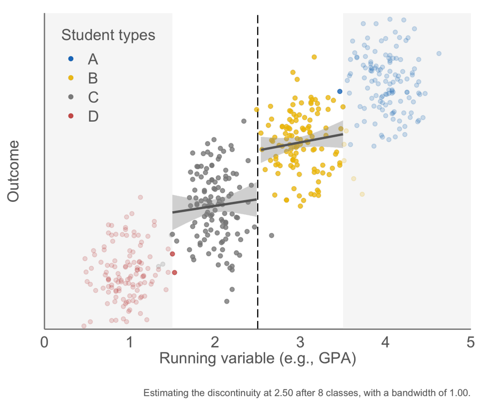
A: After one semester



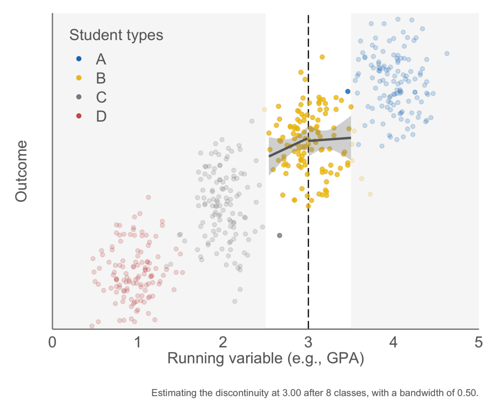
B: After one year



C: After one year with a smaller bandwidth

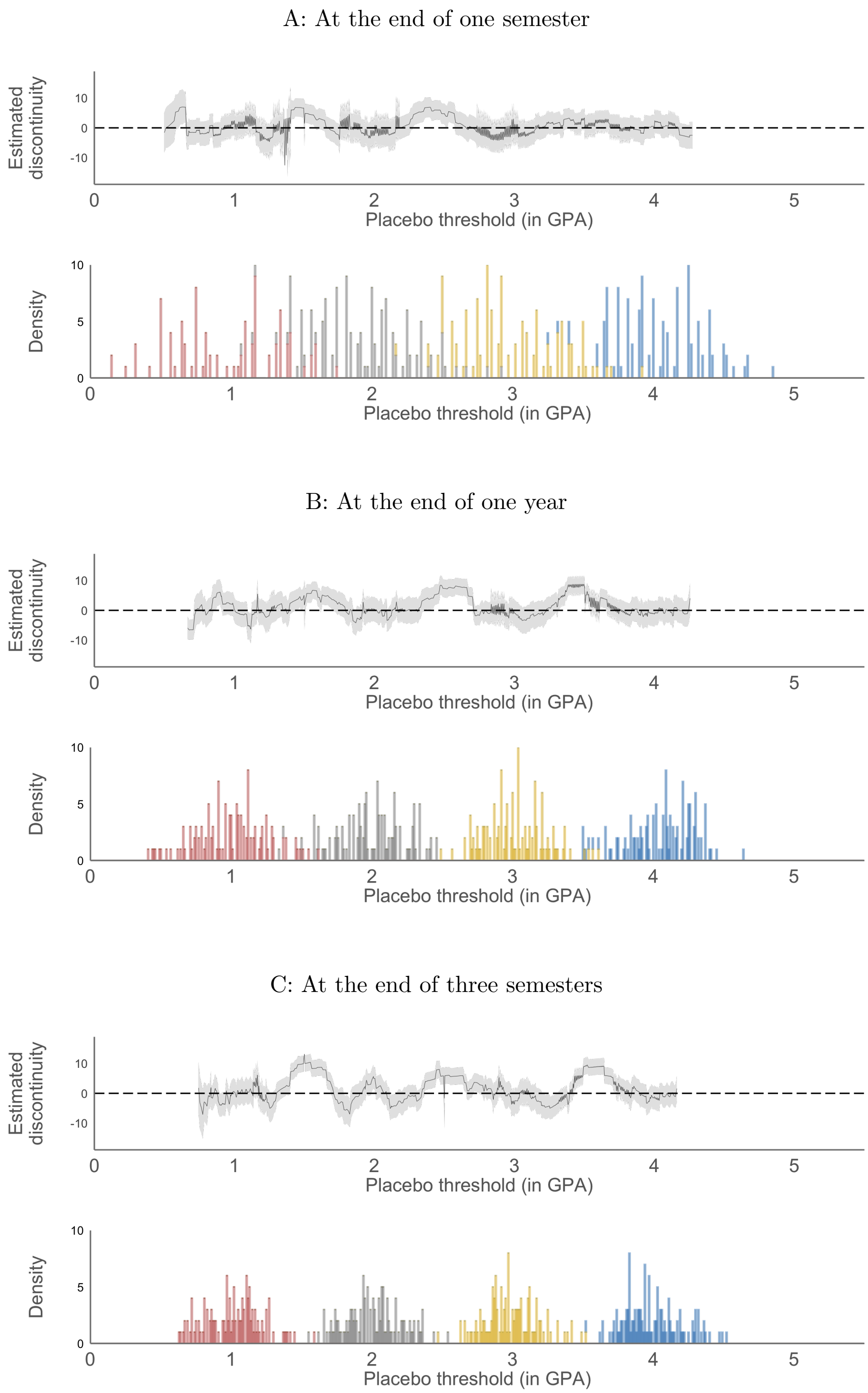


D: After one year with a smaller bandwidth where the threshold "splits" a type of student



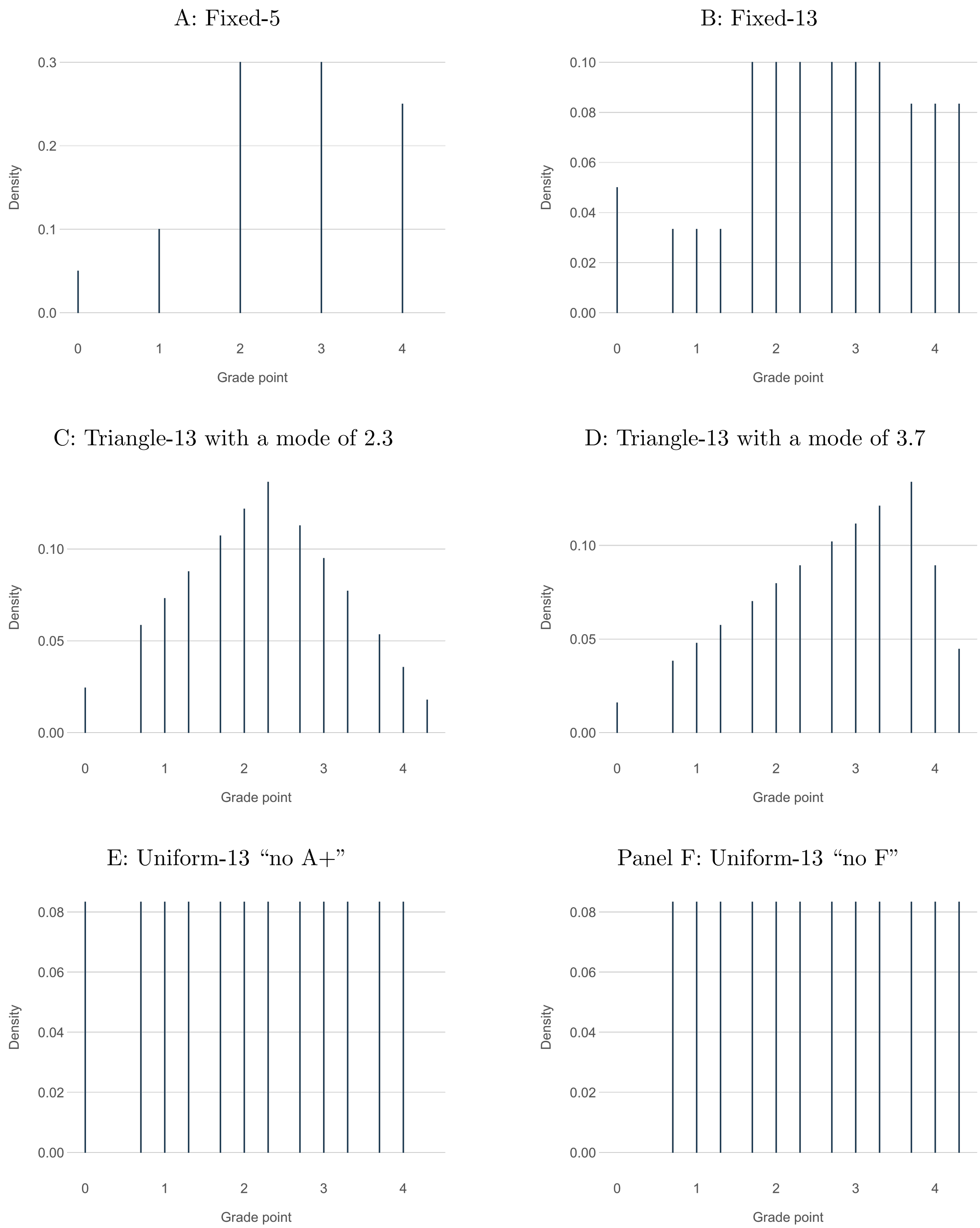
Notes: In all panels, students draw uniformly from the grades that are within ± 1 of their median grade, which is anchored by their type. In all panels there are 125 students of each type (i.e., $n = 125 \times 4 = 500$ students). Students take four classes per semester and two semesters per academic year. See Section 2.1 for related discussion. (For a more flexible environment in which to explore the variation in RD estimates in simulated GPA data, see `https://glenwaddell.shinyapps.io/RD-in-GPA-data/`.)

Figure 2: Placebo tests across GPA in the absence of any treatment (i.e., true $\beta = 0$)



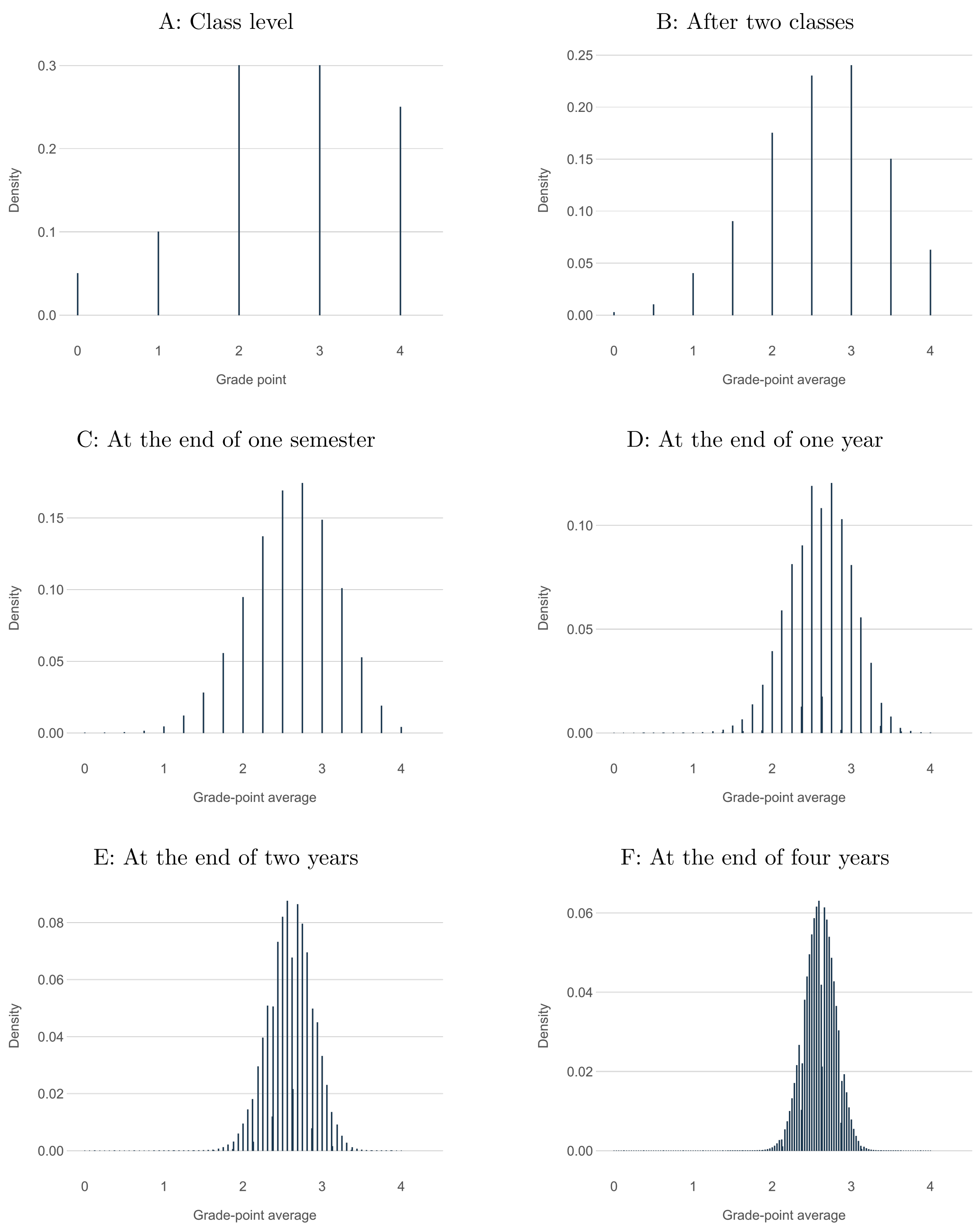
Notes: In all panels, students draw uniformly from the grades that are within ±1 of their median grade, which is anchored by their type. In all panels there are 125 students of each type (i.e., $n = 125 \times 4 = 500$ students). Students take four classes per semester and two semesters per academic year. See Section 2.1 for related discussion.

Figure 3: The combinatorics of GPA: PDFs of different grading curves



Notes: In all panels, the probability densities (at the class level) are identified in the panel titles. See Section 2.2 for related discussion.

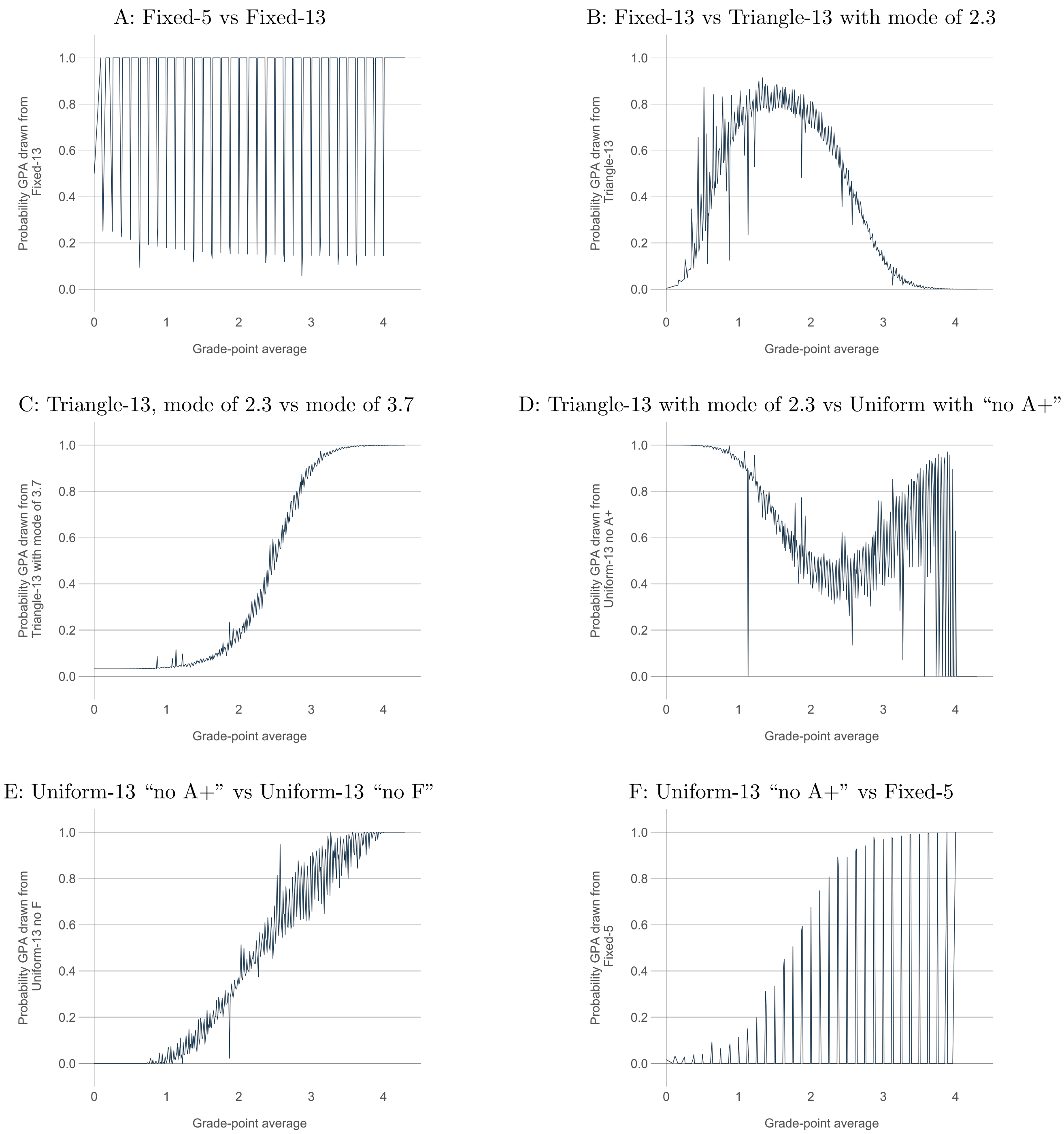
Figure 4: The combinatorics of GPA: PDFs across the number of classes (for a "Fixed-5" curve)



Notes: In Panel A we plot the underlying discrete probability density function at the class level—the student's expectation of grade-point contributions at the class level. In panels B through F we plot the probability densities of having repeatedly drawn from that class-level PDF a number of times, for each GPA between 0.00 and 4.30 in increments of 0.01. In their production, we assume four classes per semester and two semesters per academic year. Thus, across all six panels we span the equivalent of having completed 1 through 32 classes. See Section 2.2 for related discussion.

Figure 5: Non-random sorting into GPA: Pairwise comparisons across students who have experienced two different grading curves for one year of classes

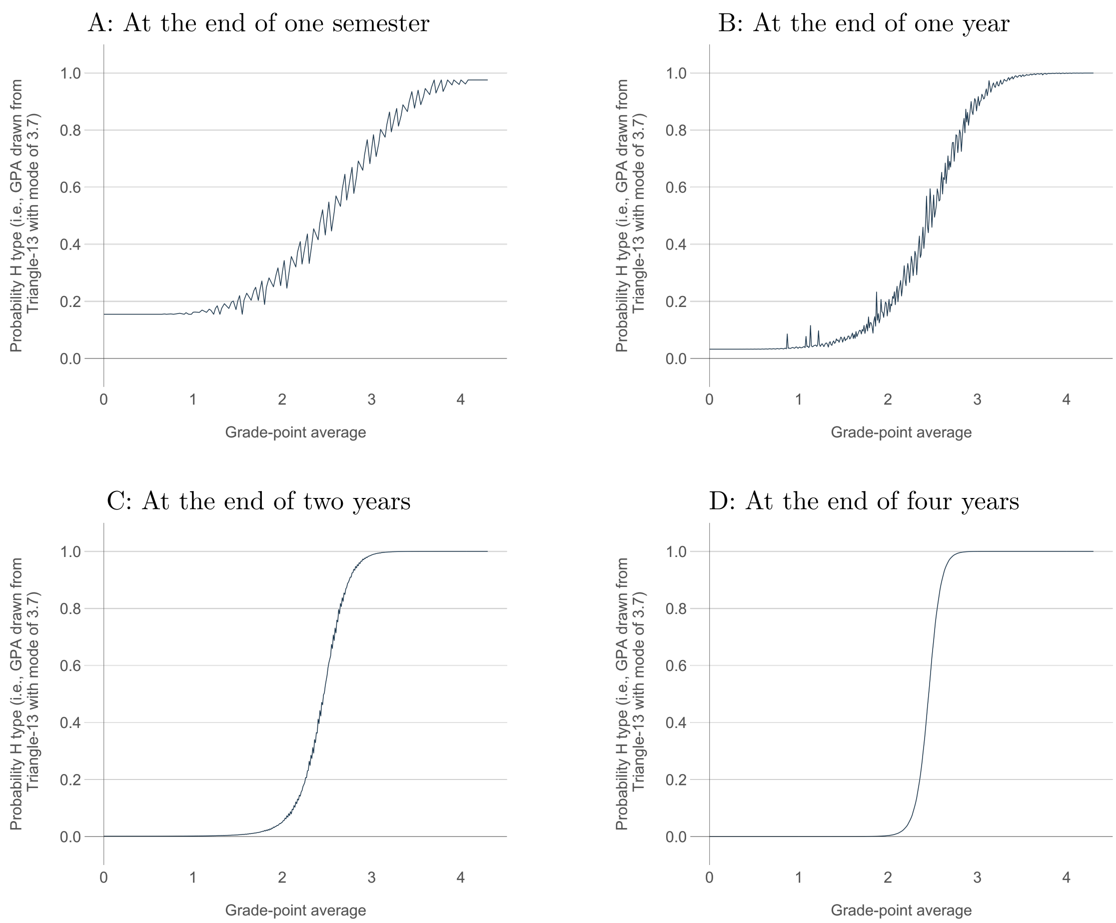
Here we evaluate how students will be distributed across GPAs if there is heterogeneity in grade accumulation across students. In each panel, we evaluate a DGP with two types of student—each draws grades from one of the grade distributions in Figure 3. Across GPAs, we ask how likely it is that a student observed at a GPA has drawn their grades from each of those distributions.



Notes: For each GPA between 0.00 and 4.30 (in increments of 0.01) we plot the probability that a student with that observed GPA was drawing grades from one of the PDFs (identified on the y axis). In each panel, the population of students is split equally between the two types. See Section 2.2 for related discussion.

Figure 6: Non-random sorting into GPA: How a pairwise comparison of H- and L-type students changes over time

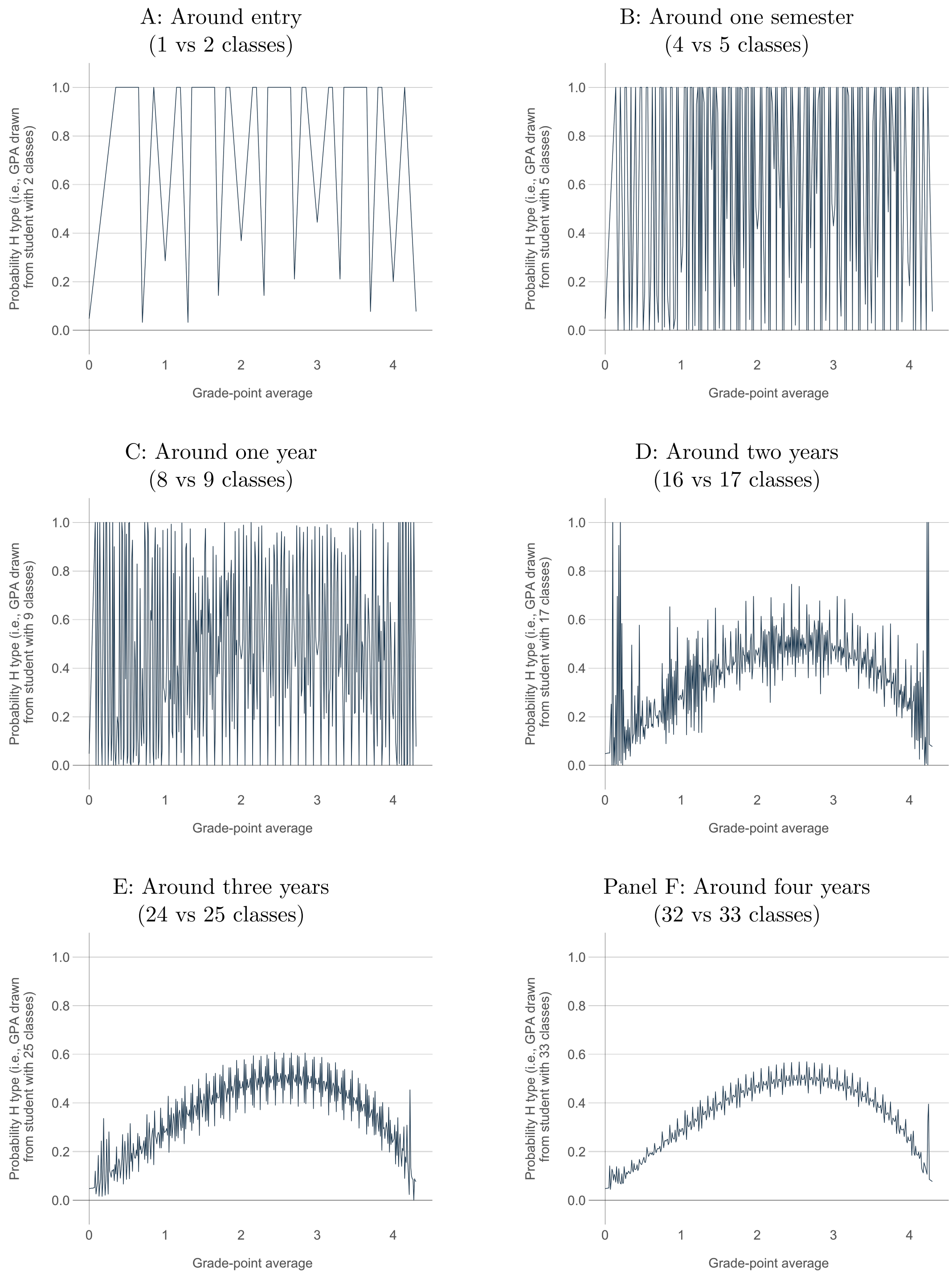
In each panel we evaluate a DGP with two types of student—they draw grades from a "triangle-13" PDF with either a mode of 2.3 (L types) or a mode of 3.7 (H types), as in Panel C of Figure 5. Across panels, we ask how the sorting of students into GPAs systematically changes as students take more classes.



Notes: For each GPA between 0.00 and 4.30 (in increments of 0.01) we plot the probability that a student with that observed GPA had taken classes with the mode of 3.7. Across panels, we reconsider this relationship as the number of classes increases. In each panel, the population of students is split equally between the two types. See Section 2.2 for related discussion.

Figure 7: Non-random sorting into GPA: What if H types just take an extra class?

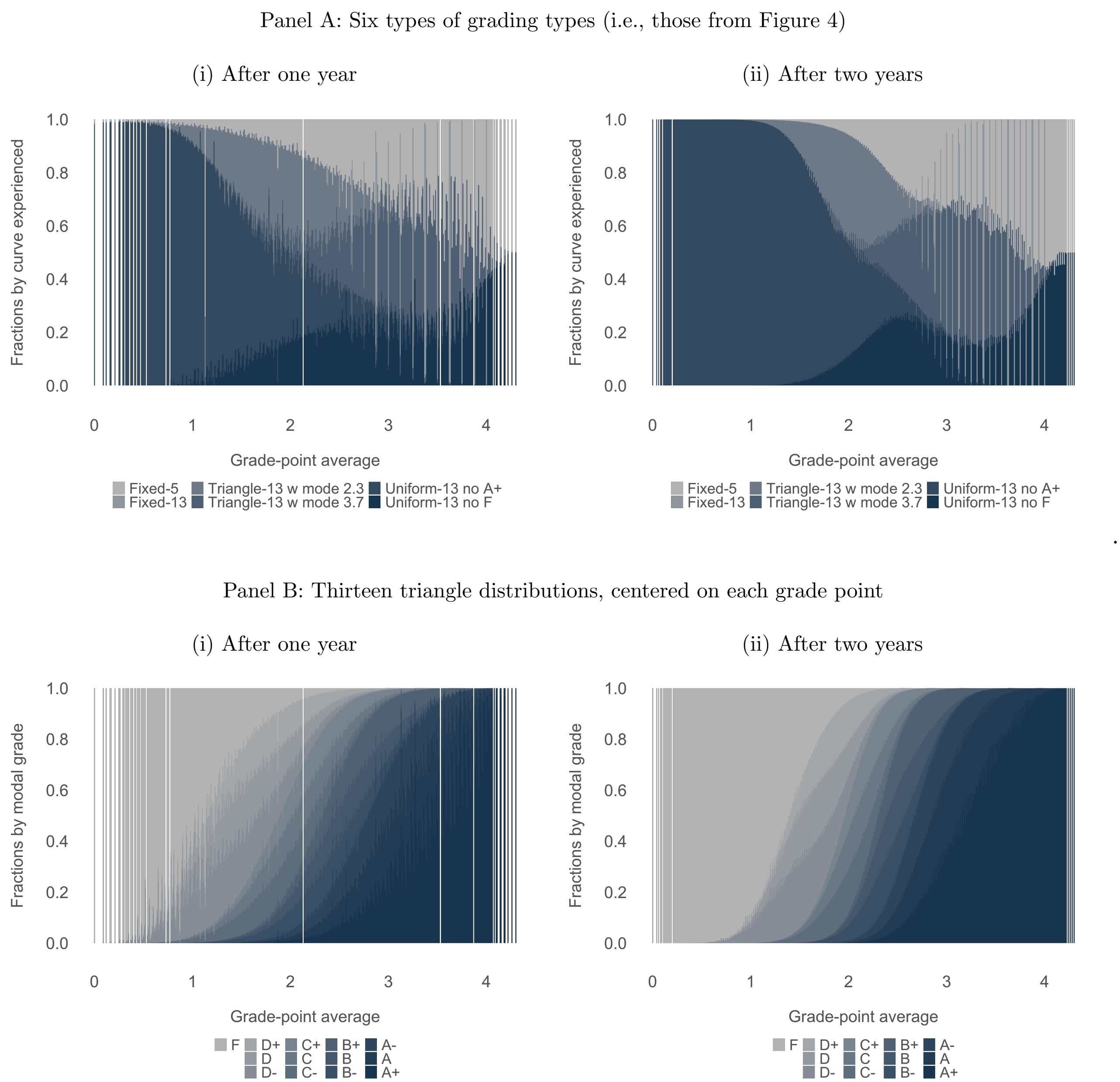
In each panel we evaluate a DGP with two types of student—they both draw grades from a "triangle-13" PDF with a mode of 3.7 (as in Panel D of Figure 3) but L types draw c grades from that PDF and an equal number of H types draw $c + 1$ grades from that PDF. Across panels, we ask how the sorting of students into GPAs systematically changes as students take more classes.



Notes: For each GPA between 0.00 and 4.30 (in increments of 0.01) we plot the probability that a student with that observed GPA had taken $c + 1$ classes. See Section 2.2 for related discussion.

Figure 8: The proportional breakdown of students sorted into GPA by curves experienced

Here, we assume that there are many student types are present in a population (instead of only two), and ask how those students will be distributed across GPAs. In Panel A we visualize the distribution of six types of students, each reflecting one of the grade PDFs shown in Figure 4. In Panel B we assume 13 types of students, each drawing grades from a triangle distribution with a modal grade that corresponds to the 13 traditional letter grades (i.e., F to A+).

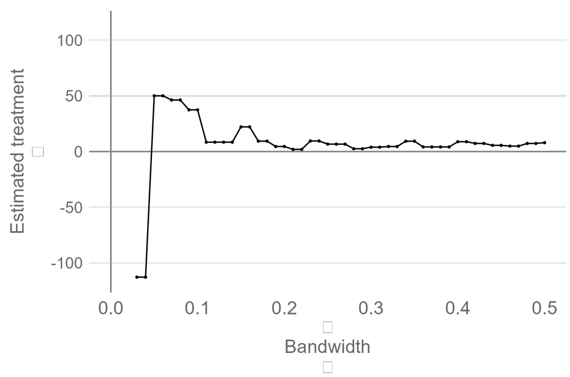


Notes: For each GPA between 0.00 and 4.30 (in increments of 0.01) we plot the stacked probabilities that a student with that observed GPA had experienced the associated grading curves. See Section 2.2 for related discussion.

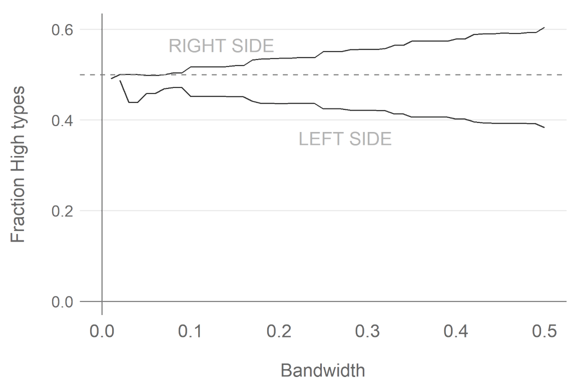
Figure 9: Bandwidth sensitivity in RD estimates evidences combinatorics

In the absence of any treatment, we retrieve estimates of the discontinuity in outcomes at GPAs at or above 2.50. Due to combinatorics, H and L types populate the domain of GPAs differently—this amounts to a violation of the smoothness assumption if this sorting is around the RD threshold. Here, as H types are level different in outcomes, this violation is transmitted through to estimated treatment effects. Estimates at smaller bandwidths are more likely to reflect the non-monotonicity in student-type across GPA. See Section 3 for related discussion.

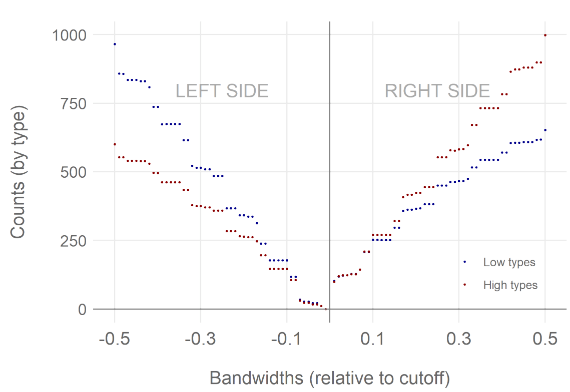
A: Bandwidth sensitivity at the end of one semester (4 classes)



B: On either side of the cutoff, the fraction of "High" types



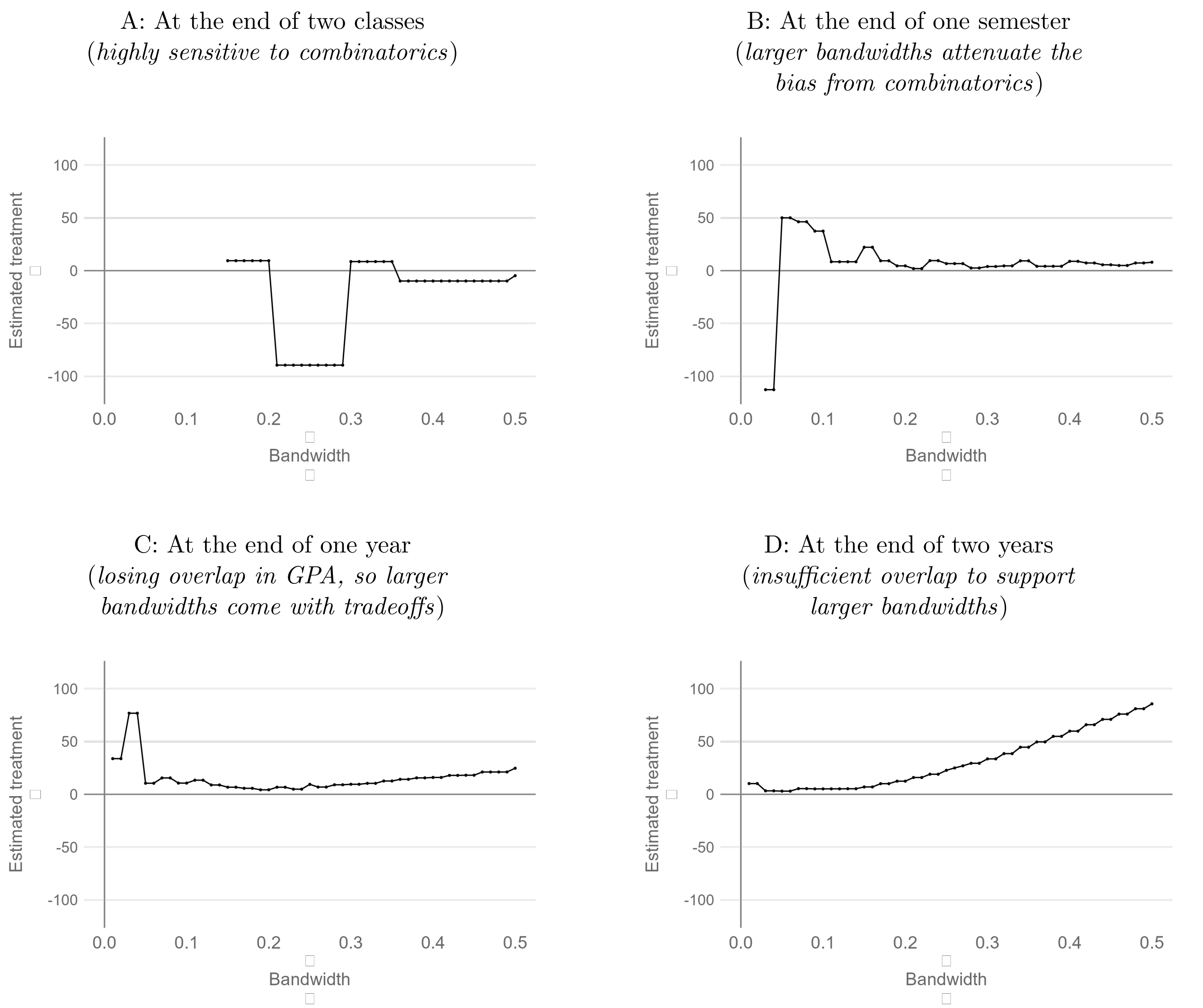
C: The combinatorics-induced sorting of types across bandwidths



Notes: In each panel we consider bandwidths in increments of 0.01 and report means across 1,000 simulations of 30,000 students.

Figure 10: Bandwidth sensitivity in RD estimates at different points in time

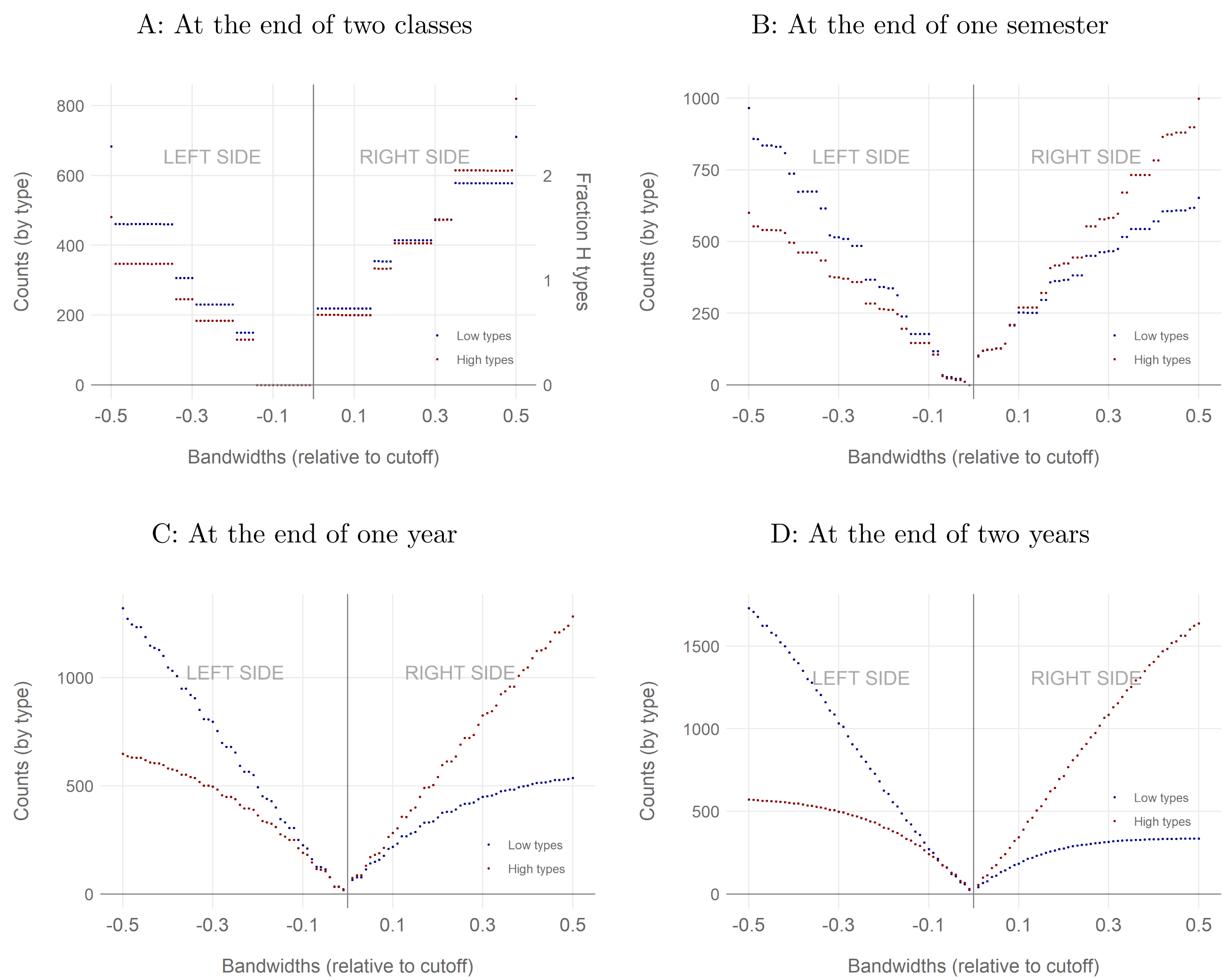
As in Panel A of Figure 9, we retrieve estimates of the discontinuity at GPAs at or above 2.50. (As there is no treatment at 2.50, these should be zero.) This demonstrates that combinatorics is more of a concern at smaller numbers of classes and at smaller bandwidths, with mean convergence becoming more of a concern at larger numbers of classes and at larger bandwidths. See Section 3.2 for related discussion.



Notes: In each we panel consider bandwidths in increments of 0.01 and report means across 1,000 simulations of 30,000 students. In the data-generating process, students take four classes per semester (i.e., eight classes per year).

Figure 11: The combinatorics-induced sorting into bandwidths at different points in time

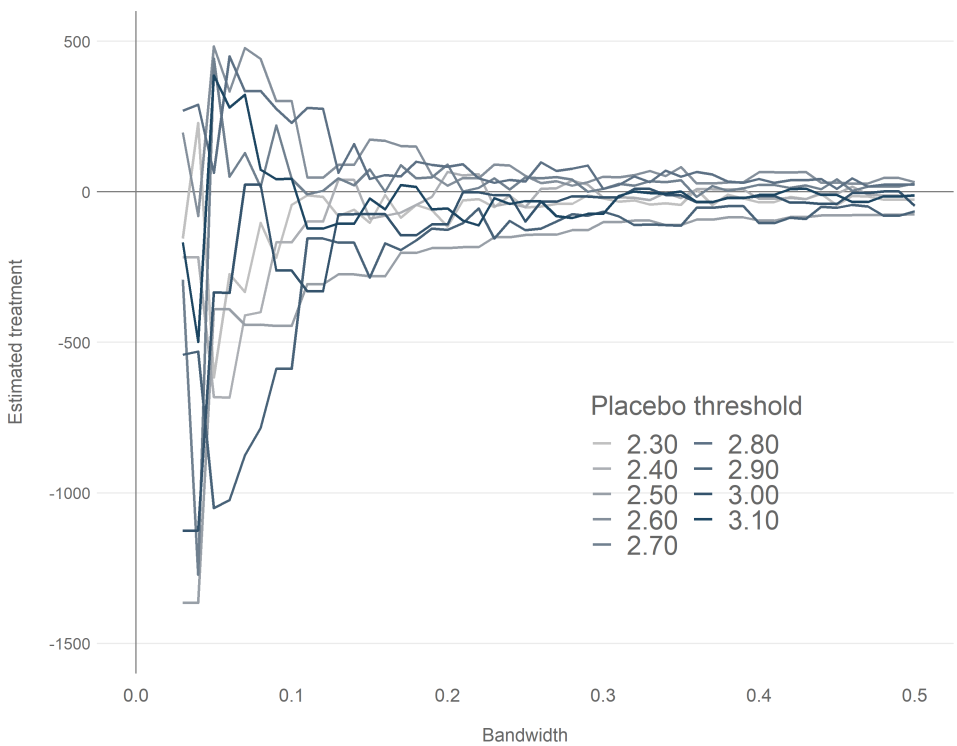
As in Panel C of Figure 9, we visualize how the population of High and Low type students are distributed around a GPA threshold. When students have taken few courses, only some GPAs can be reached, leading the sample included on either side of an RD threshold to change discretely as bandwidths grow. Smoothness in the distribution of High and Low type students only begins to appear after one year of courses have been taken (Panel C), though mean convergence is also begins to appear. See Section 3.2 for related discussion.



Notes: In all cases, the cutoff is a GPA of 2.50 and above. In each panel we consider bandwidths in increments of 0.01 and report means across 1,000 simulations of 30,000 students. In the data-generating process, students take four classes per semester (i.e., eight classes per year).

Figure 12: Bandwidth sensitivities (at various placebo thresholds) when there is student-level variation in the number of classes contributing to GPA

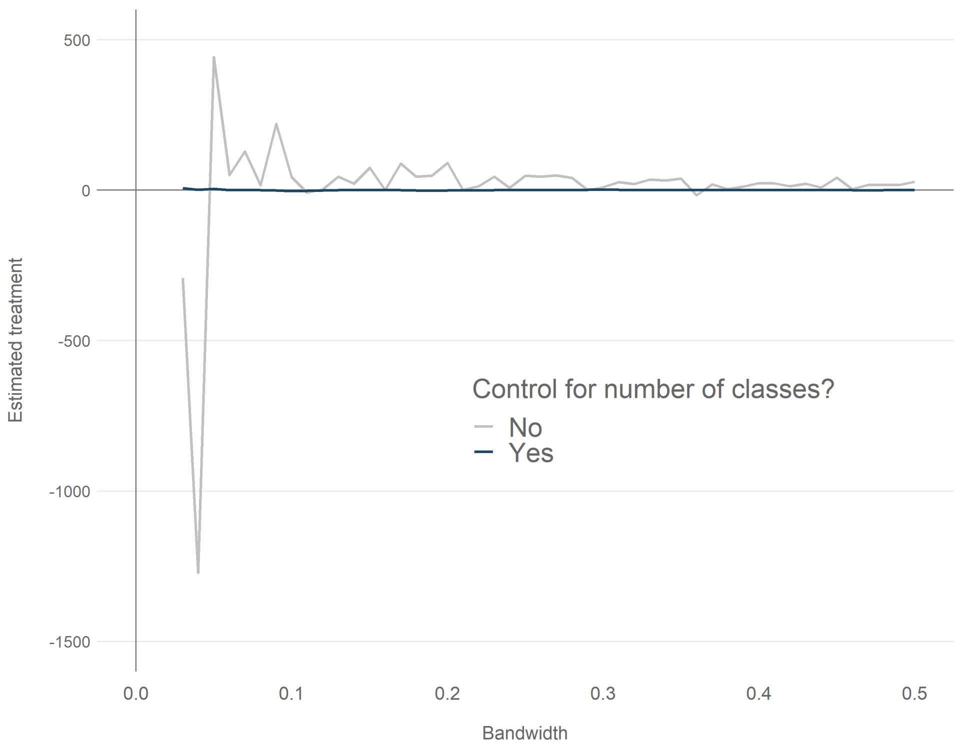
In the absence of any treatment, we retrieve estimates of the discontinuity in outcomes at various GPA thresholds. All students have drawn course grades from the same probability distribution (i.e. a triangle distribution with a modal grade of 2.7), but half of the sample has taken four classes while the other half has taken five classes. In this way we mimic the set of decisions that commonly occur for students after one semester of coursework, such as entry into a specific degree program. We demonstrate that variation across students in the number of classes taken can induce bias through combinatorics. We also evidence here that combinatorics bias is generally unsignable—the sign of point estimates here varies with bandwidth and treatment threshold. See Section 3.3 for related discussion.



Notes: In each panel we consider bandwidths in increments of increments of 0.01 and report means across 200 simulations of 5,000 students.

Figure 13: Bandwidth sensitivity controlling for student-level variation in the number of classes

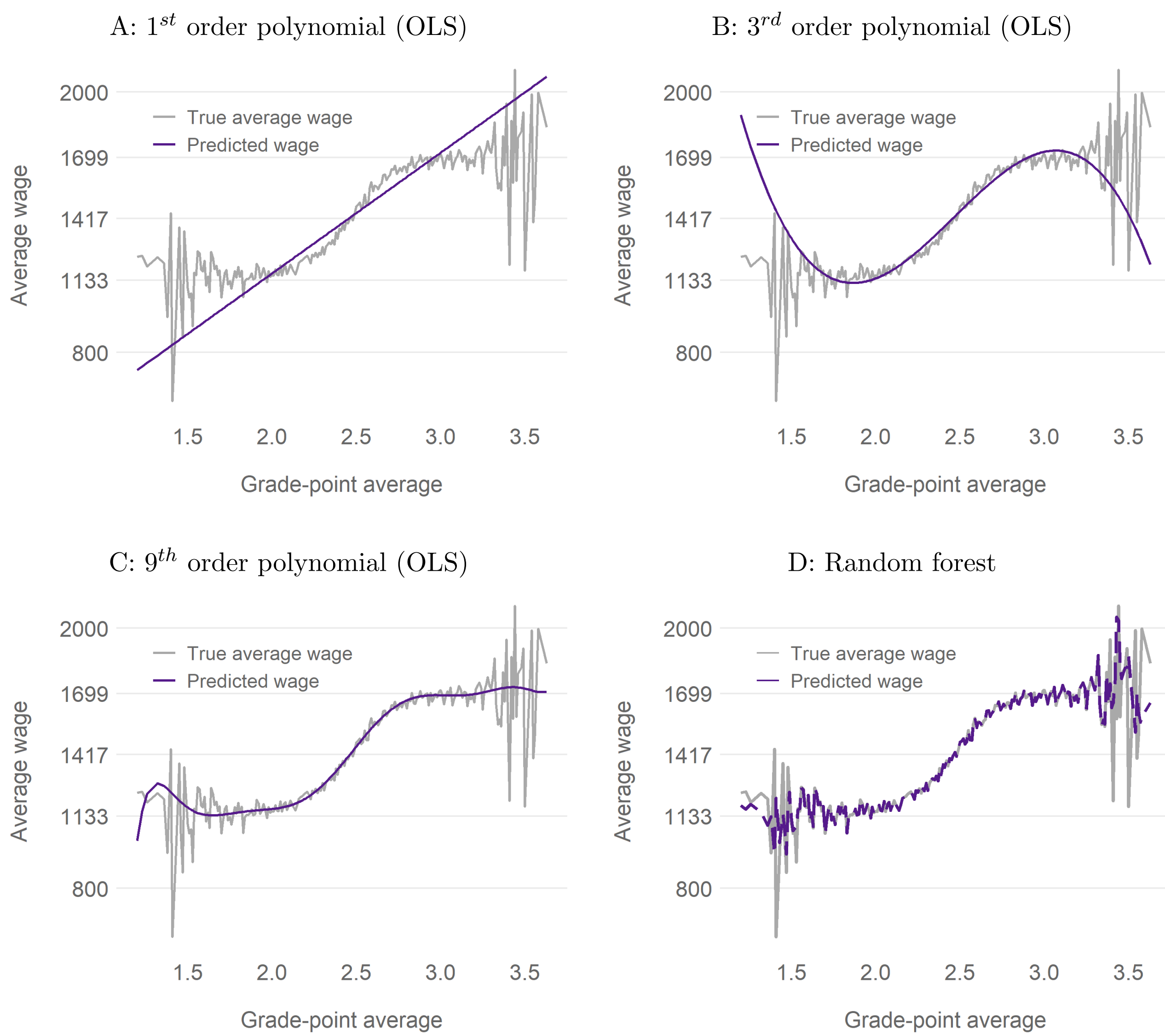
In the absence of any treatment, we evaluate the bandwidth sensitivity of an RD estimator at a 2.70 GPA threshold (using the same data as in Figure 12), with and without controlling for the number of courses taken. All students have drawn course grades from the same probability distribution (i.e. a triangle distribution with a modal grade of 2.7), but half of the sample has taken four classes while the other half has taken five classes. We demonstrate that the bias induced by combinatorics is eliminated when a visible signal of the heterogeneity between students in grade accumulation can be used as a control. See Section 3.3 for related discussion.



Notes: In each panel we consider bandwidths in increments of increments of 0.01 and report means across 200 simulations of 5,000 students.

Figure 14: How well does GPA predict outcomes when there is unobserved student heterogeneity?

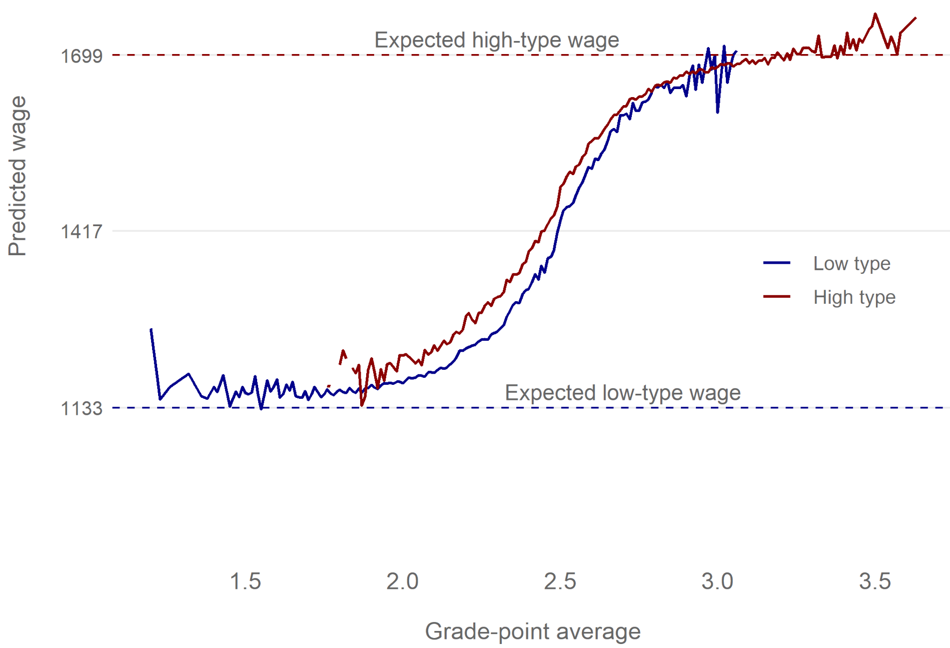
We evaluate the ability of various estimators to predict simulated wages when provided only with GPA. With enough flexibility, OLS estimators can predict the global non-linearity in outcomes, but fail to capture the local nonlinearities. In contrast, the random forest captures local nonlinearities in wages. All panels use the same sample of 30,000 students observed at the end of two years of classes. See Section 4.1 for related discussion.



Notes: L-type students draw from the PDF in Panel C of Figure 3, and H-type students draw from the PDF in Panel D of Figure 3—these PDFs are both triangles with modes of 2.3 and 3.7 respectively. We plot the predicted wage for each GPA in the domain over which we observe at least one student—GPAs between 1.20 and 3.63. Outcomes are level different between the two types and include a randomly drawn error that is normally distributed with a standard deviation of 300. Out-of-bag predictions are used for the random forest results presented in Panel D.

Figure 15: Can machine-learned methods capture student heterogeneity using transcript-level information?

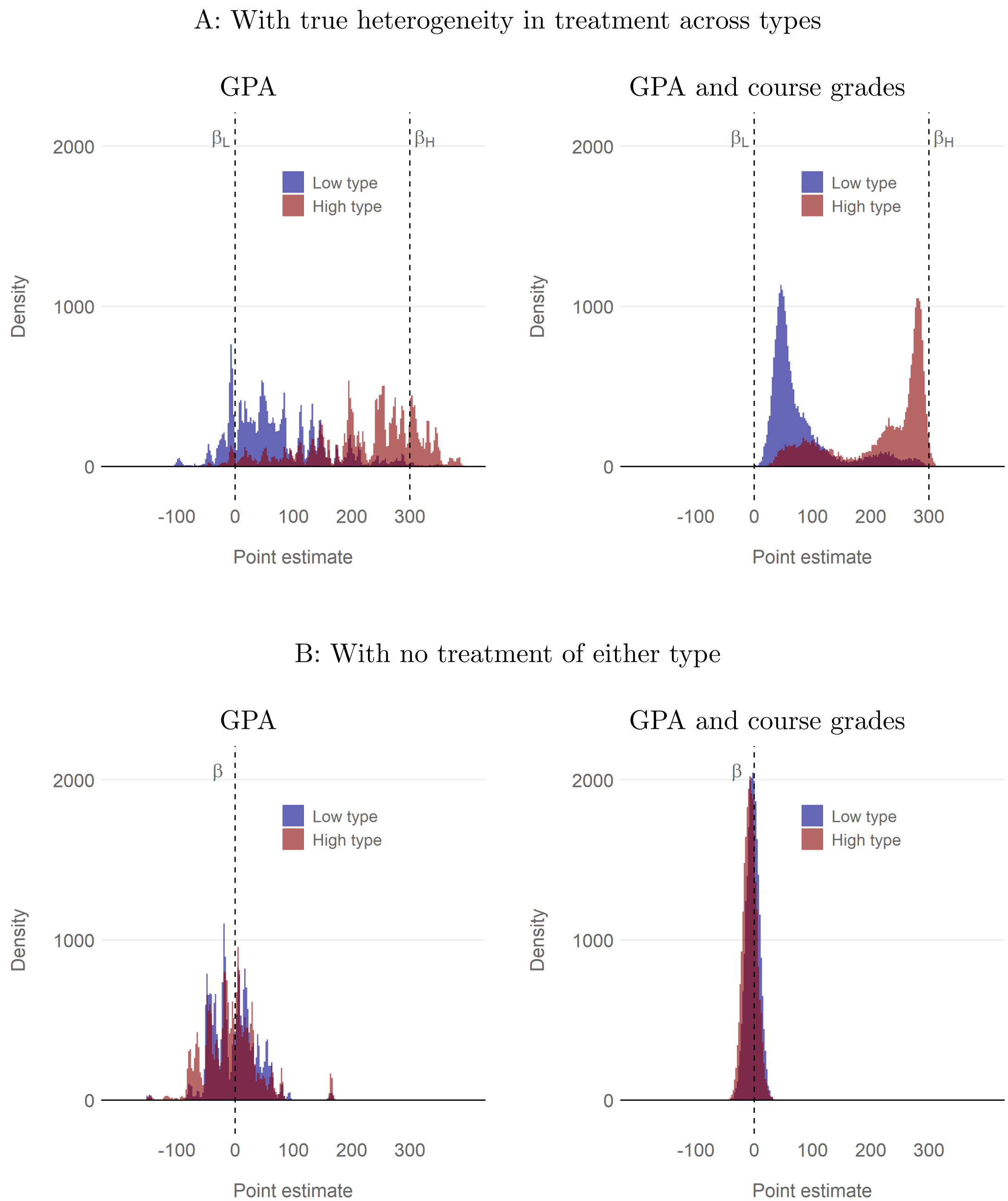
While the random forest on GPA in Figure 14 tracks average outcomes well, the addition of transcript-level data allows a random forest to identify heterogeneity at individual GPAs. This panel reflects a sample of 30,000 students observed at the end of two years of classes.



Notes: We plot the predicted wage for each GPA in the domain over which we observe at least one student—GPAs between 1.20 and 3.63.

Figure 16: How well does a causal forest distinguish heterogeneous treatment across student type?

We estimate conditional average treatment effects (CATE) with and without heterogeneity in treatment effects. The causal forest estimator identifies the presence of treatment effect heterogeneity when present, even when only provided with GPA. Adding individual course grades enhances the causal forest's ability to capture the treatment differences between types (or lack of difference in the absence of treatment). We test for heterogeneous effects in the CF estimators following Chernozhukov et al. (2020) and Athey et al. (2019). More generally, causal forests reject the null hypothesis of no heterogeneity in 100% of simulations.



Notes: In all panels we simulate 30,000 students observed at the end of two years of classes, with a GPA threshold for treatment of 2.50 and above. We plot the density of estimated treatment (in wage). See the GRF manual (link) for more details on causal forest estimation and formal tests for treatment effect heterogeneity. See Section 4.2 for related discussion.

Table 1: The natural variation in GPA exposes RD estimators to over rejecting the $\beta = 0$ null

Treatment estimates are expressed as point estimates relative to standard deviations of the dependent variable. Values in parentheses show the share of simulations in which the estimator (incorrectly) rejects the null hypothesis that $\beta = 0$. The impact of combinatorics on the same RD estimator using a smaller bandwidth is more apparent when students have taken fewer classes. However, as the number of classes taken increases the RD estimator is increasingly vulnerable to the bias induced by mean convergence. The DGP underlying these simulations is shown in Figure 5.

	Estimated effect sizes (fraction of times $p < 0.05$)		
	At the end of one semester	At the end of one year	At the end of two years
	(1)	(2)	(3)
Panel A: Regression discontinuities over-reject the $\beta = 0$ null			
Optimal bandwidth	0.0531	0.0888	0.1146
	(0.31)	(0.56)	(0.61)
Optimal bandwidth × .1	0.1340	0.1069	-0.0039
	(0.13)	(0.08)	(0.06)
Panel B: Causal forests reject the $\beta = 0$ null appropriately			
GPA	-0.0004	-0.0001	0.0005
	(0.04)	(0.05)	(0.05)
GPA and individual course grades	-0.0003	0.0000	0.0005
	(0.04)	(0.05)	(0.05)
Panel C: Causal forests, with bandwidth-restricted samples			
Sample restricted to the RD's optimal bandwidth			
GPA	-0.0001	0.0001	0.0009
	(0.04)	(0.05)	(0.04)
GPA and individual course grades	-0.0001	0.0002	0.0009
	(0.04)	(0.05)	(0.04)
Sample restricted to the RD's optimal bandwidth × .1			
GPA	-0.0015	0.0004	0.0011
	(0.05)	(0.04)	(0.06)
GPA and individual course grades	-0.0010	0.0005	0.0015
	(0.05)	(0.04)	(0.06)

Notes: Estimates are means across 1,000 samples of 30,000 students using the same DGP as in Panel A of Figure 16. Optimal bandwidth selection for the fuzzy RD estimator follows that of Imbens and Kalyanaraman (2012). If the smaller bandwidth RD estimator would lead to empty bandwidth on either side of the treatment threshold, we default to the smallest populated bandwidth. We test for heterogeneous effects in the CF estimators following Athey et al. (2019), Chernozhukov et al. (2020).

Appendices

A Performance of OLS and ML wage prediction models

In Table A1 we report quantitative measures of the performance of the wage prediction models shown in Figure 14. The data generating process through which students accumulate grades and realize post-graduation wages is unchanged from that described in Section 4.1.[31] We test the performance of these methods across 1,000 simulations when evaluating a population of 30,000 students observed at the end of 1 semester (Panel A), one year (Panel B), and 2 years (Panel C). Here we highlight the results from Panel A, where the population of students in question has taken four classes, but we note that the results and relative performance of our prediction models are similar when students have taken more classes. In Column (1) we report the frequency at which a 0.01 increase GPA is associated with a decrease in the weekly wage of the average student. At the end of one semester, a 0.01 increase in GPA predicts a *decrease* in weekly wages 43.7 percent of the time, across simulations.[32] While the fitted polynomials rarely predict such a decrease, the random forest predicts that wages rise as GPA rises 43.5 percent of the time, close to the true share in the data. Moreover, as we report in Column (2), roughly 98 percent of these predictions are accurate. As a complement to Column (2), in Column (3) we report the fraction of GPAs at which average wages decrease that are missed by each of our models. Again, the random forest performs especially well, and fails to "catch" a GPA at which average wages fall only 2 percent of the time.

In columns (1) through (3) we demonstrate that the random forest predictor is effective at capturing local non-monotonicities in average wages. But the question of how closely each estimator tracks average wages is also important. To address this question, in Column (4) we report the mean distance between the predicted and true average outcome at each GPA.[33] We express the distances in Column (4) relative to that of the linear model. For example, when evaluating students at the end of one year of classes, the random forest predicts wages that are over ten-times closer to the true average wage than the predictions of the linear model.

[31] Due to the sparsity of students in the tails of GPA, we report these measures of model performance for the inner-95 percent of observations. In the tails, this sparsity leads to erratic changes in the average wage across GPAs, which all of our models have difficulty in tracking.

[32] Absent any statistical noise, combinatorics alone would lead a 0.01 increase in GPA to predict a decrease in weekly wages 41.8 percent of the time when this population of students was observed after one semester of classes.

[33] We compute distance as the absolute difference between the predicted wage and the average actual wage at a GPA. This statistic captures how well the prediction lines in Figure 14 track the true average wage line, but in a way that can be easily averaged across multiple simulations. In plain language, Column (4) captures how far apart the purple and grey lines are in Figure 14.

Table A1: Across methods, how well does GPA predict (simulated) weekly income variation?

	How often does wage decrease? [a]	How often is the predicted decrease correct?	How often does it not catch the decrease?	Distance between the predicted and actual wage at a GPA, relative to 1^{st} order polynomial [b]
	(1)	(2)	(3)	(4)
Panel A: At the end of one semester (4 classes)				
DGP	43.5%	–	–	–
OLS				
1^{st} order polynomial	0.0%	-	100%	1.000
3^{rd} order polynomial	0.1	53.1	99.9	0.911
9^{th} order polynomial	0.3	57.5	99.7	0.898
Random forest	43.5	98.1%	2.0	0.082
Panel B: At the end of one year (8 classes)				
DGP	44.7%	–	–	–
OLS				
1^{st} order polynomial	0.0%	-	100%	1.000
3^{rd} order polynomial	0.1	46.4	99.9	0.864
9^{th} order polynomial	0.3	46.8	99.7	0.836
Random forest	44.6	97.6%	2.6	0.119
Panel C: At the end of two years (16 classes)				
DGP	44.8%	–	–	–
OLS				
1^{st} order polynomial	0.0%	-	100%	1.000
3^{rd} order polynomial	5.5	50.3%	93.9	0.629
9^{th} order polynomial	3.8	50.2	95.8	0.469
Random forest	44.8	98.5	1.5	0.021

Notes: Based on 1,000 simulations of 30,000 students. We calculate all values using only the students with GPAs in the central 95% of the data. At 8 classes, the average minimum GPA in the central 95% of the data is 1.64and the maximum is 3.29, while at 16 classes the minimum GPA is 1.82 and the maximum is 3.11. Actual wage decreases are defined as instances in the simulated data wherein the average wage at a GPA falls compared to the average wage at the next lower and at which we observe students.
[a] In the actual wage data generated by the DGP, and then in the wages predicted at each GPA by each of the methods.
[b] In plain language, Column (4) captures the average distance (across 1,000 simulations) between the purple and grey lines shown in Figure 14, normalized by the value of the first-order polynomial.