Waddell, Glen R.; Putz, Jenni

Working Paper

What Can We Learn from Student Performance Measures? Identifying Treatment in the Presence of Curves and Letter Grades

IZA Discussion Papers, No. 15321

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Waddell, Glen R.; Putz, Jenni (2022) : What Can We Learn from Student Performance Measures? Identifying Treatment in the Presence of Curves and Letter Grades, IZA Discussion Papers, No. 15321, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/263537

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15321

What Can We Learn from Student Performance Measures? Identifying Treatment in the Presence of Curves and Letter Grades

Glen R. Waddell
Jenni Putz

MAY 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15321

What Can We Learn from Student Performance Measures? Identifying Treatment in the Presence of Curves and Letter Grades

Glen R. Waddell
University of Oregon and IZA

Jenni Putz
University of Oregon

MAY 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

What Can We Learn from Student Performance Measures? Identifying Treatment in the Presence of Curves and Letter Grades*

Grade-based performance measures are often relied on when considering the efficacy of education-related policy interventions. Yet, it is common for measures of student performance to be subjected to curves and discretized through letter-grade transformations. We show how transformed grades systematically challenge causal identification. Even without explicit curving, transformations to letter grade are particularly problematic and yield treatment estimates that are weighted combinations of inflated responsiveness around letter thresholds and "zeros" away from these thresholds. Curving practices can also introduce false patterns of treatment heterogeneity, attenuating measured responses to treatment among high-performing students, for example, or inflating measured responses among low-performing students.

JEL Classification: I21, I26, C21

Keywords: program evaluation, grades, curves, gpa, education

Corresponding author:
Glen R. Waddell
Department of Economics
University of Oregon
Eugene, OR 97403-1285
USA

E-mail: waddell@uoregon.edu

* We thank Joshua Angrist, Peter Arcidiacono, Timothy Bond, Joshua Goodman, Jason Lindo, Grant McDermott, Tyler Ransom, Edward Rubin, and Isaac Swensen for helpful comments.

1 Introduction

Engaging with letter-grade measures of student performance is common for both practitioners or researchers. If discrete letter classifications are not attached to individual assignments or tests, they are very often the summary record that students receive at the end of a class. It's also typical for measures of student performance to be transformed by a "curve" of some kind as they're being discretized. Thus, while one often looks for systematic variation in student grades to evidence the efficacy of policy interventions, it is rare to observe *untransformed* measures of student performance—and it is in these pre-curved, untransformed performance of performance that the benefits of interventions would be best evidenced. However, knowing how treatment-induced changes in the latent performance of students identifies the average treatment effect is important, as is the interpretation of treatment estimates in this context.

Taking the practice itself as given, we provide a taxonomy of sorts—how we should consider our ability to asses the performance returns to policy intervention within the various mappings of student performance to letter-grade records of performance. In the end, we demonstrate how letter-grade transformations *systematically* distort our perceptions of treatment efficacy, and fundamentally jeopardized the ability of researchers to estimate unbiased treatment effects.

Letter-transformed grades present unique empirical challenges that are largely unexplored in the literature. Letter grading does share some of the challenges that are more-typically associated with ordinal measures of happiness and subjective notions of well-being (Bond and Lang, 2013; Schrödera and Yitzhakib, 2017). However, there are unique aspects to grading practices that go beyond those of ordinal-measures data, and with important implications. For example, unlike happiness scales (where one's happiness need not displace another's) letter-grade assignments often impose zero-sum conditions. In short, if an untreated student *must* be displaced in letter grade for a treated student to be able to improve in letter grade, "treated minus control" differences in outcomes double count the effect of treatment.[1] More generally, though, *any* displacement brought about by the allocation of scarce letter grades amounts to a violation of the stable unit treatment value assumption (SUTVA),

[1] Note that expressing grading norms in terms of fractions of classes that can receive various letter grades directly implies zero-sum competition. Princeton's university-wide grade deflation policy did this, for example, by recommending that each department should "award no more than 35% of A-range grades for course work," and "no more than 55% of A-range grades" for independent research typical of juniors and seniors. (Source: https://odoc.princeton.edu/faculty-staff/grading-princeton.) Likewise, imposing maximums on the average grade imposes similar tradeoffs (e.g., Wellesley's anti-grade-inflation policy, as described in Butcher, McEwan, and Weerapana, 2014).

and in this context manifests in inflated treatment effects.

Even in the absence of any displacement around letter grades, however, letter grades should not regularly retrieve unbiased estimates of treatment as there is another equally fundamental challenge to identification associated with the discretized measurement of performance. Namely, to the extent that treated students are far from letter-grade distinctions—far enough that they don't increase in letter grade despite higher levels of performance—treatment estimates must attenuate. In essence, if the only thing we attribute to treatment is changes or differences in letter grade, then treated students who do not perform "better enough" for their improvement to be observable in letter grades will be accounted for as unresponsive to treatment.

It is in these two ways that we characterize letter-based grades as problematic. Ultimately, when we look to letter grades for evidence of treatment we are retrieving estimates that are weighted combinations of (i) inflated responsiveness around letter-grade thresholds (due to the potential displacement of untreated students) and (ii) "zeros" for those away from thresholds (due to the latent performance gains).

Given that these are in tension, across varieties of typical curving practices we demonstrate the systematic ways in which the resulting net biases can materialize. In particular, we demonstrate the implications of four different approaches to curving, each having their own shape parameters that change the degree to which they bind. Some instructors will adopt these curves explicitly—an internet search for "how to curve grades" will quickly identify these and guide instructors through applying them. However, these are merely representative curves meant more as abstractions that cover the sorts of *shapes* implied by the broader set of potential curving practices. That said, two of these curves are used in popular learning management systems. For example, Canvas defaults into one of these explicitly, and Blackboard uses another of these mechanisms as their example when they walk instructors through how they can curve their grades.[2] In the end, we've captured a fairly wide range of practice without needlessly exhausting readers—in practice one of these is quite likely to capture

[2] The curve option in Canvas is a "two-point transformation," where the instructor is given the opportunity to chose a new mean score for a given test or assignment. In Blackboard, the instructions imply that instructors might want to adopt a mechanism that adjusts scores so that the highest-scoring student receives a score of 100 (i.e., a "high grade to 100" rule).

the shape of the various curves we employ, and it is exactly the "shape" that drives the result.[3]

In each curving environment we simulate the random treatment of students and consider our ability to retrieve an unbiased estimate of the effect of that treatment. In the end, the ability to identify treatment effects will depend on the curve that was adopted, which should be troubling to applied researchers given the lack of transparency around the application of curves. Moreover, even in cases where the curve is known and the bias is signable (which is not always the case), curving mechanisms can introduce false patterns of treatment heterogeneity. This itself has significant implications for policy, which we discuss in our closing remarks.[4] The main point remains, however—treatment estimates are sensitive to letter-grade and curving practices, and we rarely know which curve has been applied (or with what tuning parameters) so our confidence should be conditioned when relying on letter-grade data in the evaluation of policy interventions.

As we proceed, we will remain agnostic as to the interventions the researcher is attempting to identify, though one could imagine treatments or policy experiments such as providing academic-support services or financial incentives to students to improve academic performance (Levitt, List, and Sadoff, 2016; Barrow, Richburg-Hayes, Rouse, and Brock, 2014; Angrist, Oreopoulos, and Williams, 2014; Angrist, Lang, and Oreopoulos, 2009), or academic and behavioral interventions for disadvantaged youth (Cuellar and Dave, 2016; Cook et al., 2014). Likewise, one could have in mind the manipulation of peer characteristics in the estimation of peer effects or social spillovers, as in Angrist (2014).[5] Of the many interventions one could have in mind—the interventions are not what is important here—our main concern is that researchers routinely look to transformed grades for evidence of treatment.

In Section 2 we discuss the implications associated with letter-grade transformations. The most-significant takeaways likely exist here. In Section 3 we define four representative curving mechanisms.

[3] It turns out that in our own curving practices, our less-formal approaches indirectly result with something quite like the "root curve" we will describe, being more generous to the lower-performing students. Of course, the multitude of informal curving methods will introduce its own source of variation into the heterogeneity estimated treatment effects. (For example, though we will not consider their implications here, giving extra help to low-performing students or offering to re-grade assignments could arguably be considered "curves.")

[4] There are several margins around which one might imagine "fixes" existing. We've collected some of our intuition in Appendix A. In most cases, the intuition is in seeing how the proposed fixes are merely approximations of the fundamental mechanisms we've reported above. (For example, "percentile ranks" can be thought of as letter-grade transformations with many thresholds, and thus does not offer a fix at all.)

[5] Angrist (2014) describes clean inference being achievable in experiments where peers "are a mechanism for causal effects but not themselves subjects for study." For example, one can imagine subjects having been randomly allocated peers but the researcher's focus remaining on the *original subjects who are treated to those new peers.* In this way, "treatment" in our simulated environments can also include otherwise-well-identified models of social interaction. As such, we will suggest that even in these refined environments where one can learn of social interactions, curve transformations threaten identification. (In other attempts to identify peer effects, the appropriate inference is often unclear even in the absence of curve transformations. Adding curves to those attempts to identify peer effects complicates inference even further.)

These will act as examples of the transformations we anticipate being nested within any letter-grade regime, and further complicate the inference problems associated with letter grades. Here, we also describe what is known about the practice of grade curving, though the literature is surprisingly light. In Section 3 we demonstrate the implications of curving on treatment-effect estimates across each curving rule in simulated environments that assure random treatment assignment. To be clear, our intention is not to have letter-grade transformations and the wide variety of "curving" mechanisms be thought of as substitutes. By discussing them separately, we merely hope to address the challenges to identifying treatment that are driven by the discretization of performance into letters, and then the particulars that depend on the curving mechanism adopted within that regime (i.e., on the particular shape of the transformation of raw performance into letters). In Section 4 we offer some related thoughts and important implications for policy moving forward, concluding in Section 5.

Before we continue with our focus on the researcher's interest, it is likely right to also acknowledge that students use grades to inform themselves of relative aptitudes and behave systematically with this performance feedback (Arcidiacono, Aucejo, Maurel, and Ransom, 2016). As such, even as we discuss the implications from the researcher's or policy maker's point of view, we can imagine that students, in their own attempts to identify the effect of their own efforts on grade outcomes, could also find their ability to interpret their own standing hampered by the effects of grade transformations.[6,7]

[6] For example, any student who has ever legitimately shared the sentiment of having "tried so hard this term" belies an experimentation with the degree to which their efforts have been rewarded with higher (curved) grades. We believe that high grading standards matter to student outcomes at both the elementary and secondary levels (Figlio and Lucas, 2004; Betts and Grogger, 2003)—to the extent curves obstruct the ability of students to update their own priors, we might also worry about the implications from that perspective.

[7] Moreover, our concern that transformed grades do not reliably inform us about student-level treatments should not be interpreted as implicitly sanctioning other group-level interactions. For example, the fundamental non-comparability we describe here is of the sort that can also give rise to non-comparability at higher levels of aggregation. Along these lines, Bond and Lang (2013) identifies the role of scaling in ordinal measures of performance, suggesting caution when using test scores to determine when black-white test-score gaps first emerge, and whether or how they widen throughout early school years. Bond and Lang (2018) shows that measurement error in test scores underestimate the black-white test-score gap (and adjust the gap by using an instrumental-variables approach). Lang (2010) also addresses a related concern in the use of value-added measures to determine teacher compensation and retention, lamenting that "economic studies of education commonly proceed as if the intervals between scores always mean the same thing, as if top- and bottom-coding did not exist, and as if fourth- and fifth-grade test scores are really comparable." Likewise, Nielsen (2019) finds dramatic decreases in the achievement gap between youth from high- and low-income households using scale-independent tests that maintain cardinal and inter-group comparability.

2 Letter-grade transformations

2.1 The identification problem

In the end, transformations to letter grade will yield treatment estimates that are weighted combinations of "zeros" for those away letter-grade thresholds and inflated responsiveness for those close to thresholds. Moreover, these weights will be endogenous to curving practices, which leaves treatment estimates both sensitive and manipulable. To see the origins of these competing weights, consider that treatment of size β (in raw scores) can only ever be evidenced in letter grades if the treated individual is already within distance β of a letter-grade cutoff in the absence of treatment. By implication, when derived from letter grades, treatment estimates do not reflect the treatment of *all* treated individuals in a class. (This is most clear for the highest-performing students, who may well perform better with treatment but simply cannot *measurably improve* with treatment. However, the same is true of any treated student who is more than β below a letter-grade distinction.) In letter-grade spaces, then, we should always reflect on treatment estimates being contributed to differently by those who are more than β away from letter-grade distinctions and those who are *within* β of a letter-grade distinction.[8]

The first of these two mechanisms is more obvious—if the empirical design retains treated students who are more than β from a letter-grade distinction, they will only contribute weight to zero in the estimator, as though they were unresponsive to treatment. Thus, average treatment effects will be attenuated in designs that are unable to distinguish students by how far they are from letter-grade distinctions in the absence of treatment. (Below we will discuss related implications further.)

At the same time, a more-complex interaction contributes to identifying treatment "within βs" of letter-grade distinctions, and ultimately leads to a violation of the stable unit treatment value assumption (Rubin, 1980, 1986). Namely, this is due to the zero-sum competition among students that accompanies all relative grading system, and materializes precisely around letter-grade thresholds. For example, consider a grading rubric in which the top-30 percent of a class receive As and the next-highest 30 percent receive Bs, etc.[9] For a given class size, then, the number of As is fixed, and for

[8] In Section 3 we will consider that, before the transformation to letter grades, raw scores x_i can also be acted on by a curve $g(\cdot)$. In the presence of curved scores, then, these distances would be characterized in some $g(\beta)$-denominated distances around cutoffs, with the distance varying with the curve applied and the densities of students there. Given any non-linearity in $g(\cdot)$, for example, they need not be symmetric around a given threshold or common across thresholds.

[9] This roughly approximates the established norms in our own upper-division Economics classes, but also reflects the way that grade deflation policies often communicate grading norms (e.g., recall Princeton's "no more than 35% of A-range grades for course work" rule).

any treated student "within a treatment effect" of an A there must therefore be an untreated student that is at the margin of being *displaced* by treatment—a student who would have received an A in the absence of treatment but receives a B when treatment affords other students an advantage.[10] Any zero-sum environment faces this challenge. Furthermore, in a strictly enforced zero-sum rubric, for treatment to have changed the letter grade of a treated student, it must have also changed the letter grade of an untreated student. In that way, treatment is inseparable from *untreated* students coincidently suffering an equal but opposite change in grade, which implies that this upward bias will be by a factor of two.[11] As the degree of displacement is often an instructor's choice (and unobservable to the researcher), we conclude more generally as follows: For the students who are within β of letter-grade distinctions, their contributions to treatment will be biased upward when there are any limits imposed on the number of students who are to receive letter grades, as it is these limits that force the displacement of untreated students and thereby inflate contributions to treatment estimates. In short, any amount of displacement violates the "all-else-equal" condition that is necessary to make a causal claim.

We take a first pass at illustrating these biases in Figure 1 by visualizing a zero-sum scheme in which the top-30 percent of students receive As, the next-highest 30 percent receive Bs, and the bottom-40 percent receive Cs. In Panel A we see the distribution of performance for treated students, which shifts to the right with treatment from $x_i \sim N(70, 10)$ to $x_i \sim N(72, 10)$. In Panel B the performance of students in the control group is $N(70, 10)$ before and after treatment.[12] In both, we identify the cutoff that separates Cs from Bs and the cutoff that separates Bs from As. Due to the zero-sum nature of grades, the performance required in order to be awarded a B or an A increases after treatment—this is true for students in both the teated or and untreated group, and is proportional to how many treated students now compete better for the higher letter grades. (We've assumed that half of students are in the treated group.) The source of attenuation is clear in Panel A, as 92.7 percent of

[10] Similar sources of bias have been considered in other contexts. For example, Crépon, Duflo, Gurgand, Rathelot, and Zamora (2013) refers to the potential for labour-market interventions to induce "a game of musical chairs among unemployed workers," that would result in overstating the impact of treatment by comparing a treated worker to a non-treated worker in a given area—the SUTVA violation is that the employment rate among workers in the control group is lower than it would have been absent the program.

[11] See Appendix B for a mathematical demonstration of this "doubling" effect.

[12] $N(70, 10)$ has an inner-99.9 percent ranging in expectation from 46.737 to 93.263. This will later serve as the baseline when we consider the influence of curving in a simulated environment.

treated students do not realize a change in letter grade in response to treatment.[13] Here we strictly enforce the grading rubric, so the source of double counting is evident in Panel B, which identifies the displacement of control students. Just as 3.9 percent of treated students were induced from C to B and 3.5 percent were induced from B to A, a full 3.9 percent of control students are displaced from B to C and 3.5 percent from A to B.

In general, then, we characterize opposing forces at play. To the extent that the mass of treated individuals is not within "a treatment effect" of a letter-grade threshold, the estimated average treatment will necessarily attenuate. After letter-grade transformations, responses to treatment are only partially observable to the econometrician, and the treatment of individuals outside of these β-related distances puts weight on zero in proportion to the mass of treated individuals who fall outside of these intervals. Within these intervals, however, zero-sum grading implies that the individual contributions to the identification of treatment are biased upward. In zero-sum grading environments, when treatment is large enough to close the gap between the highest-performing B student and the lowest-preforming A student, treatment will be double counted in a way. Thus raising the possibility that double-counting treatment within these intervals offsets that there is weight being put on zero outside of these intervals. However, retrieving something like the true effect of treatment on student performance would only be by chance.[14]

2.2 Implications

If the number of letter grades is fixed (e.g., teachers do not respond to treatment by rewarding "better" classes with more As) then contributions to estimated treatment effects are double counted at each letter-grade margin. Alternatively, if the generosity of a grading regime is endogenous to treatment

[13] In this case, 36.1 percent receive a C both before and after treatment, while 30 percent receive an A both before and after treatment. The B category has more turnover within it—26.5 percent of students who were in the B category remain in the B category, despite improving in their raw performance.

[14] Though not representative of the educational environments one typically experiences, an easy special cases to envision (where the weighted changes in grade point perfectly offset in a way that leaves the average treatment effect identified) assumes (i) that the density of raw performance is uniformly distributed, (ii) that internal grade categories divide students into bins of equal size (given uniformity, this implies equal mass), (iii) that the true treatment effect (i.e., the distance treated students improve in their raw performance) is exactly one third of the distance between letter-grade thresholds, and (iv) that the two outermost letter-grade categories (e.g., F and A+) each account for two-thirds of the mass that is accounted for in each of every other letter-grade category. This set of conditions implies that the mass of treated students who are "double counted" is perfectly offset by *twice* the mass of treated students who are too far from the next-higher letter grade for treatment to close that gap (i.e., and thereby contribute to attenuating treatment estimates). While there's good intuition in envisioning this and other such special cases, they are unlikely to exist in practice. Moreover, our main point remains—that we are unlikely to know that a student was evaluated in such an environment, or which students were evaluated in such environments and which were not. In the end, treatment estimates are sensitive to the number of letter-grade categories, and (in non-trivial ways) the mass therein.

(e.g., if instructors reward better classes with more As) then competition for grades is not strictly zero-sum and estimated treatment effects will depend on *how* sensitive teachers are to student performance. In terms of direction, treatment estimates will be lower where instructors are more generous in response to treatment (as generosity implies that there is less double counting). However, as these distinctions are largely unobservable in practice, asking "How zero-sum is the environment treatment is occurring in?" is relevant to any analysis or interpretation of policy evaluations that rely on letter-transformed student performance.

To illustrate several ways in which "who is at the margin" might matter, in Figure 2 we've plotted simulated student performance to illustrate the influence of two different letter-grade mappings. (We've again assumed that the continuous performance of student i in class c, x_{ic}, is distributed $N(70, 10)$.) In Panel A of Figure 2 we plot the CDF of these raw scores, demonstrating their mapping into letter grades assuming that As are assigned to the top-30 percent of the class, Bs are assigned to the next-highest 30 percent, and Cs are assigned to those between the 10th and 40th percentiles (i.e., a zero-sum regime). In Panel B we assume a more-generous rule—As assigned to the top-40 percent, and Bs assigned to the next-highest 40 percent. In each panel, we've highlighted the margins within which treatment-induced movements in the measured performance of students will have even the potential to identify estimates of treatment. With this context in mind, we find several noteworthy implications of letter-grade measures of student performance deserving of mention.

The larger is β itself, the larger is the proportion of students who fall into a β-determined interval around a letter-grade cutoff. This itself introduces an awkward association between the magnitude of treatment and the ability to identify treatment without bias. Where treatment is more effective at increasing student performance (i.e., where β is larger) there is less attenuation in estimated treatment effects (as unmeasurable improvements in performance now become measurable) but also more potential for double counting (where treatment was measurable)—mechanically this is so. Thus, the bias in estimated treatment is itself a function of the magnitude of treatment.

The more generous is the grading rubric, the more likely treatment will be identified off of lower-performing students. This is evident in comparing panels A and B of Figure 2, for example, as the students contributing to measuring treatment at the B and A margins in Panel B are to the left

of those in Panel A.[15] Moreover, even if there are norm-setting rules for generosity—a department governing the proportion of students who receive As or Bs, for example—any underlying variation in student performance will drive changes in the *type* of student ending up within β of a threshold, and therefore the type that ends up away from them (i.e., adding weight to zero).

The larger is the number of letter-grade thresholds, the larger will be the fraction of students within a β of a letter-grade cutoff. As such, the less attenuation bias there will be and, in zero-sum environments, the more "double" counting there will be. While the bias is generally unsignable within schools, we can anticipate systematic distinctions across schools with different grading practices. For example, parameter estimates identified in schools that allow partial-letter distinctions in letter grade (compared to schools only offering A/B/C distinctions) will suffer from less attenuation bias (as there are more students for which their treatment is measurable). Assuming a zero-sum competition for letter grades, there will also be more double counting with additional letter-grade distinctions. Thus, treatment effects should clearly be higher in regimes that offer students plus/minus-modified letters, for reasons other than the efficacy of treatment. We illustrate this in Figure 3.[16] Identical policy initiatives should appear better in institutions that allow plus/minus distinctions in their grading rubrics.[17]

There is a potential confounding of returns to effort at the margin of letter-grade distinctions. Insofar as the behavior of those around letter-grade margins are not representative of all students—our intuition has us anticipating that, if anything, incentives to perform are heightened around letter-grade distinctions—then contributions to identifying treatment in the vicinity of cutoffs will confound treatment and effort. Assuming that proximity to letter-grade margins is known to students, treatment is at least confounded by "effort effects." If students respond similarly on both sides of the cutoff then there could be an argument made that this effort response drops out of the identifying variation (i.e., if students above and below both work harder, then neither gains relative to the other). However, if students respond differently on the different sides of letter-grade thresholds—and there is evidence (Main and Ost, 2014; Oettinger, 2002) that this is the case—signing the bias is made even

[15] In this environment, the groups able to identify positive treatment effects when the top 60 percent receive As and Bs are those within β of 67.5 and 75.2. When the top 80 percent receive As and Bs positive treatment is identified off of those within β of 61.6 and 72.5.

[16] In Appendix B we offer a mathematical demonstration of this same comparison.

[17] If anything, our read of the available information suggests that there is a move toward rubrics that allow plus/minus grades. This implies that similar policy may also appear to perform better over time, not due to any change in the efficacy of the treatment but simply due to the change in rubric.

more problematic.[18] Namely, it will depend on the returns to those marginal increases in effort and whether additional effort moves students to within β of the letter-grade threshold in the absence of treatment, after which treatment then bumps them above the threshold. (The larger that return, the larger is estimated treatment relative to true treatment.) As a rule, then, we might anticipate that any incentive-type effects that are active around letter-grade cutoffs generally, are going to confound treatment estimates—upward bias in treatment estimates if those just below are more responsive to the potential increase in letter grade, and downward bias in treatment estimates if those just above are more responsive to the potential decrease in letter grade.[19]

3 Curves

3.1 Motivating framework

We have not proposed that letter grades are bad practice. Likewise, we will not be proposing that curving grades is bad practice. We are instead proposing that researchers should maintain a broad skepticism and exercise caution when anticipating that treatment effects can be cleanly identified in grade-based measures of student achievement. As it turns out, this will be especially true across students of different ability.

In order to consider the implications of various curving mechanisms on the treatment estimates that are recovered from letter grades *and* curves, a common comparator is necessary—a space in which achievement (and treatment) measures are observable absent any curve. We therefore define a common measure of achievement across students and classes in "raw" scores, and consider the implications of attempts to measure the effect of treatment that changes these raw scores directly but is only evidenced by changes to the associated curved scores. Doing so, we make clear that treatment estimates can vary with how generous a teacher's curve is to toward low-performing students, for example, and that identifying variation can be influenced by how well top-performing students perform. Neither of these is desirable, obviously—especially so given that we can rarely observe "teacher generosity" or the raw

[18] Main and Ost (2014) use a regression-discontinuity approach to evaluate how student performance changes around letter-grade cutoffs. Specifically, they use raw numerical scores on exams and find that students scoring slightly below an 80 (the cutoff for a B) on the first exam perform five percentage points better, on average, on the second exam compared to students who score slightly above an 80 on the first exam. Oettinger (2002) finds that students tend to perform at levels just above the minimum thresholds for various grades and that students who are closer to a grade boundary going into the final exam tend to perform better on the final exam.

[19] In Appendix A we discuss the challenges to identification in percentile rankings, in pass/fail rules, and in gains measures.

scores of one's classmates, for example. More generally, though, as classrooms adopt different curves, or parameterize common curves differently, or realize different inputs that endogenously drive the non-linearities that some curves are shaped by, defining both achievement and treatment in a common space serves to enable the legitimate consideration across various curving environments.

That said, we do care about *post-transformation* achievement measures—we are not positing that one is more important than the other, *per se*, or that it is right to record only raw scores. Merely, we are demonstrating that *we can learn something* about the role of curve- and letter-grade transformations, tools commonly employed in the classroom and measures that we heavily rely on in the economics literature.[20] It strikes us as relevant to ask whether raw or curved scores are the stronger predictors of later outcomes—though, to the best of our knowledge, there is no answer to that question in the literature.[21] There is little evidence identifying the causal role of grades on outcomes, unfortunately. One exception to this rule is Tan (2020), which identifies off of discrete differences in letter grades for students with similar (underlying, continuous) scores at the National University of Singapore—better letter grades result in higher earnings post-graduation. Thus, locally, where raw scores are smooth through the threshold, apparent returns to letter-grade distinctions is consistent with the returns to letter grades being higher than the returns to underlying continuous measures. Arcidiacono et al. (2016) also well identifies a role for grades, in the learning that goes on in students as information is revealed to them—grades play a role in determining choices and therefore outcomes. Thus, to the extent variation in letter grades are inducing changes in outcomes, we should take seriously their construction and the implications of transformations that can imply systematic differentials in the efficacy of policy across students who experience different curves (across classes) or experiences different fallout from the application of curves (within classes).

3.2 Representative transformations

In general, consider some function $g(x_{ic})$, where x_{ic} is student i's continuous raw score in class c, and $g(x_{ic})$ is that student's curved score. In our context, $g(\cdot)$ should be rank preserving—if $x_A \geq x_B$ then

[20] To be clear, we can imagine several motivations for the adoption of a curve, and therein for curved scores to fulfill some worthwhile objectives. For example, instructors may apply curves to keep students happy, to comply with long-standing grade-distribution policies, to manage student expectations, or to manage teaching evaluations, for example. Grade inflation has also motivated norm-setting policies—controlling the fraction of As or Bs received in a class, for example. (Many law schools in the United States now mandate explicit curves in first-year courses. See `https://en.m.wikipedia.org/wiki/List_of_law_school_GPA_curves` for a list of published grading standards in law schools.)

[21] We do know that test scores are predictive of outcomes—the long tradition of including the AFQT in wage equations comes to mind, where it is the percentile score that is typically entered, as this is provided in the NLSY.

$g(x_A) \geq g(x_B)$, which forbids that student A scores higher than student B *before* the curve, but lower than student B *after* the curve. However, as a matter of policy, we have found little conformity around curve adoption other than the adoption of a curve of some sort seeming to be nearly universal.[22]

In Figure 4 we demonstrate the transformation associated with each of these four families of curves, with the mapping of raw scores x_{ic} into curved scores $g(x_{ic})$. In each, we will also convey some of the typical sensitivity available through various parameterizations within these families.[23]

In Panel A we depict "*slope-flattening curves*" (e.g., referred to as a "four-fifths plus 20" curve when parameterized as $.8x_{ic} + 20$) captured generally with

$$f(x_{ic}, a) = ax_{ic} + (1 - a)100 \ , \tag{1}$$

where x_{ic} is the raw score of individual student i in classroom c and $a \in (0, 1)$ is a shape parameter. Generally considered favorable to the lower end of the class, a controls just how generous to the lower end of the class the f-type curve will be.

In Panel B we depict several "*high grade to 100*" rules—curves that move the highest-scoring student to 100, with all others moving by the same degree. This implies that other students' grades are computed as the percentage of the maximum, or, in general,

$$h(x_{ic}) = \frac{100x_{ic}}{\max_c(x_{ic})} \ . \tag{2}$$

Though often employed, some object to the use of "high-grade-to-100" mechanisms as they are more generous to stronger students than to weaker students. (For example, where $\max_c(x_{ic}) = 90$, a student with a raw score of 90 percent, gets a 10-point curve, while a student with a raw score of 60 percent only gets a 7-point curve.) That said, it is our impression that they are reasonably common.

Root curves, which we depict in Panel C, have the property that students with raw scores of 0 or

[22] We will not drill down as far as to reflect on the implications of allowing students to submit work for regrading, or for extra credit, or the implications of variation in withdrawal dates. However, these practices (which can also be implicated as examples of "curves") each have the potential to likewise obscure our ability to identify treatment effects in student-performance data. Note also that Diamond and Persson (2016) documents bunching in Swedish test-score distributions that is interpreted as instructors inflating students who have "bad test day." To the extent systematic, this would again influence our ability to retrieve unbiased estimates of treatment.

[23] In a survey of 119 law school programs, Kaufman (1994) finds that 64 percent of schools implemented a curve. When asked about the type of curve used, many programs report using some type of curve to set a fixed mean for the course and few programs set a fixed percentage of each letter grade. (Gordon and Fay, 2010) likewise reports on the use of curves, reporting that undergraduates students imagine that 63 percent of their courses had curves that moved low grades upward. See Brookhart et al. (2016) for a review of teachers' grading practices and perceptions, which evidences massive variation across teachers.

100 receive no curve, while those with lower scores receive a larger boost than do those with higher scores. A root curve can be captured in

$$r(x_{ic}, b) = 100^{1-b} x_{ic}^{b} \ , \tag{3}$$

where $b \in (0, 1)$ is a shape parameter.

Last, we consider a family of curves that allow instructors to take any two points in a distribution of scores and move one or both of them to known places. For example, suppose that one was interested in moving the minimum and mean scores—these are functions of the x_{ic} in class c, so we notate them as $\min_c^{raw}(x_{ic})$ and $\mu_c^{raw}(x_{ic})$. Defining their new locations as parameters $\min_c^{new}$ and μ_c^{new}, the *two-point transformation* that accomplishes this can be written as

$$t(x_{ic}, \min_c^{new}, \mu_c^{new}) = \min_c^{new} + \frac{\mu_c^{new} - \min_c^{new}}{\mu_c^{raw}(x_{ic}) - \min_c^{raw}(x_{ic})}(x_{ic} - \min_c^{raw}(x_{ic})) \ . \tag{4}$$

While not all applications of this rule will be as formally defined, (4) captures the essence of what might be accomplished in many different ways as instructors aim to manage the minimum and/or mean scores of a classroom. The flexibility of such a transformation is made evident in Panel D of Figure 4.

3.3 Identifying variation in the presence of common transformations

It is often in curved environments such as these that researchers are interested in retrieving estimates of treatment—wanting to evaluate the effect treatment that originated in raw score of student i in class c, x_{ic}, while having only the curved score, $g(x_{ic})$. Below, we consider the effect of f, h, r, and t transformations on the researcher's ability to interpret estimates of treatment from specifications of the form

$$g(x_{ic}) = \alpha + \beta \mathbb{1}(\text{Treated}_i) + \epsilon_{ic} \ , \tag{5}$$

given that treatment is operative in raw scores, x_{ic}. In the presence of curved scores, even when $\mathbb{1}(\text{Treated}_i)$ is exogenous with respect to ϵ_{ic}, treatment estimates will not reliably retrieve the causal effect of treatment across individuals.

As a general rule, signing the bias introduced across the individual contributions to treatment will

require diagnosing the gradient of the curve. Given that curved scores follow some general function $g(\cdot)$, any individual for which $g' > 1$, treatment-induced variation in x_{ic} will identify something greater than the true treatment effect. Likewise, anywhere that $g' < 1$, variation in x_{ic} will identify in $g(x_{ic})$ something smaller than that true underlying variation. In other words, where slopes in Figure 4 are steeper (flatter) than the 45-degree line, treatment-induced contributions to *curved* outcomes will be amplified (attenuated) relative to their true values—since true treatment lives in the domain space of raw scores, x_{ic}, we are generally unable to recover treatment from data on curved scores, $g(x_{ic})$.[24]

In Figure 5 we demonstrate the contributions to treatment-identifying variation across students' raw scores. In panels A and B we consider the two simple linear environments—"flattening" mechanisms, and "high-grade-to-100" mechanisms. In such environments, the bias is knowable (subject to parameters) and follows this g' rule quite cleanly. For example, $f'(x_{ic}) = a < 1$ and $h'(x_{ic}) = 100/\max_c(x_{ic}) \geq 1$. That is, no student in an $f(\cdot)$ or $h(\cdot)$ environment contributes to identifying treatment without bias—"flattening" mechanisms attenuate treatment estimates, and "high-grade-to-100" mechanisms inflate treatment estimates.[25] By implication, only an upper bound on treatment can be identified in an $f(\cdot)$ environment, and only a lower bound on treatment can be identified in an $h(\cdot)$ environment.

Where $g(\cdot)$ is itself non-linear in raw scores, as is the case with a root curve, for example, we find a more-complex series of treatment-identifying biases. As was illustrated in Figure 4, root curves tend to boost lower scores more, and taper out gradually for higher raw scores—the measured contributions to identifying treatment are therefore *inflated* among low-performing students, while *deflated* among high-performing students. In general, the bias in the estimated average treatment effect is therefore unsignable, as

$$r'(x_{it}) = b\ 100^{1-b} x_{ic}^{b-1}\ , \tag{6}$$

which crosses one, given $b \in (0, 1)$. This is evident directly in Panel C of Figure 5. In practice, however, as treatment estimates are amplified only among the lowest performers, we imagine that the practical concern from root-curve transformations will be one of attenuating estimated treatment effects and

[24] An exception, of course, is the "flat curve," through which instructors add a fixed number of points to each student's raw score, leaving the slope coefficient unchanged. In such an environment (maybe this is true of oxymorons?) the ability of the researcher to retrieve the causal parameter is unabated.

[25] Note the special case of treatment in a high-grade-to-100 environment where $x_{ic} = 100$ —any such treatment must attenuate treatment estimates if the raw scale is strictly bound to $x_{ic} \in [0, 100]$, since there are no measurable gains in x_{ic} for one who has already achieved $x_{ic} = 100$.

inflating type-II errors.

In root curves, we also uncover something of a general implication—nonlinear curves can mimic heterogeneity in the efficacy of treatment *where there is no such heterogeneity.* By extension, where one boosts the scores of low-performing students more than the scores of high-performing students—not an uncommon practice, we gather—efficacy tests that rely on grade-based measures of student performance may falsely motivate investing the marginal dollar on lower-performing students, and away from high-performing students.

Two-point transformations are still-more complex, as the unsignable biases resulting from them is likely to remain in practice. Some of this complexity was evident in our earlier representation of the mapping from x_{ic} to $t(x_{ic})$, in Figure 4. In those figures, it was clear that reasonable parameterizations could yield both $t' > 1$ or $t' < 1$, driven by the curve's shape parameters (i.e., instructor preferences). What was not so apparent, though, was the endogenous effects coming from the distribution of raw scores themselves, through the classroom mean. This complexity sets t-type transformations apart from other transformations in interesting ways.

Under $t(\cdot)$ mechanisms, individual contributions to identifying treatment again depend on where in the distribution of raw scores i resides—contributions in t environments are higher (though possibly still biased down relative to true treatment) among low-performing treated students and lower among high-performing treated students. (Like root curves, t-type curves produce a false heterogeneity.) However, individual contributions to identifying treatment are also *indirectly* influenced in t environments, *whether or not individual i is among the treated.* The treatment-induced β-change itself induces changes in the class mean, μ_c^{raw}, which then indirectly influences $t'(\cdot)$ for all i. To better diagnose the contributions to identification, assume for example that treatment falls randomly on $N_c^\star$ of the N_c students in class c, which allows us to separate the implications of β and x_{ic} on post-treatment mean scores as

$$\begin{aligned} \mu_c^{raw}(x_{ic}) &= \left(\textstyle\sum_{j=1}^{N_c^\star}(x_{jc}+\beta) + \sum_{j=N_c^\star+1}^{N_c} x_{jc}\right)/N_c \\ &= \left(\textstyle\sum_{j=1}^{N_c} x_{jc} + N_c^\star\beta\right)/N_c \\ &= \mu_c^{raw(-\beta)} + p\beta \ , \end{aligned} \tag{7}$$

where $p = N_c^\star/N_c$ is the fraction of classroom c that is treated. It will also help to make the role of treatment in mean-shifting mechanisms more explicit, which we attempt to do by redefining $t(\cdot)$ as

$$t^\star(x_{ic}, \min_c^{new}, \mu_c^{new}, p, \beta) = \min_c^{new} + \frac{\mu_c^{new} - \min_c^{new}}{\mu_c^{raw(-\beta)} + p\beta - \min_c^{raw}}(x_{ic} + \beta\mathbb{1}(\text{Treated}_i) - \min_c^{raw}) \ , \tag{8}$$

where we separate the roles of raw score (x_{ic}), treatment magnitude (β), and the fraction treated (p). While untransformed scores change one-to-one with treatment (i.e., $\partial x_{ic}/\partial\beta = 1$ for all treated i), among treated individuals, t-transformed scores respond to β as

$$\frac{\partial t^{\star}(\cdot)}{\partial\beta}|_{\text{Treated}_i=1} = \frac{\mu_c^{new} - \min_c^{new}}{\mu_c^{raw(-\beta)} + p\beta - \min_c^{raw}} - p\frac{\mu_c^{new} - \min_c^{new}}{(\mu_c^{raw(-\beta)} + p\beta - \min_c^{raw})^2}\left(x_{ic} + \beta - \min_c^{raw}\right) \ . \tag{9}$$

This highlights what amounts to a contaminated-control problem, through the influence of β on treated individuals who then contribute to the class mean, which then influences individuals in the control group according to

$$\frac{\partial t^{\star}(\cdot)}{\partial\beta}|_{\text{Treated}_i=0} = -p\frac{\mu_c^{new} - \min_c^{new}}{(\mu_c^{raw(-\beta)} + p\beta - \min_c^{raw})^2}\left(x_{ic} - \min_c^{raw}\right) \ . \tag{10}$$

As treated individuals experience β-increases in raw scores, which are transmitted through $t(\cdot)$ to the control group—specifically through the inclusion of $\mu_c^{raw}(x_{ic})$—untreated students do not merely experience the absence of β increases, but are also hurt by $\beta > 0$ through the curve itself.[26]

In Figure 6 we plot this identifying variation across raw scores—given the contamination of control groups induced by t-type mechanisms, we plot the identifying variation coming from treated and control individuals separately. The estimated treatment effect is then the weighted average of the two, where the weights are merely the fractions of c that are in the teated and control groups. (We report these differences in each panel, where it is the *difference* between the treatment and control lines that is identifying treatment.) While random treatment implies that this weighted average is constant across x_{ic} for a given class, treatment estimates are likely to differ across classes, as the distribution of students across x_{ic} itself differs across c.

In panels A and B we plot this relationship as it relates to the instructor's preference for the new mean, as it relates to average raw score of students. In general, the lower is the instructor's preferred mean (μ_c^{new}) relative to the raw mean, the lower will be estimated treatment effect. Though we do not show the potential for estimated treatment to flip sign, to reduce the mean further would eventually produce *negative* treatment effects.[27] (We attempt to avoid confusion in the theoretical plots, but this

[26] Note that we assume that the minimum score originates in control group. In the simulated environment below we relax this assumption.

[27] Note that there is some bias in the retrieval of treatment estimates even when the $t(\cdot)$ transformation leaves means scores unchanged. Specifically, it can be shown that if $\mu_c^{raw(-\beta)} + p\beta = \mu_c^{new}$ (i.e., no change in mean), and $\min_c^{raw} = \min_c^{new} = 0$ (for simplicity), then for any x_{ic},

$$\frac{\partial t^{\star}}{\partial\beta}|_{\text{Treated}_i=1} - \frac{\partial t^{\star}}{\partial\beta}|_{\text{Treated}_i=0} = 1 - \frac{p\beta}{\mu_c^{new}} < 1 \ . \tag{11}$$

This is evident in Panel D, in particular, where the attenuation is smaller for smaller p.

phenomenon is evident in the simulation results that follow.)

In panels C and D of Figure 6 we leave the raw mean unchanged (at 70) and demonstrate the roles played by variation in minimum scores and in the fraction of students treated. In each case, the t-transformation (even without mean shifting) continues to retrieve unbiased estimates of treatment—note also, that the weighted average of treated and control contributions tends to be more negative at higher levels of raw performance. In Panel D we see that only as $p \to 0$ are unbiased estimates of treatment available to researchers—in the limit, treating just one student in a class minimizes the potential feedback through that student's influence on the class mean, but that influence is still not zero.

3.4 Estimating treatment in the presence of curve transformations

Even in environments where the type of curve is known, estimating the effect of treatment in grades should clearly be considered with great care. Moreover, variation in the application of curves across classes—or even just in the endogenous inputs into a commonly applied curve that vary by class mean, or class minimum, or class maximum—leaves researchers largely unequipped to identify treatment. For example, under both root curves and two-point transformations, even diagnosing whether one has retrieved estimates that are too low or too high requires information about the shape of $r(\cdot)$ or $t(\cdot)$ that the researcher is not likely to have. The point is less about these particulars, though, than it is about the potential for curves of *unknown* mechanism to impart bias of unknown sign and magnitude, leaving researchers unable to identify the effects of treatment.

We next take these diagnoses of individual contributions to the identification of treatment into forming something of an aggregate expectation of the estimated treatment effect (i.e., the weighted average of student contributions across classes). To keep things simple, we will again assume that the raw score is a percentage (between 0 and 100) and that a curved or scaled grade is again a score between 0 and 100. We have in mind that the practitioner applies this curved score to a straight scale, where scores between 90 and 100 are awarded As, scores between 80 and 90 are awarded Bs, scores between 70 and 80 are awarded Cs, and so on—this is similar to that which we have depicted in Figure 4.

Where treatment varies at the individual level, we define raw scores of student i in class c as

$$Raw_{ic} = \alpha + \beta \mathbb{1}(\text{Treated}_i) + \epsilon_{ic} \ . \tag{12}$$

If half of individuals are treated to a true treatment of $\beta = 1$, then $\alpha = 69.5$ will yield an expected class average of 70 out of 100 points—without loss of generality, this will be a convenient benchmark, so we make this assumption unless otherwise noted. (This convenience becomes relevant, in particular, where we consider the "two-point transformation" environment, which is a function of the mean raw score.) As part of the data-generating process, we allow ϵ_{ic} to be additively separable in unobserved heterogeneity at the student level, $\epsilon_i \sim N(0, 5)$, and classroom level, $\epsilon_c \sim N(0, 5)$. In expectation, then, raw scores are distributed $N(70, 10)$, with an inner-99.9 percent ranging in expectation from 46.737 to 93.263.

Having illustrated student-level mappings from raw to curved scores, we have argued for the inability to retrieve an unbiased estimate of the treatment parameter β. In figures 7 through 10 we show the aggregate effects of the curve-induced biases—biases entering at the individual level, as in figures 5 and 6, but here reflected in the aggregate identification of the estimated treatment effect. Under each curving mechanism, we plot the distribution of estimated treatment parameters from models of Raw_{ic} and of $g(Raw_{ic})$, showing estimates derived from models with and without the inclusion of classroom fixed effects.[28] As identifying variation exists within classes, and the biases are introduced across students within classes, absorbing classroom heterogeneity into the error term will not "fix" the problem—researchers will retrieve biased estimates of treatment in both modeling approaches.

In Figure 7 we plot estimates from the first of these four environment, a flattening environment, like $f(\cdot)$ described above. Given the signable biases associated with slope-flattening and high-grade-to-100 curves, we find now the evidence of such expressed in the associated simulations. In Figure 7 treatment estimates are biased down. The most common application of f-type mechanisms is the $.8x + 20$ rule, in which case estimated treatment is $.8\beta$. (As a general rule, where $f(x_{ic}, a) = ax_{ic} + (1 - a)100$, recall that $\hat{\beta} = a\beta$.) In Figure 8 we find the opposite—treatment is biased upward throughout the distribution of students, so produces aggregate estimates that are unambiguously biased upward.

[28] Specifically, in panels A of figures 7 through 10 we've modeled curved scores $g(Raw_{ic})$ without regard for the c-specific component, as in $g(Raw_{ic}) = \alpha + \beta \mathbb{1}(\text{Treated}_i) + \epsilon_{ic}$, and in panels B we've accounted for the c-specific component estimating $g(Raw_{ic}) = \alpha + \beta \mathbb{1}(\text{Treated}_i) + \epsilon_i + \epsilon_c$.

While root curves can, in theory, produce very large treatment effects relative to β, in practice we imagine this being unlikely, as the "very large" contributions are also very low in the distribution of raw scores. As such, what is evident in the simulated environment reflects where in the distribution of scores the mass is. In this case, the mass is largely above the low scores that contribute upward bias—recall from Panel C of Figure 5 that upward bias occurred among students below 25 out of 100. Thus, the plots of treatment estimates in Figure 9 are all biased downward. With identifying variation within classes, these biases are again robust to classroom fixed effects.

In Figure 10 we produce two sets of plots—in Panel A we vary instructor preferences for a new mean (μ_c^{new}) while varying student scores, and in Panel B we vary student scotes (μ_c^{raw}) holding instructor preferences constant. Here, the simulated environment evidences the weighted average of the identifying variation coming from the treated units and (given t-type curving) the set of contaminated controls. Recall from Figure 6 that the higher is raw performance, the lower is the weighted contribution to the treatment estimate—Figure 10 makes evident that this can be sufficient to flip the sign of treatment estimates. (This is precisely the case in our simulated environment when the instructor moves raw scores with mean 70 to curved scores with mean 50.) As before, absorbing classroom heterogeneity into the error term is insufficient to identify treatment without bias—false heterogeneity (also within classroom) in the effect of treatment on outcomes remains, regardless of any estimation of level differences in classroom performance.

We then plot estimates of β from one additional environment—the random assignment of a curve at the classroom level, with equal weight on a "no curve" condition in which only the raw scores is used. In Figure 11, then, we note that when classes adopt a curve randomly, the inclusion of classroom fixed effects also fails to recover the unbiased parameter. This is anticipated, of course, as curves introduce within-classroom heterogeneity in the individual contributions to the identification of β—the problem is not escapable merely though absorbing "the curve" into unobserved classroom heterogeneity.

4 Policy relevance

In terms of going forward, we find ourselves reevaluating policy interventions in light of this fundamental sensitivity we consider. For example, while not always knowable (e.g., the implications of two-point transformations are maybe the most complex), the biases introduced by some curves are at least signable. This, we believe, elevates the value of institutional knowledge as researchers contem-

plate policy evaluations that rely on grade-based outcomes. For example, if we know that the method used to transform grades likely causes the measured gains of a program to appear lower than they otherwise would appear, this is relevant to how we interpret any cost-benefit analysis of the program.

In place of grade-based outcomes for policy evaluation, this analysis also implicitly elevates the importance of other outcomes—high-school graduation and postsecondary education enrollment (Rodríguez-Planas, 2012), the decision to enroll and remain in college (Carrell and Sacerdote, 2017), college entry, college choice, and degree completion (Dynarski, Hyman, and Schanzenbach, 2013), earnings and employment (Deming, Cohodes, Jennings, and Jencks, 2016). However, note that these other outcomes could have a lag of many years after the intervention and, as a result, present their own challenges to causal inference. Likewise, we can imagine scenarios where treatment effects could be re-characterized as lower bounds, given that outcomes were GPA-based.[29] While we also acknowledge the potential for curve-transformed grades to themselves drive other outcomes, variation in non-grade outcomes are possibly more meaningful to be evaluating the efficacy of policy.

Standardized tests also take on added importance in this way, as comparability is maintained and one could imagine systematic differences in performance on such tests being informative. That said, where testing services "equate" raw scores or otherwise adjust scores in such a way as to maintain comparability across testing dates, care again should be taken. In short, transformations need not leave behind a common slope coefficient—given whatever is comparable to $g'(\cdot)$ in this space—from year to year.[30] Even prior to equating, however, there are often transformations implicit in standardized tests. For example, in Panel A of Figure 12 we reproduce the mapping of raw scores (number of questions answered correctly in this case) into scaled scores that the College Board suggests one use when "Scoring Your SAT® Practice Test #1." In Panel B we note changes in the reward (in scaled score) available with one-question improvements in the raw score. Clearly, this is suggestive that even standardized-testing might suffer similarly, and induce heterogeneity depending on where in the

[29] As but one example, Lindo, Sanders, and Oreopoulos (2010) measure the variation in subsequent GPA associated with students having been put on academic probation. Arguably, the measured gains associate with probation would be higher still were treated students all within close-enough proximity to letter-grade distinctions for performance gains to actually materialize in GPA.

[30] In their description of scaled scores, the College Boards describes the process to test takers as follows: "Your raw score is converted to a scaled score of 200 to 800 points, the score you see on your score report. We use a process that adjusts for slight differences in difficulty between various versions of the test (such as versions taken on different days). We do this to make sure there's no advantage in taking the test on a particular day. A score of 400, for instance, on one day's test means the same thing as a 400 on a test taken on a different day—even though the questions are different," (Source: `https://collegereadiness.collegeboard.org/sat/scores/how-sat-is-scored`). Relatedly, Penney (2017) and Ost, Gangopadhyaya, and Schiman (2017) caution against the use of z-standardized test scores, in particular, noting that distributions of test scores can vary across contexts.

distribution of raw performance the treated individuals are. Moreover, with the heaviest mass of scores in the 540 range, we should anticipate attenuation bias for the average student on the SAT (where one-question improvements can be rewarded with no-point scaled improvements), while upward bias among those in the lower tail (where many one-question improvements can be rewarded with 20-point scaled improvements).[31]

Stepping back from the specifics of measurement and identification for a moment, note that the false heterogeneity produced by some curve transformations also suggests that we should be cautious of the results from pilot experiments, in particular. Imagine the political economy surrounding policy innovation, for example. In one world, policy initiatives that are targeted toward students deemed most in need may appear strong among early initiates while failing at scale—the existence of root curves would set this in motion, for example. On the other hand, imagine selection into program participation having early adoption tip toward higher-performing classrooms—if in these classrooms instructors have also chosen curving mechanisms that are generous to low-performing students, contributions to identifying treatment among the high-performing students are biased down. By extension, there is the very real opportunity for early experimental results to appear to have failed efficacy tests and never make it to scale, all the while producing real gains in actual student performance that went unmeasurable simply due to curving practices. In this way, one might imagine policy makers lamenting their fears that "It didn't work on our best students, how could it possibly work at scale?" may be curve-enabled. In a world where we already trade off type-I and type-II errors out of necessity—we're describing a situation in which curves may be tipping us toward excessive type-II error—it is quite possible that there are initiatives worth revisiting.[32]

In Figure 13 we return to our simulated environment—5,000 draws of 150 classrooms of 25 ($n = 3,750$), with treatment randomly imposed on half of students. However, here we assign a curve randomly to each classroom and plot treatment-effect estimates by type of curve adopted by the classrooms. As anticipated, estimated treatment effects are not curve invariant. This is troubling, of course, but also highlights that treatment estimates can be influenced by instructors through their

[31] Calsamiglia and Loviglio (2019) discusses an interesting environment in which there is access to both internal and external performance measures, and document negative externalities associated with having high-performing students in one's class–this is interpreted as curve induced. They suggest caution when using internal grades (i.e., those subject to curve-like transformations) to compare students across schools and classes. Recall again, as they are also relevant here, the examples of Lang (2010), Bond and Lang (2013), and Nielsen (2019).

[32] On a related note, see Davis, Guryan, Hallberg, and Ludwig (2017) for an approach to modeling experiments that informs the researcher about how well a program is likely to work at scale without having to actually test it at scale.

choice of curving mechanism. Where teacher interests align with demonstrating gains with treatment, for example, a "high-grade-to-100" curve could be chosen which would tip toward demonstrating gains (though, two-point transformations that raise the mean are clearly superior). Where there is a misalignment of those incentives, choosing a flattening or root curve would attenuate estimates. Alternatively, consider ranking instructors by their ability to encourage treatment responsive among underperforming students—without regard for transformation, those adopting "high-grade-to-100" curves would excel, with root curves yielding something akin to having a comparative advantage with lower-performing students. At the very least, it would be wise to exercise an additional layer of care when considering apparent heterogeneity across environments, as variation in curving practices can confound heterogeneous treatment effects.

5 Conclusion

Ideally, grades should reflect performance on specific learning criteria and thereby inform students and educators about relative aptitudes. Grades allow students to make better decisions amid uncertainty, and allow policy makers to evaluate the merits of pedagogy or benevolent interventions aimed in one way or another at improving the lives of youth and investing in their futures. Motivated in part by the importance of human capital in promoting health and welfare, researchers often look to grades to measure progress and benchmark policy innovations. However, as part of common practice, performance metrics are typically transformed by both letter-grade assignments and curves.

With respect to letter-grade transformations we highlight two concerns, in particular. First, letter-grade transformations induce attenuation bias to the extent treated individuals are not close enough to letter distinctions (in the absence of treatment) for treatment itself to induce a change in letter grade. Second, identification is further challenged by the contamination of the "control" group when untreated students are displaced in letter grade when those in the treated group (who now outperform them) compete better for scare grades. This is a crowding out, in effect, and amounts to a SUTVA violation that biases treatment estimates upward. In the end, letter-grade transformations imply a complex and unknowable re-weighting of students that yields average-treatment effects that are a function of (i) whether the grading regime implies zero-sum tradeoffs across student grades, (ii) the generosity of the letter-grade assignment generally, and (iii) how many letter-grade distinctions are adopted.

We also document several curve-type environments that capture the essence of what is often less formal in practice, and several letter-grade transformations that represent various levels of generosity. Collectively, we demonstrate their implications on the researcher's ability to retrieve unbiased estimates of treatment. In the end, these transformations interfere with our ability to reliably inform ourselves about the implications of treatment. Biases can be large in magnitude, and estimates of treatment can be of the wrong sign. In common transformations (e.g., moving the mean of a distribution) the direction of bias is often unsignable.

Possibly most troubling, researchers should anticipate that even with approaches to curving that seem quite commonplace (e.g., setting a curve that is more generous to students at the bottom of the class than to those at the top) we find curve-induced patterns of treatment heterogeneity. Some curves attenuate apparent responses to treatment among high-performing students, while exaggerating apparent responses among low-performing students. Among other things, this sort of false heterogeneity raises concern that efficacy tests that rely on curved measures of student performance might encourage spending the marginal dollar inefficiently, to the detriment of student welfare.

References

Angrist, J. D. (2014). The perils of peer effects. Labour Economics 30(C), 98–108.

Angrist, J. D., D. Lang, and P. Oreopoulos (2009). Incentives and services for college achievement: Evidence from a randomized trial. American Economic Journal: Applied Economics 1(1), 136–63.

Angrist, J. D., P. Oreopoulos, and T. Williams (2014). When opportunity knocks, who answers? New evidence on college achievement awards. Journal of Human Resources 49(3), 572–610.

Arcidiacono, P., E. Aucejo, A. Maurel, and T. Ransom (2016). College attrition and the dynamics of information revelation. Working Paper 22325, National Bureau of Economic Research.

Barrow, L., L. Richburg-Hayes, C. E. Rouse, and T. Brock (2014). Paying for performance: The education impacts of a community college scholarship program for low-income adults. Journal of Labor Economics 32(3), 563–599.

Betts, J. R. and J. Grogger (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. Economics of Education Review 22(4), 343–352.

Bond, T. N. and K. Lang (2013). The evolution of the black-white test score gap in grades K–3: The fragility of results. The Review of Economics and Statistics 95(5), 1468–1479.

Bond, T. N. and K. Lang (2018). The black–white education scaled test-score gap in grades k-7. Journal of Human Resources 53(4), 891–917.

Brookhart, S. M., T. R. Guskey, A. J. Bowers, J. H. McMillan, J. K. Smith, and L. F. Smith (2016). A century of grading research: Meaning and value in the most common educational measure. Review of Educational Research 86(4), 803–848.

Butcher, K., P. McEwan, and A. Weerapana (2014). The effects of an anti-grade-inflation policy at Wellesley College. Journal of Economic Perspectives 28(3), 189–204.

Calsamiglia, C. and A. Loviglio (2019). Grading on a curve: When having good peers is not good. Economics of Education Review 73.

Carrell, S. E. and B. I. Sacerdote (2017). Why do college-going interventions work? American Economic Journal: Applied Economics 9(3), 124–51.

Cook, P., K. Dodge, G. Farkas, R. Fryer, J. Guryan, J. Ludwig, S. Mayer, H. Pollack, and L. Steinberg (2014). The (surprising) efficacy of academic and behavioral intervention with disadvantaged youth: Results from a randomized experiment in chicago. Working Paper 19862, National Bureau of Economic Research.

Crépon, B., E. Duflo, M. Gurgand, R. Rathelot, and P. Zamora (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. The Quarterly Journal of Economics 128, 531–580.

Cuellar, A. and D. M. Dave (2016). Causal effects of mental health treatment on education outcomes for youth in the justice system. Economics of Education Review 54, 321–339.

Davis, J. M., J. Guryan, K. Hallberg, and J. Ludwig (2017). The economics of scale-up. Working Paper 23925, National Bureau of Economic Research.

Deming, D. J., S. Cohodes, J. Jennings, and C. Jencks (2016). School accountability, postsecondary attainment and earnings. Review of Economics and Statistics 98(5), 848–862.

Diamond, R. and P. Persson (2016). The long-term consequences of teacher discretion in grading of high-stakes tests. Working Paper 22207, National Bureau of Economic Research.

Dynarski, S., J. Hyman, and D. W. Schanzenbach (2013). Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion. Journal of Policy Analysis and Management 32(4), 692–717.

Figlio, D. and M. E. Lucas (2004). Do high grading standards affect student performance? Journal of Public Economics 88(8), 1815–1834.

Gordon, M. E. and C. H. Fay (2010). The effects of grading and teaching practices on students' perceptions of grading fairness. College Teaching 58(3), 93–98.

Kaufman, N. H. (1994). A survey of law school grading practices. Journal of Legal Education 44(3), 415–423.

Lang, K. (2010). Measurement matters: Perspectives on education policy from an economist and school board member. Journal of Economic Perspectives 24(3), 167–82.

Levitt, S. D., J. A. List, and S. Sadoff (2016). The effect of performance-based incentives on educational achievement: Evidence from a randomized experiment. Working Paper 22107, National Bureau of Economic Research.

Lindo, J. M., N. J. Sanders, and P. Oreopoulos (2010). Ability, gender, and performance standards: Evidence from academic probation. American Economic Journal: Applied Economics 2(2), 95–117.

Main, J. B. and B. Ost (2014). The impact of letter grades on student effort, course selection, and major choice: A regression-discontinuity analysis. The Journal of Economic Education 45(1), 1–10.

Nielsen, E. (2017). How sensitive are standard statistics to the choice of scale? Working paper.

Nielsen, E. (2019). The income-achievement gap and adult outcome inequality. Working paper.

Oettinger, G. S. (2002). The effect of nonlinear incentives on performance: Evidence from "Econ 101". The Review of Economics and Statistics 84(3), 509–517.

Ost, B., A. Gangopadhyaya, and J. C. Schiman (2017). Comparing standard deviation effects across contexts. Education Economics 25(3), 251–265.

Penney, J. (2017). A self-reference problem in test score normalization. Economics of Education Review 61, 79–84.

Rodríguez-Planas, N. (2012). Longer-term impacts of mentoring, educational services, and learning incentives: Evidence from a randomized trial in the united states. American Economic Journal: Applied Economics 4(4), 121–39.

Rubin, D. B. (1980). Comment on: "Randomization analysis of experimental data in the Fisher randomization test". Journal of the American Statistical Association 75, 591–593.

Rubin, D. B. (1986). Which ifs have causal answers? Comment on: "Statistics and causal inference". Journal of the American Statistical Association 81, 961–962.

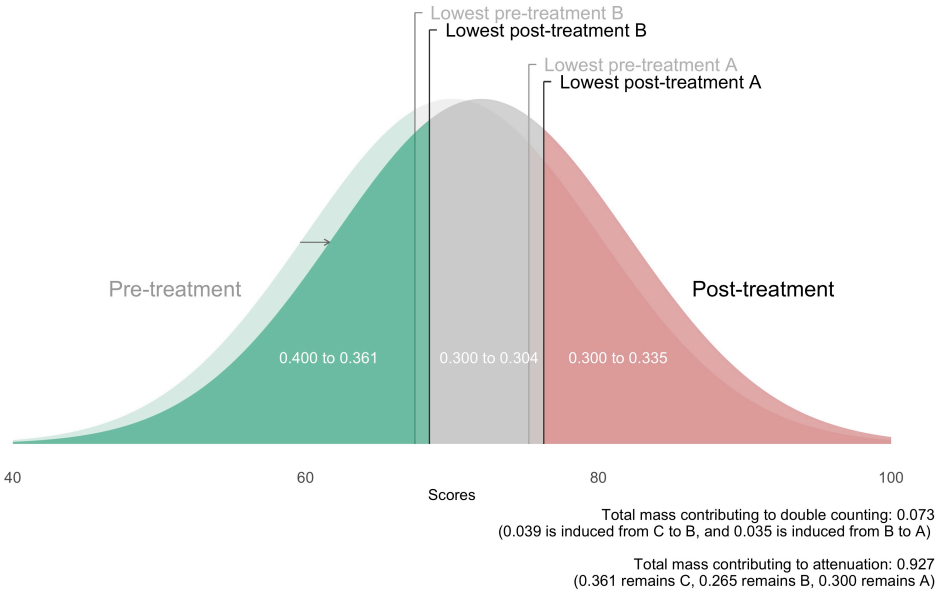
Schrödera, C. and S. Yitzhakib (2017). Revisiting the evidence for cardinal treatment of ordinal variables. European Economic Review 92, 337–358.

Tan, B. (2020). Grades as noisy signals. Working paper.

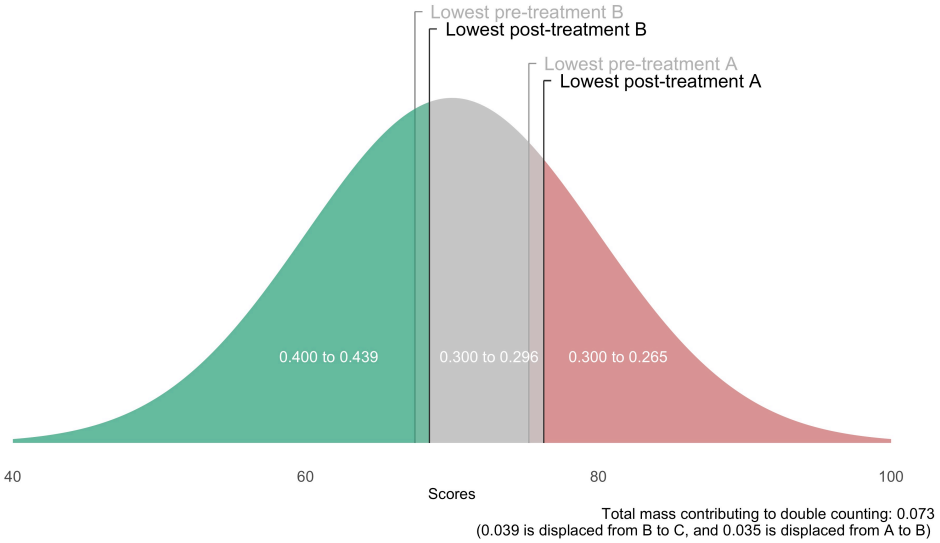
Figure 1: How treatment is measured (and not measured) in letter-grade distributions

The density of treated-student performance (in Panel A) shifts with treatment, yet increases in letter-grade are only experienced by some treated students. The performance of untreated students (in Panel B) does not change, but the letter-grade thresholds increase due to the increase in treated student performance (which is transmitted due to restrictions on the number of each letter). In this strict zero-sum example, for every treated student who experiences an increase in letter grade there is an untreated student who experiences an offsetting decrease in grade (i.e., a STUVA violation that here leads to double counting).

Panel A: Treated students



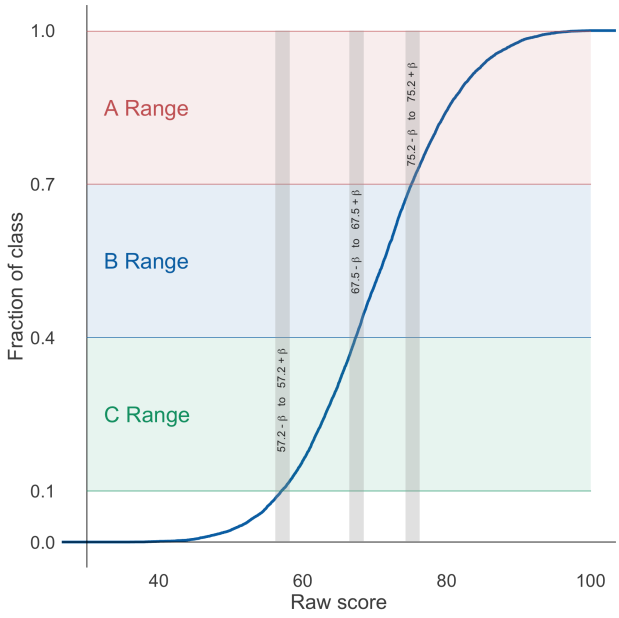
Panel B: Untreated students



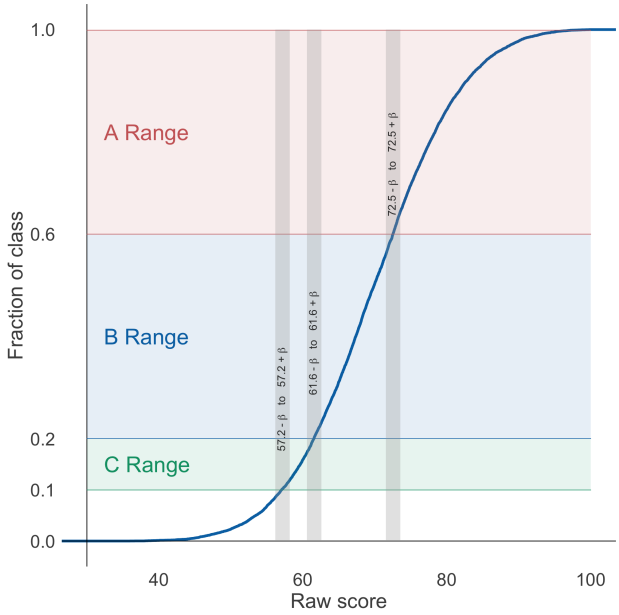
Notes: In each panel we plot the PDF of $x_{ic} \sim N(70, 10)$, separated (by color) into letter-grade categories according to a rule in which the top-30 percent of the class receive As, the next-30 percent receive Bs, and those below the 40th percentile receive Cs. In Panel A we also plot the post-treatment PDF of $x_{ic} \sim N(72, 10)$.

Figure 2: The intervals of domain space that identify treatment effects in letter-grade transformations

Panel A: The CDF of a GPA-transformation with top 60% splitting As and Bs



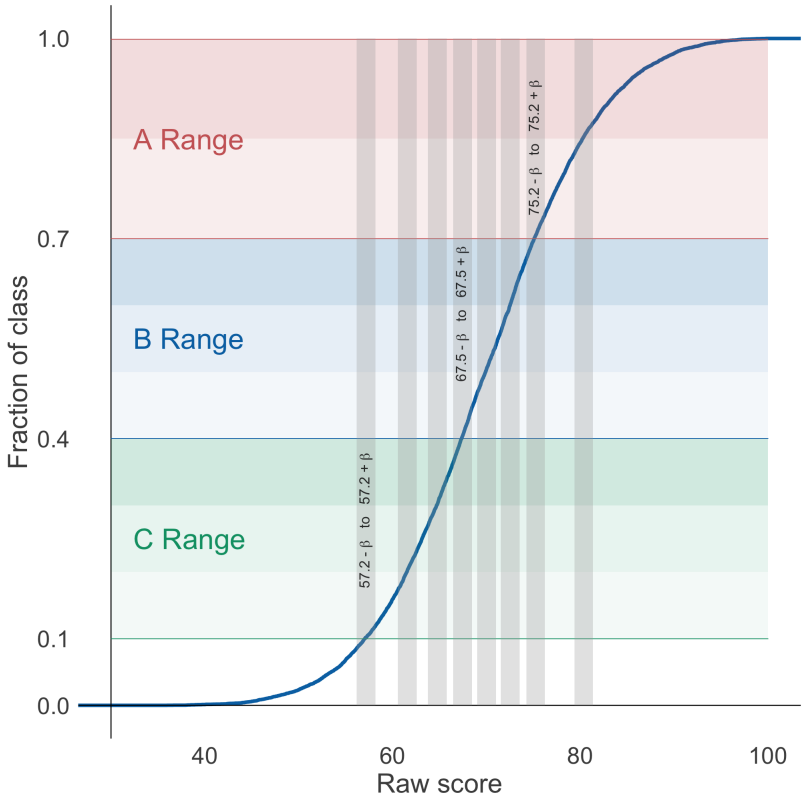
Panel B: The CDF of a more-generous GPA-transformation with top 80% splitting As and Bs



Notes: In each, we plot the CDF of $x_{ic} \sim N(70, 10)$, with letter-grade cutoffs and the students that contribute to treatment (i.e., those within a treatment effect of letter-grade cutoffs) indicated by the shaded regions. In Panel A the top-30 percent of the class receive As, the next-30 percent receive Bs, and those between the 10th and 40th percentiles receive Cs. In Panel B the top-40 percent of the class receive As, the next-40 percent receive Bs, and those between the 10th and 20th percentiles receive Cs. At each letter-grade distinction there exists an interval $\pm\beta$ (shaded in gray) within which treatment is measurable.

Figure 3: Allowing for plus/minus letter grades increases the domain space that identifies treatment

The CDF of a GPA-transformation with top 60% splitting As and Bs, with plus/minus

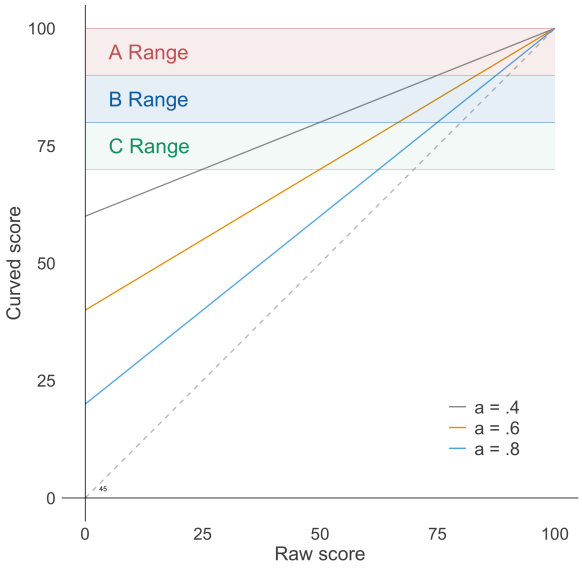


Notes: We plot the CDF of $x_{ic} \sim N(70, 10)$, with letter-grade cutoffs and the students that contribute to treatment (i.e., those within a treatment effect of letter-grade cutoffs) indicated by the shaded regions. The top-30 percent of the class receive As, the next-30 percent receive Bs, and those between the 10th and 40th percentiles receive Cs. At each letter-grade distinction there exists an interval $\pm\beta$ (shaded in gray) within which treatment is measurable.

Figure 4: Various transformations of raw scores into curved scores

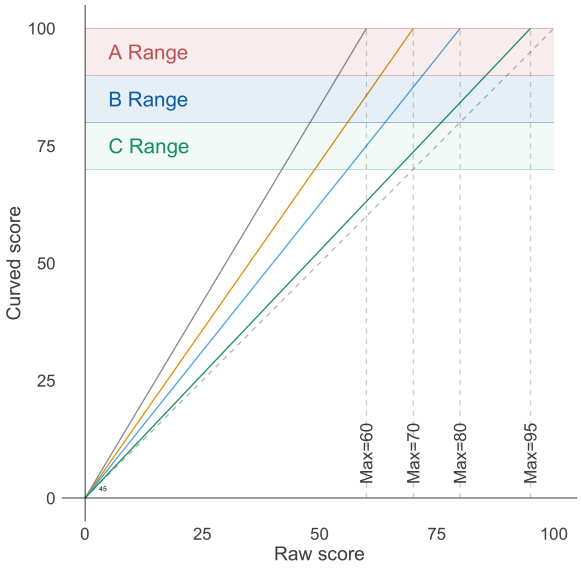
Panel A: Flattening curves

$f(x_{ic}, a) = ax_{ic} + (1-a)100$



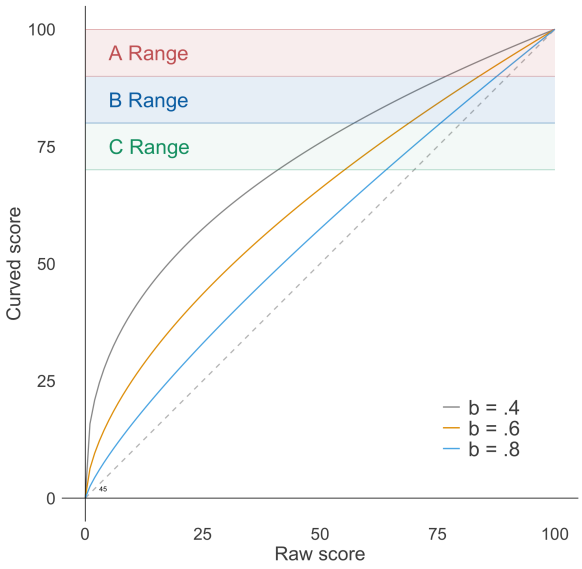
Panel B: High-grade-to-100 curves

$h(x_{ic}) = \frac{100x_{ic}}{\max_c(x_{ic})}$



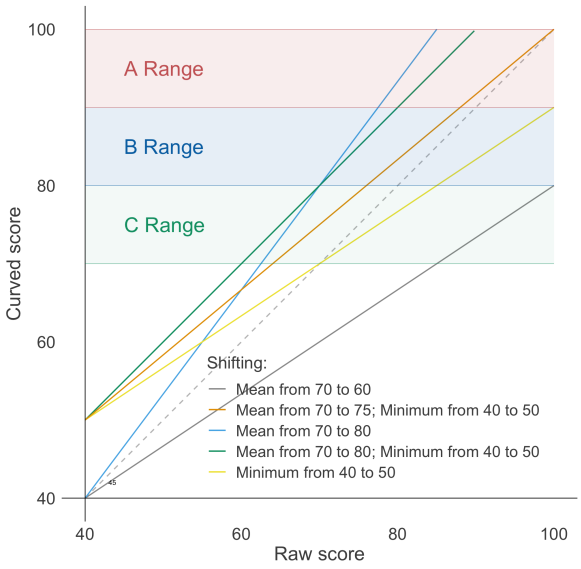
Panel C: Root curves

$r(x_{ic}, b) = 100^{1-b}x_{ic}^{b}$



Panel D: Two-point transformations

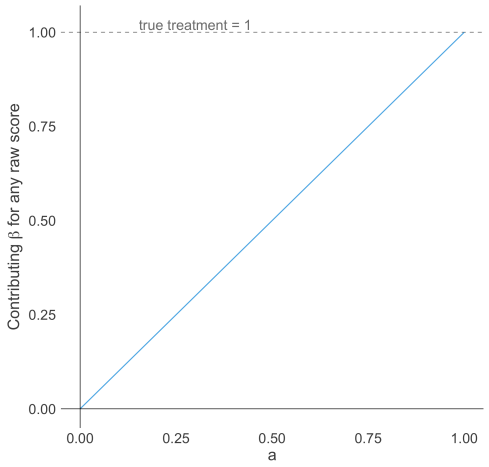
$t(x_{ic}, \min_c^{new}, \mu_c^{new}) = \min_c^{new} + \frac{\mu_c^{new} - \min_c^{new}}{\mu_c^{raw} - \min_c^{raw}}(x_{ic} - \min_c^{raw})$



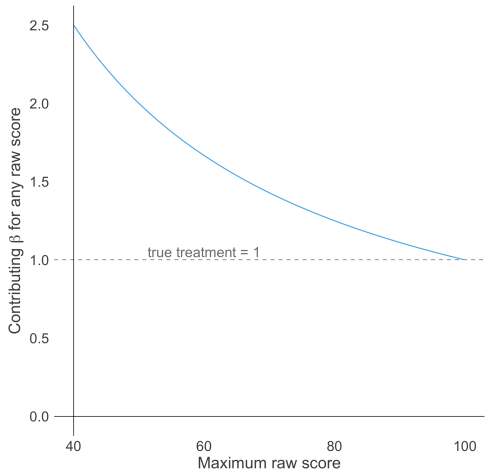
Notes: See Section 3.2 for related discussion. (In Panel D we assume a minimum raw score of 40.)

Figure 5: Individual contributions to identifying treatment in $f(\cdot)$, $h(\cdot)$, and $r(\cdot)$ environments

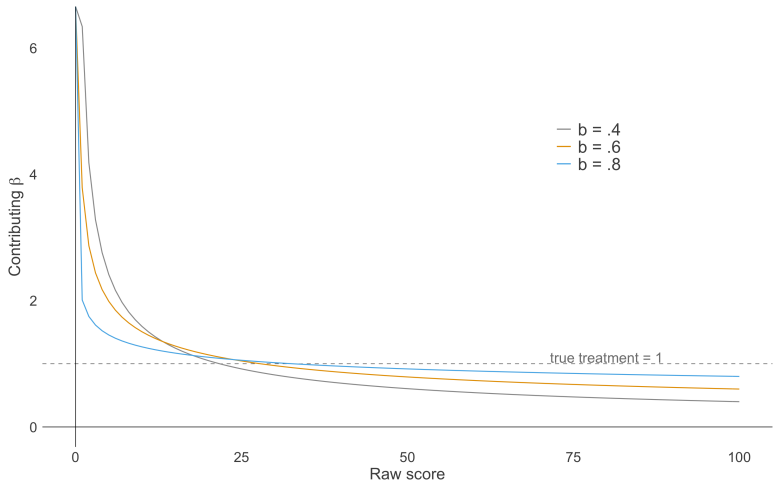
Panel A: Flattening curves attenuate estimates of treatment



Panel B: High-grade-to-100 curves amplify estimates of treatment



Panel C: Root curves yield unsignable biases in estimated treatment and induce false heterogeneity

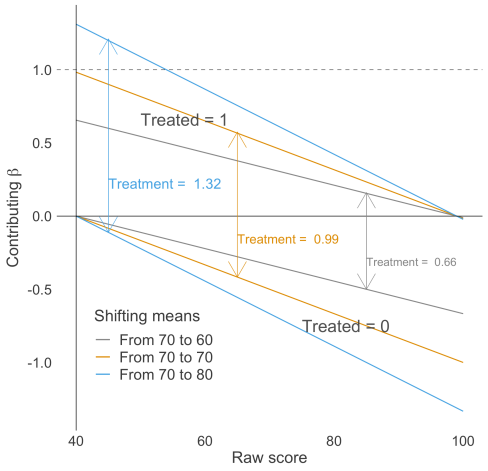


Notes: As in Figure 4, flattening curves are captured in $f(x_{ic}, a) = ax_{ic} + (1-a)100$, high-grade-to-100 curves are captured in $h(x_{ic}) = \frac{100x_{ic}}{\max_c(x_{ic})}$, and root curves are captured in $r(x_{ic}, b) = 100^{1-b}x_{ic}^{b}$.

Figure 6: Individual contributions to identifying treatment for two-point transformations, $t(\cdot)$, and the contamination of controls

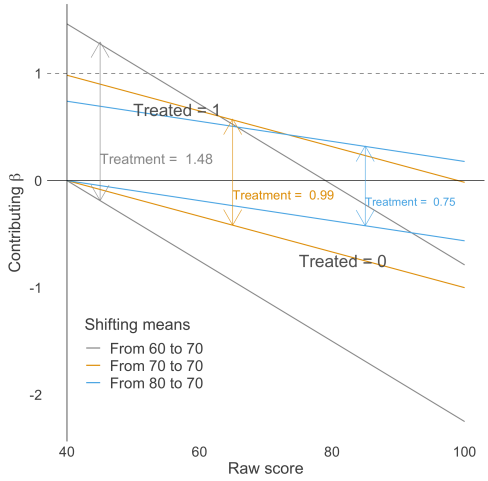
Panel A: Varying instructor preference for mean

(For a given distribution of student performance, treatment estimates are larger the higher is the chosen mean)



Panel B: Varying student performance

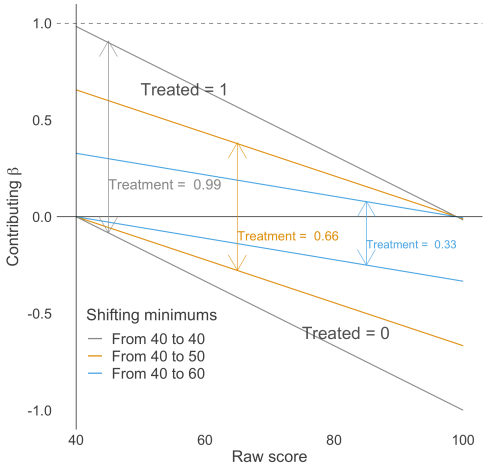
(For a given instructor preference, treatment estimates are larger the lower was student performance)



In panels C and D we assume a mean raw score of 70, unaltered by instructor preference.

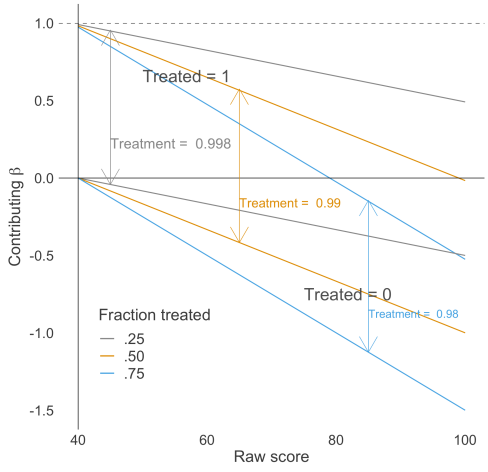
Panel C: Varying instructor preference for minimum

(For a given distribution of student performance, treatment estimates are larger the higher is the chosen minimum)



Panel D: Varying the fraction of students treated

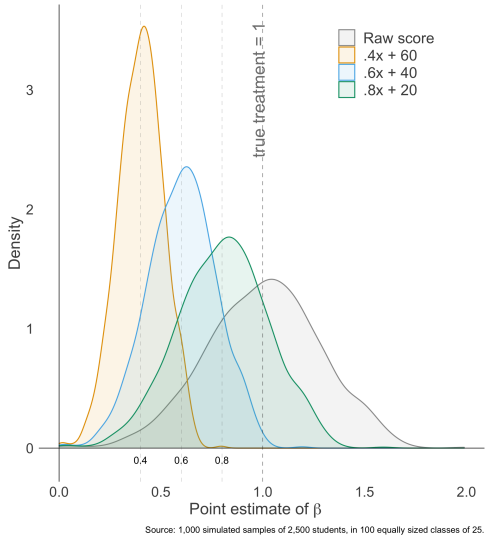
(Treatment estimates are smaller, the larger is the fraction of students treated)



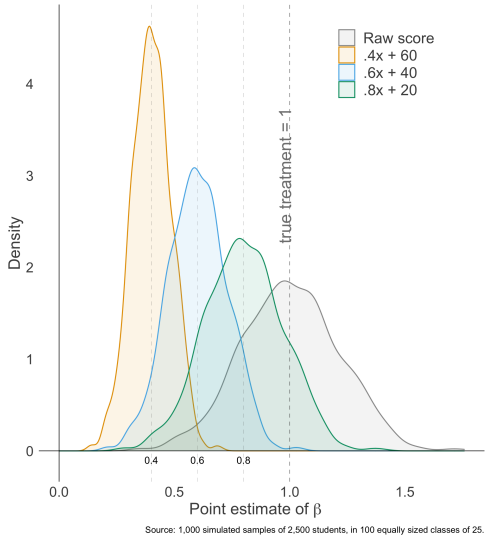
Notes: In all panels, the estimated treatment is the average of treatment- and control-unit contributions, from equations (9) and (10). In reality, this would be weighted by the density of students in x_{ic}. As in Figure 4, two-point transformations are captured in $t(x_{ic}, \min_c^{new}, \mu_c^{new}) = \min_c^{new} + \frac{\mu_c^{new} - \min_c^{new}}{\mu_c^{raw} - \min_c^{raw}}(x_{ic} - \min_c^{raw})$.

Figure 7: "Flattening" curves attenuate treatment-effect estimates

Panel A: Without classroom FE



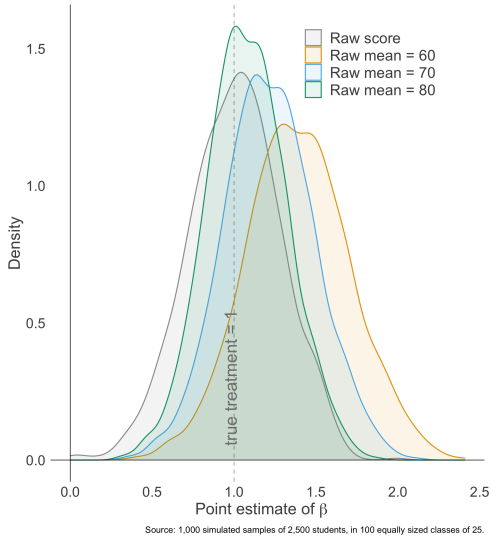
Panel B: With classroom FE



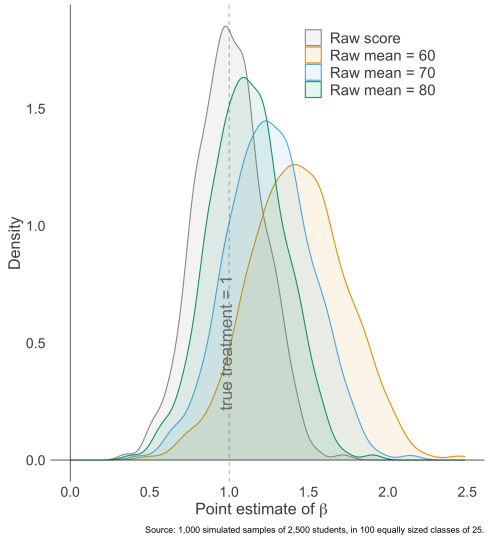
Notes: In all cases, we plot the distribution of estimated treatment effects from models of f-transformed raw scores of individuals i in classrooms c, as in $f(\text{Raw score}_{ic}) = \alpha + \beta \mathbb{1}(\text{Treatment}_i) + \epsilon_{ic}$, where $f(x_{ic}, a) = ax_{ic} + (1-a)100$.

Figure 8: "High-grade-to-100" curves amplify treatment-effect estimates

Panel A: Without classroom FE

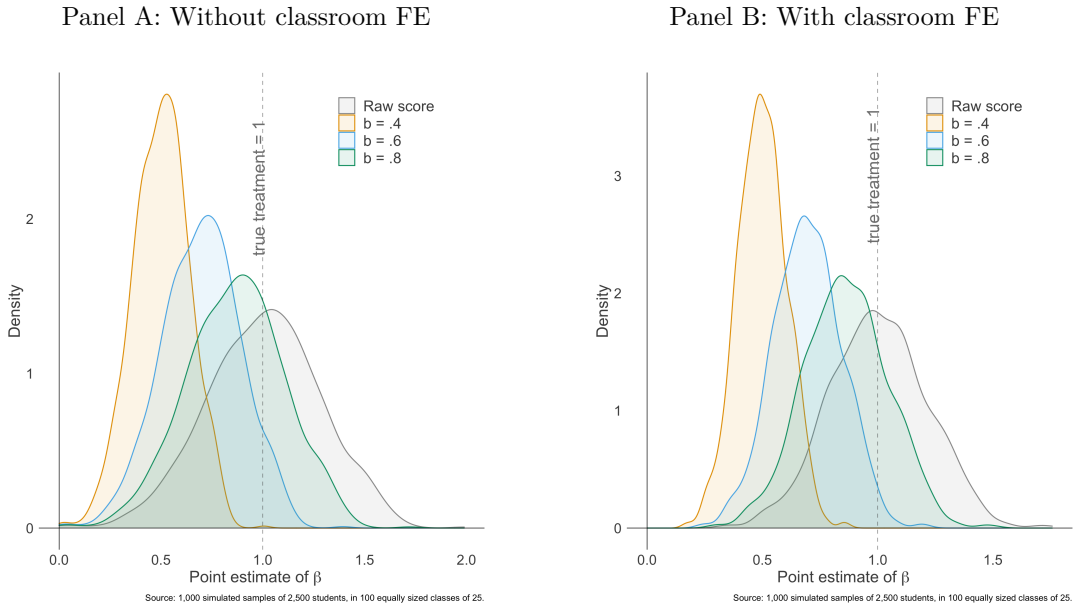


Panel B: With classroom FE



Notes: In all cases, we plot the distribution of estimated treatment effects from models of h-transformed raw scores of individuals i in classrooms c, as in $h(\text{Raw score}_{ic}) = \alpha + \beta \mathbb{1}(\text{Treatment}_i) + \epsilon_{ic}$, where $h(x_{ic}) = \frac{100x_{ic}}{\max_c(x_{ic})}$.

Figure 9: In practice, root curves attenuate treatment-effect estimates

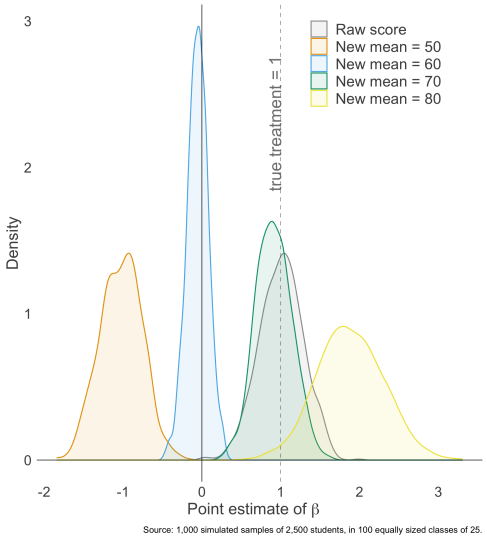


Notes: In all cases, we plot the distribution of estimated treatment effects from models of r-transformed raw scores of individuals i in classrooms c, as in $r(\text{Raw score}_{ic}) = \alpha + \beta\mathbb{1}(\text{Treatment}_i) + \epsilon_{ic}$, where $r(x_{ic}, b) = 100^{1-b}x_{ic}^{b}$. Only when the mass of students falls roughly below 25 percent on a raw scale of 0–100 will root curves *amplify* treatment effects. It is in this way that we imagine the implications of root curves largely leading to attenuation bias.

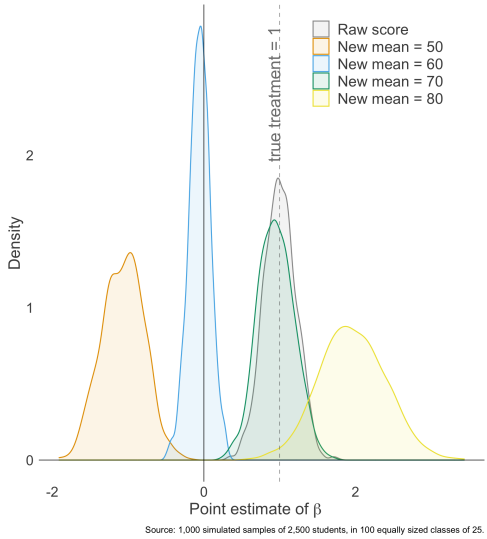
Figure 10: Two-point transformations lead to unsignable biases in treatment-effect estimates

Panel A: Variation in instructor preference around student performance

(i) Without classroom FE

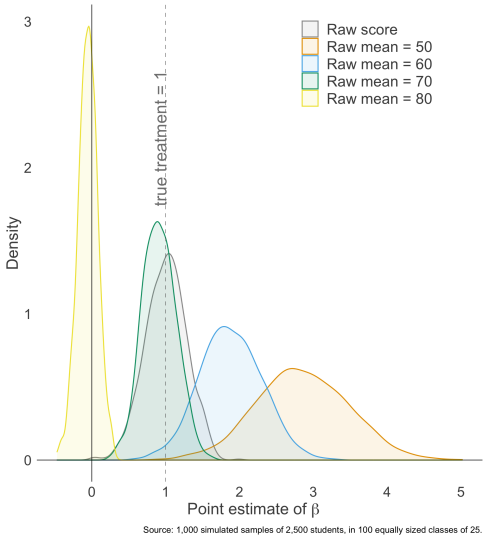


(ii) With classroom FE

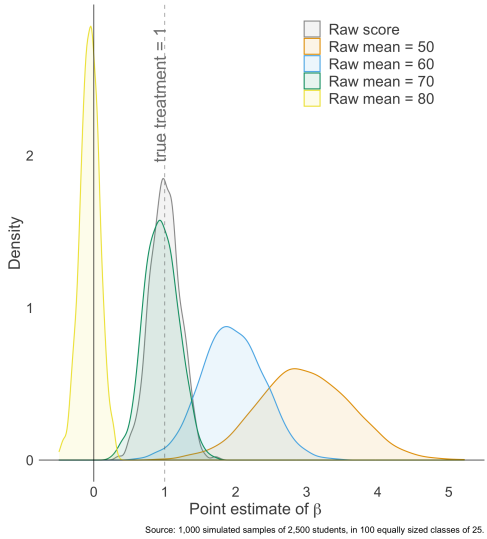


Panel B: Variation in student performance around instructor preference

(i) Without classroom FE

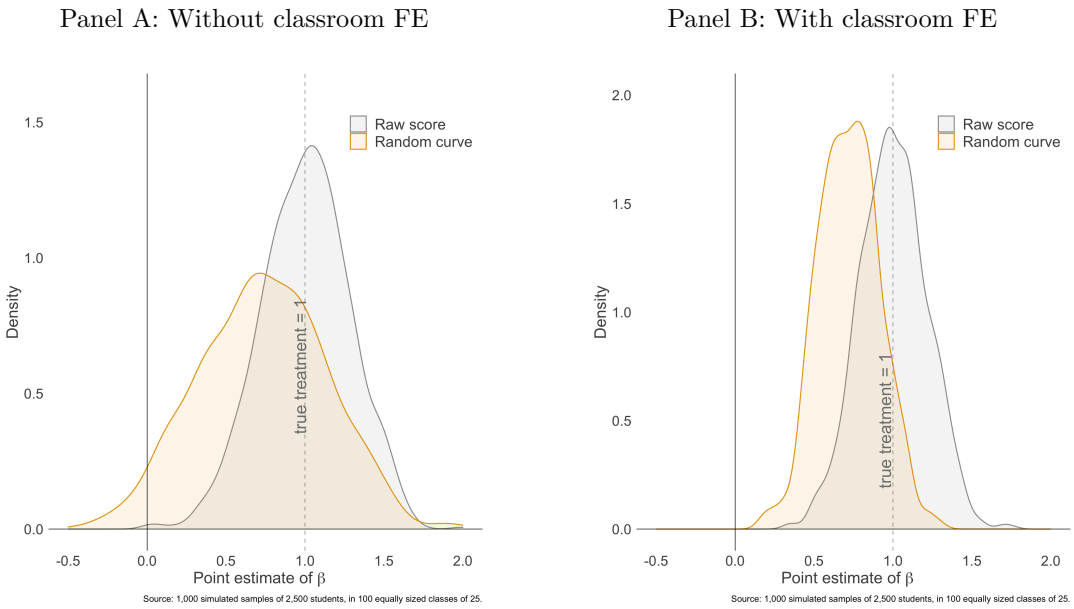


(ii) With classroom FE



Notes: In all cases, we plot the distribution of estimated treatment effects from models of t-transformed raw scores of individuals i in classrooms c, as in $t(\text{Raw score}_{ic}) = \alpha + \beta\mathbb{1}(\text{Treatment}_i) + \epsilon_{ic}$, where $t(x_{ic}, \min_c^{new}, \mu_c^{new}) = \min_c^{new} + \frac{\mu_c^{new} - \min_c^{new}}{\mu_c^{raw}(x_{ic}) - \min_c^{raw}(x_{ic})}(x_{ic} - \min_c^{raw}(x_{ic}))$.

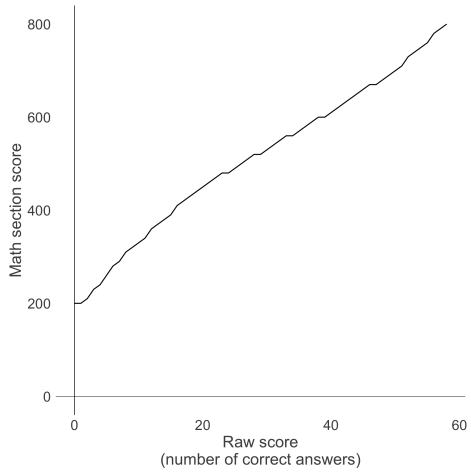
Figure 11: Random curves, and the bias in treatment-effect estimates



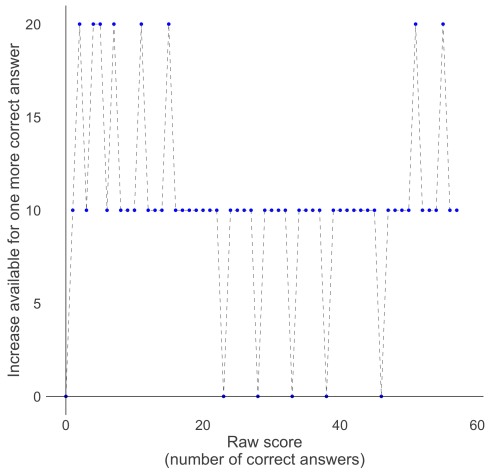
Notes: In all cases, we plot the distribution of estimated treatment effects from models where a curve was assigned randomly to classes (with an equal weight given to "no curve" being assigned).

Figure 12: How raw scores on the SAT Practice Test are converted into scaled scores

Panel A: The transformation of correct responses into scaled math score

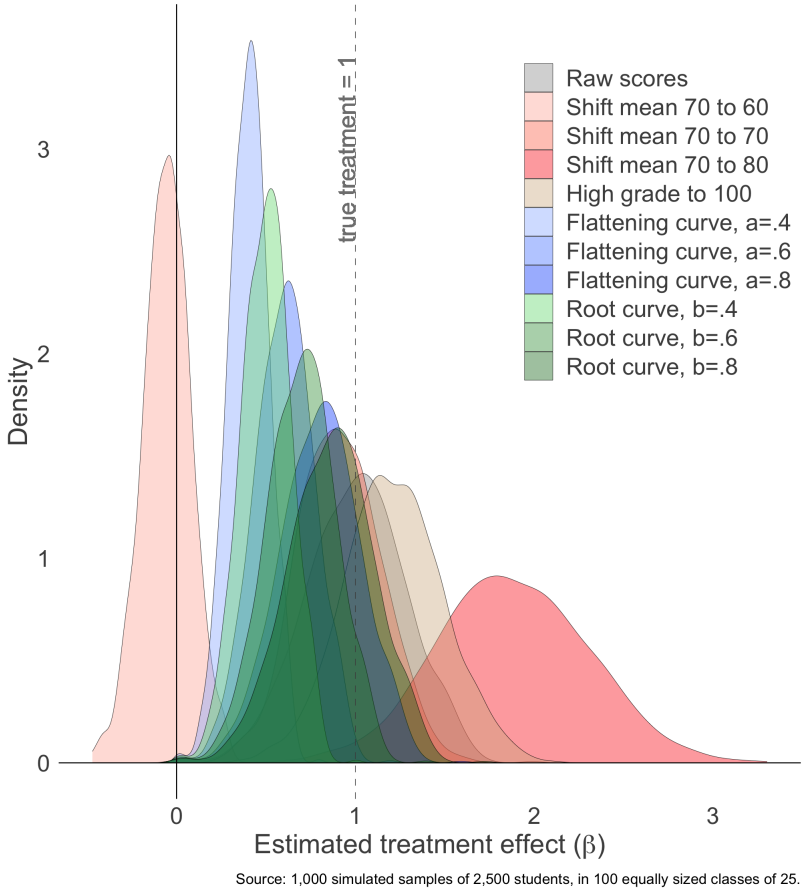


Panel B: The change in scaled SAT math score with a one-question improvement in raw score



Notes: For more, see the "Raw Score Conversion Table" at `https://collegereadiness.collegeboard.org/pdf/scoring-sat-practice-test-1.pdf`.

Figure 13: Estimated treatment effects across classrooms adopting different curves



Notes: We plot distributions of treatment-effect estimates from each of the curve mechanisms and parameterizations described in Section 3.2.

A Recasting the biases in other environments

There are several margins around which some may imagine "fixes." Here, we offer some of our intuition.

A.1 Is this problem avoided by using percentile ranks?

Some may have the intuition that percentile ranks avoid the challenge to identifying the effect of intervention in curved and letter-grade transformations—indeed, this intuition has been shared with us. Yet, percentile rankings are simply an example from the large set of zero-sum, discretized measures of student performance—as one can imagine five or thirteen letter grades, one can likewise imagine 100 percentiles.

In Section 2 we discuss two margins that combine to bias estimates of the effect of treatment in letter-grade-transformed performance measures. First, we note that discretized transformations limit identifying variation to those who are within β of a threshold—this attenuates estimated treatment effects. On this margin, percentile ranks likely mitigate bias, as they relax the "within β of a letter-grade" constraint in favor of a "within β of a percentile." In that way, the opportunity to advance in *rank* is easier, on average, than the opportunity to advance in *letter grade*. As researchers, we are therefore apt to see the evidence of treatment in rank-based metrics in a way that we could miss in letter grades.

That said, the second margin that is in play in rank-based metrics is one around which percentiles are challenged *to an even-greater extent* than are simpler letter-grade transformations. Namely, the zero-sum-driven "double counting" within β of every letter-grade distinction potentially occur more often in a percentile-ranking regime. The implications here will depend on the class size, as the distances that matter are fundamentally driven by the proximity (in x) of nearest-neighbour students—the probability that a treatment of size β is likely to induce treated students into leap frogging students in the control group. In expectation, this source of bias is *exaggerated* in percentile ranks.

In general, the more ranks there are available to distinguish students (e.g., percentiles measured to one decimal of precision allow more "ranks" than strict integer percentiles) the less we should worry about attenuation and the more we should expect the estimated treatment effect to reflect a "double" counting.

A.2 Is this problem avoided by using pass/fail measures?

As with traditional letter-grade transformations, collapsing to pass/fail distinctions does not alleviate the threat to identification. In expectation, pass/fall outcomes behave like one-margin letter-grade transformations—the effect of treatment is only measurable within β of the pass/fail margin. Those who experience treatment *and are within β of a passing grade* will contribute to identifying variation. As no other treated individual is positioned in such proximity to have treatment move them across the pass/fail threshold, treatment occurring outside of this interval only attenuates estimates of the average treatment effect. Within β of the threshold, we again experience the "double-counting" issue—for those within β of the pass/fail threshold, the zero-sum implications lead to inflated estimates of treatment. With treatment effects double counted only at one threshold, though, we it would be reasonable to anticipate that collapsing around pass/fail tends to under estimate the true effect of treatment.

A.3 Does this problem also exist in "gains" measures?

It is reasonable to inquire into the extent models based on students' gains are equally problematic. Inputs into gains-type models rely on inputs, though, so nest multiple observations of students across time. Fundamentally, then, they nest observations across curves. Even if the curves students are exposed to across classes and/or time are common, the implications of their application are endogenous

to the students contributing to those classes. Thus, to reflect on the efficacy of a given intervention in a model of gains is to subject researchers to the same pitfalls we have shown. Outside of the assumption that all curves in all classes for all students in the model, inclusive of the parameters of all of those curves, are common, *and all raw measures of performance for all students contributing to those curves are also common*, treatment is not well-identified.

While we have sided with parsimony in the manuscript, we have simulated a panel of observations, with treatment falling in the middle of the time series. With or without the use of the simulated control group, we find patterns of bias similar to those we report. As the transformation-induced perturbations are within the identifying variation, difference-in-differences designs do not retrieve an unbiased estimate of treatment.

A.4 Are bounds on β informative?

We have already shown that it is an upper bound on β that is identified in an$f(\cdot)$ environment, and a lower-bound on β that is identified in an $h(\cdot)$ environment. We have also argued that retrieving unbiased estimates of treatment is challenged by non-linear curve transformations. However, given $g(x_{ic}+\beta)$ for treated units and $g(x_{ic})$ for control units, it is tempting to imagine that there is still interpretable information in the within-classroom differences available to the econometrician. So we do want to spend just a little time with the potential to bound β in this environment before moving on, and possibly highlight some intuition at the same time. (Note that with the transformation to letter grade, still to come, no matter the hope offered here the researcher will have little ability to retrieve an unbiased estimate of treatment.)

Recall, then, that if $g(x_{ic}+\beta)$ is additively separable then the *difference* between $g(x_{ic}+\beta)$ and $g(x_{ic})$ would contribute to identifying something at least proportional to β. A flattening curve, for example, would allow for the identification of $a\beta$, and a high-grade-to-100 curve would allow for the identification of $\frac{100}{max_c(x_{ic})}\beta$. Thus, even additive separability is insufficient to identify treatment separate from a shape parameter. More relevant, however, is that researchers rarely know the even family from which curves are adopted, or the mixture of curves adopted across classrooms, never mind their various shape parameters. And among the likely curves are nonlinear transformations.

In Figure A1 we plot the PDFs of treated and control units for a simulated sample. We plot the raw scores in Panel A, which we simply draw from $N(70,10)$. Implicit here is that the mean difference (given random treatment assignment) identifies β, and it does. In Panel B we plot the first transformed version of this sample—a flattening curve with $a=.6$, which yields an estimate of $\hat{\beta}=.6\beta$. This illustrates that we identify only an upper bound of β in a flattening environment, as the parameter is proportional to β and governed by $a\in(0,1]$.[33] Without knowledge of a, in a flattening environment we are limited to identifying only bounds on treatment—namely, $\beta\in[a\hat{\beta},\ \hat{\beta}]$.

Unlike in a linear rule, however, a nonlinear transformation does not offer that same ability to bound β. In Panel C of Figure A1 we plot the raw scores transformed by a root curve—this one happens to have been parameterized as $b=.4$. Notably, while the control group will again have greater density at the bottom of the performance distribution, and the treated group more at the top, the non-linearity in $r(\cdot)$ has also perturbed the shape of the treatment distribution—the additional loss of symmetry in the PDFs around their means is evident, for example. We've seen this already (in Figure 5C)—as root curves are less generous at higher x_{ic}, treatment-induced changes in raw score (by β) are implicitly taxed away at a higher rate, the higher is x_{ic}. (Clearly, this leads to downward bias here, though recall from earlier that one could construct an example with sufficient mass in the distribution at low x_{ix} that the bias was positive.)

[33] Recall, $f(x_{ic})=ax_{ic}+(1-a)100$. Therefore, for treated units $f(x_{ic}+\beta)=a(x_{ic}+\beta)+(1-a)100=ax_{ic}+(1-a)100+a\beta=f(x_{ic})+a\beta$. As such, a "*treated minus control*" difference in this environment yield $\hat{\beta}=a\beta$.

While the shape of any $g(x_{it})$ is itself perturbed in non-linear transformations, it is where $g(\cdot)$ acts on x_{it} and on $x_{it} + \beta$ *differently* that we lose the ability to identify β. Let us imagine applying mean-difference estimator to the data in Panel A of Figure A1. Specifically, the mean difference in g-scaled scores is

$$\int_0^{100} x \left[g^T(x+\beta) - g^C(x) \right] dx, \tag{13}$$

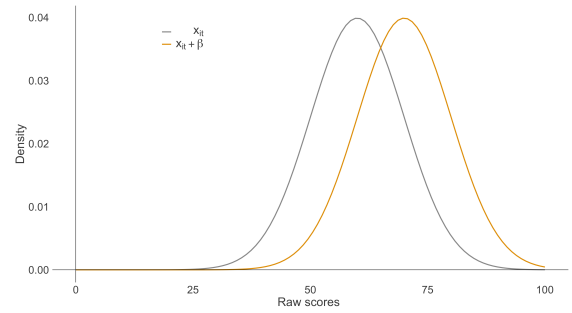
where we use $g(\cdot)$ to capture some generic probability density of scaled scores, adding the notation g^T and g^C to make clear that we are subtracting densities associated with treated and control units. With positive treatment, the density of the treated group has moved to the right, leaving the control group with greater density at the bottom of the distribution. However, just as the shape of the density reflects g, the *change* in the density of the treated group reflects a g-transformed β—it is these perturbations in particular that challenge identification, and that they potentially vary *across* raw scores.

Nielsen (2017) characterizes bounds in a related environment by a re-weighting of outcomes that makes the treatment group's lead on the control group appear either as small or as large as possible. We have considered similarly motivated re-weighting schemes—bounding the difference from below using the distance between the two densities (see Panel B of Figure A1) to increase the weight on low outcomes (where the control group is dominant) and decrease the weight on higher values (where the treatment group is dominant. Or, alternatively, bounding the difference from above by making the treated group's lead appear larger by overweighting high outcomes and underweighting low outcomes. However, in this environment we see little benefit to producing bounds, as they often fail to resolve even the sign of treatment with any confidence when the curve is known. Nielsen (2017) produces something of a "worst-case bounds" scenario by re-weighting according to the *sup norm* of the distance between $g^T(\cdot)$ and $g^C(\cdot)$. However, bounds based on the *sup norm* (across all x) of the distance between $g^T(x+\beta)$ and $g^C(x)$ must be larger than the bounds based on the observed distances between $g^T(x+\beta)$ and $g^C(x)$ at each x, so this again has us inclined to adopt the safer and more-conservative conclusion—we see reason to question inference statements based on treatment estimates that are retrieved from environments in which performance was either curved or transformed by letter grade.

Figure A1: PDFs by treatment classification, and their (treated – control) differences

Panel A: Raw scores ($x_{ic} \sim N(70, 10)$)

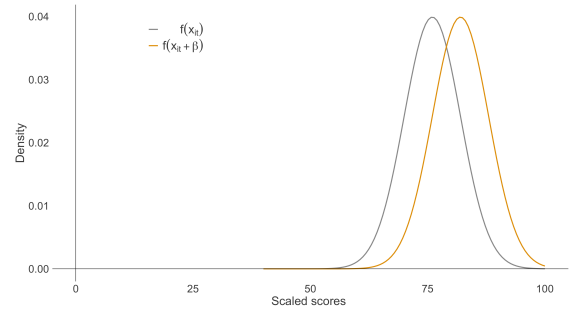
Densities of "treatment" ($x_{ic} + 1$) and "control" (x_{ic})



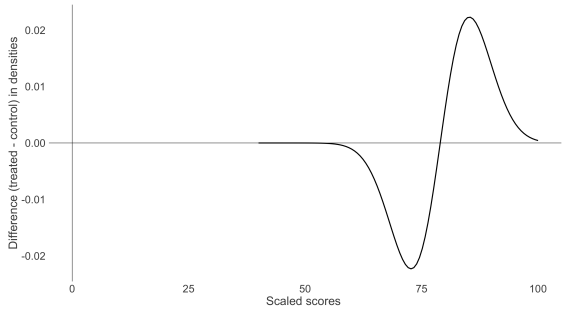
The "$x_{ic} + 1$" less "x_{ic}" difference

Panel B: A flattening curve ($a = .6$)

Densities of $f(x_{ic} + 1)$ and $f(x_{ic})$

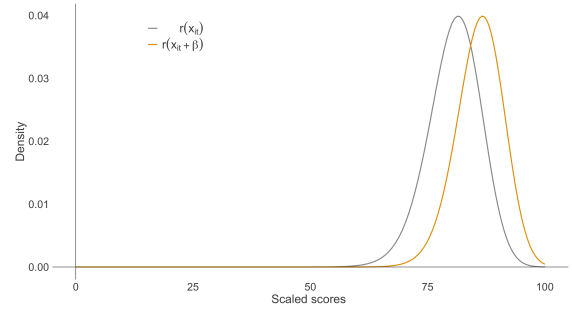


The $f(x_{ic} + 1) - f(x_{ic})$ difference

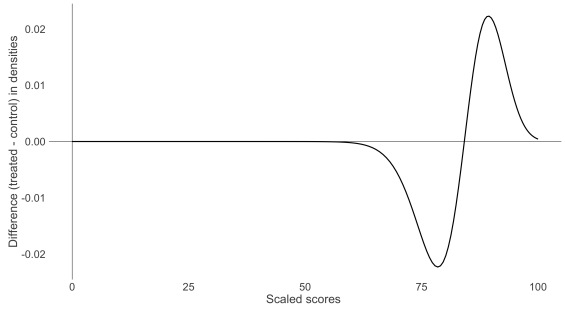


Panel C: A root curve ($b = .4$)

Densities of $r(x_{ic} + 1)$ and $r(x_{ic})$



The $r(x_{ic} + 1) - r(x_{ic})$ difference



Notes: In Panel A we plot the PDFs of $x_{ic} \sim N(70, 10)$ for the control group and $x_{ic} \sim N(71, 10)$ for the treated group. In Panel B we apply those data to the a flattening curve (with $a = .6$) and in Panel C we apply those data to a root curve (with $b = .4$). See Section 3.2 for related discussion of curves.

B ATE estimates in letter grades

B.1 Contributions to estimated treatment effects

In our analysis we have imagined a world in which instructors have grading standards that are exogenous to treatment—with grading cutoffs held constant, treatment induces improvements in the raw performance, x_{ic}, of some students i in class c. In all such regimes, there is an attenuation of estimates of treatment in proportion to the number of treated students not within β of a letter-grade threshold. More formally, in a five-letter regime (i.e., grades of F, D, C, B and A) we can express contributions to the estimated average treatment effect as a function of the induced changes in letter grade, given induced changes in x_{ic},

$$\begin{aligned}
&\overbrace{\int_{D-\beta}^{D}}^{\text{close enough to change letter grade}} \overbrace{f(x_i)\mathbb{1}(T_i=1)}^{\text{the treated}} \overbrace{(D-F)}^{\text{change in grade}}\ dx\ + \\
&\int_{C-\beta}^{C} f(x_i)\mathbb{1}(T_i=1)(C-D)dx\ + \\
&\int_{B-\beta}^{B} f(x_i)\mathbb{1}(T_i=1)(B-C)dx\ + \\
&\int_{A-\beta}^{A} f(x_i)\mathbb{1}(T_i=1)(A-B)dx\ ,
\end{aligned} \tag{A5.1}$$

where all other treated students (i.e., those not within β of the next-higher letter-grade threshold) contribute to the weight on zero. Implicit in (A5.1), these weights can be formally defined as

$$\begin{aligned}
&\int_{0}^{D-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx\ + \int_{D}^{C-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx\ + \int_{C}^{B-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx\ + \\
&\int_{B}^{A-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx\ + \int_{A}^{100} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx\ .
\end{aligned} \tag{A5.2}$$

In (A5.1), then, we have the unambiguous result that estimates of the average treatment effect are attenuated in letter-grade transformations. With additional letter-grade distinctions (e.g., plus/minus letter grades), some among those who were attenuating estimates of the average treatment effect will now find themselves within β of letter-grade margins, and treatment—though just as real in raw performance as without plus/minus letters—will now be evidenced in letter grades. Thus, we conclude that in the absence of zero-sum tradeoffs around letter, treatment estimates are higher in regimes with more letter-grade distinctions.

Many grading regimes are zero-sum, however. For example, any regime that awards letter grades with the top-30 percent of the class sharing in the available As, with Bs assigned to the next-highest 30 percent, and so on, induces zero-sum competition among students (sometimes referred to as *relative grading*).

Where the grading regime is zero-sum, there is a "double counting" of treatment around each letter-grade threshold, which inflates estimates of ATE—this is due to treated students displacing control students within these β-determined intervals. (Recall our discussion of the SUTVA violation implied by the grades of some untreated students being lowered as treated students receive the benefit of treatment, where we argued that "where treatment is large enough to be measurable and letter-grades assignments are zero sum, it is double counted.")

In a five-letter regime with zero-sum tradeoffs, contributions to the estimated average treatment

effect can be expressed as differences between treated and control students,

$$\overbrace{\int_{D-\beta}^{D}}^{\text{close enough to pass}} \overbrace{f(x_i)\mathbb{1}(T_i=1)}^{\text{the treated}} \overbrace{(D-F)}^{\text{change in grade}} dx \underbrace{-}_{\text{their difference}} \overbrace{\int_{D}^{D+\beta}}^{\text{close enough to be passed}} \overbrace{f(x_i)\mathbb{1}(T_i=0)}^{\text{the untreated}} \overbrace{(F-D)}^{\text{change in grade}} dx +$$
$$\int_{C-\beta}^{C} f(x_i)\mathbb{1}(T_i=1)(C-D)dx - \int_{C}^{C+\beta} f(x_i)\mathbb{1}(T_i=0)(D-C)dx +$$
$$\int_{B-\beta}^{B} f(x_i)\mathbb{1}(T_i=1)(B-C)dx - \int_{B}^{B+\beta} f(x_i)\mathbb{1}(T_i=0)(C-B)dx +$$
$$\int_{A-\beta}^{A} f(x_i)\mathbb{1}(T_i=1)(A-B)dx - \int_{A}^{A+\beta} f(x_i)\mathbb{1}(T_i=0)(B-A)dx \ , \tag{A5.3}$$

where all other treated students (i.e., those not within β of the next-higher letter-grade threshold) again contribute to the weight on zero.

As (A5.3) makes clear, the estimated average treatment effect is contributed to by treatment-induced changes in letter grades for treated students (i.e., F to D, D to C, C to B, or B to A) but also by the coincident changes in the letter grades of untreated students (i.e., D to F, C to D, B to C, or A to B). Distributing the negative, (A5.3) can be expressed as

$$\overbrace{\int_{D-\beta}^{D} f(x_i)\mathbb{1}(T_i=1)(D-F)dx}^{\text{treated students}} + \overbrace{\int_{D}^{D+\beta} f(x_i)\mathbb{1}(T_i=0)(D-F)dx}^{\text{untreated students}} +$$
$$\int_{C-\beta}^{C} f(x_i)\mathbb{1}(T_i=1)(C-D)dx + \int_{C}^{C+\beta} f(x_i)\mathbb{1}(T_i=0)(C-D)dx +$$
$$\int_{B-\beta}^{B} f(x_i)\mathbb{1}(T_i=1)(B-C)dx + \int_{B}^{B+\beta} f(x_i)\mathbb{1}(T_i=0)(B-C)dx +$$
$$\int_{A-\beta}^{A} f(x_i)\mathbb{1}(T_i=1)(A-B)dx + \int_{A}^{A+\beta} f(x_i)\mathbb{1}(T_i=0)(A-B)dx \ , \tag{A5.4}$$

which further illustrates the upward biased in the estimated average treatment effect due to the untreated students also experiencing changes coincident with treatment. From here, there are two convenient ways of arguing that this source of bias is a literal "doubling."

First, the notion of zero-sum grading is alone a give away, intuitively. That treatment increases the performance of treated students (both real and measured) is indistinguishable from it coincidently decreasing the measured performance of untreated students—zero-sum grading schemes imply that for every treated student who receives a higher grade there must be an untreated student who now receives a lower grade. That's the only thing treatment can do in such an environment—it has some treated students switch letter grade with untreated students. This is a SUTVA violation that creates a wedge between treated and untreated students, inflating estimated treatment by a factor of two. As a second approach to seeing the doubling of these contributions, one can infer from treatment being random with respect potential outcomes that treated and untreated students are distributed similarly in their pre-treatment raw performance. In expectation, then, the number of treated students who are close enough to a higher letter grade for β to induce a letter-grade difference will equal the number of untreated students who are close enough to the same margin to be overtaken by treated students. That is, $\int_{\lambda-\beta}^{\lambda} f(x_i)\mathbb{1}(T_i=1)dx = \int_{\lambda}^{\lambda+\beta} f(x_i)\mathbb{1}(T_i=0)dx$ for all $\lambda \in \{\text{A, B, C, D, F}\}$. As an equal number of untreated students on the right side of a letter-grade cutoff contribute to estimates of the average treatment effect by exactly the same degree as their treated counterparts do on the left side of the same letter-grade cutoff (in the opposite direction), this amounts to a double counting.

B.2 How do plus/minus distinctions affect estimates of treatment?

With the additional letter-grade distinctions associated with a plus/minus regime, the component parts of the estimated average treatment effect can be expressed as

$$\int_0^{Dm-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Dm-\beta}^{Dm} f(x_i)\mathbb{1}(T_i=1)(Dm-F)dx - \int_{Dm}^{Dm+\beta} f(x_i)\mathbb{1}(T_i=0)(F-Dm)dx +$$
$$\int_{Dm+\beta}^{D-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{D-\beta}^{D} f(x_i)\mathbb{1}(T_i=1)(D-Dm)dx - \int_{D}^{D+\beta} f(x_i)\mathbb{1}(T_i=0)(Dm-D)dx +$$
$$\int_{D+\beta}^{Dp-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Dp-\beta}^{Dp} f(x_i)\mathbb{1}(T_i=1)(Dp-D)dx - \int_{Dp}^{Dp+\beta} f(x_i)\mathbb{1}(T_i=0)(D-Dp)dx +$$
$$\int_{Dp+\beta}^{Cm-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Cm-\beta}^{Cm} f(x_i)\mathbb{1}(T_i=1)(Cm-Dp)dx - \int_{Cm}^{Cm+\beta} f(x_i)\mathbb{1}(T_i=0)(Dp-Cm)dx +$$
$$\int_{Cm+\beta}^{C-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{C-\beta}^{C} f(x_i)\mathbb{1}(T_i=1)(C-Cm)dx - \int_{C}^{C+\beta} f(x_i)\mathbb{1}(T_i=0)(Cm-C)dx +$$
$$\int_{C+\beta}^{Cp-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Cp-\beta}^{Cp} f(x_i)\mathbb{1}(T_i=1)(Cp-C)dx - \int_{Cp}^{Cp+\beta} f(x_i)\mathbb{1}(T_i=0)(C-Cp)dx +$$
$$\int_{Cp+\beta}^{Bm-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Bm-\beta}^{Bm} f(x_i)\mathbb{1}(T_i=1)(Bm-Cp)dx - \int_{Bm}^{Bm+\beta} f(x_i)\mathbb{1}(T_i=0)(Cp-Bm)dx +$$
$$\int_{Bm+\beta}^{B-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{B-\beta}^{B} f(x_i)\mathbb{1}(T_i=1)(B-Bm)dx - \int_{B}^{B+\beta} f(x_i)\mathbb{1}(T_i=0)(Bm-B)dx +$$
$$\int_{B+\beta}^{Bp-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Bp-\beta}^{Bp} f(x_i)\mathbb{1}(T_i=1)(Bp-B)dx - \int_{Bp}^{Bp+\beta} f(x_i)\mathbb{1}(T_i=0)(B-Bp)dx +$$
$$\int_{Bp+\beta}^{Am-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Am-\beta}^{Am} f(x_i)\mathbb{1}(T_i=1)(Am-Bp)dx - \int_{Am}^{Am+\beta} f(x_i)\mathbb{1}(T_i=0)(Bp-Am)dx +$$
$$\int_{Am+\beta}^{A-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{A-\beta}^{A} f(x_i)\mathbb{1}(T_i=1)(A-Am)dx - \int_{A}^{A+\beta} f(x_i)\mathbb{1}(T_i=0)(Am-A)dx +$$
$$\int_{A+\beta}^{Ap-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Ap-\beta}^{Ap} f(x_i)\mathbb{1}(T_i=1)(Ap-A)dx - \int_{Ap}^{Ap+\beta} f(x_i)\mathbb{1}(T_i=0)(A-Ap)dx +$$
$$\int_{Ap+\beta}^{100} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx\ . \tag{A5.5}$$

To compare this to that in a five-letter regime, we assume that the cutoff for an A (or B or C or D) in the five-letter scale is the same as the cutoff for an Am (or Bm or Cm or Dm) in the comparable 13-letter regime—this facilitates an all-else-equal comparison of average treatment effect estimates

with and without plus/minus grades. Exploiting this notational convenience, we express (A5.3) as

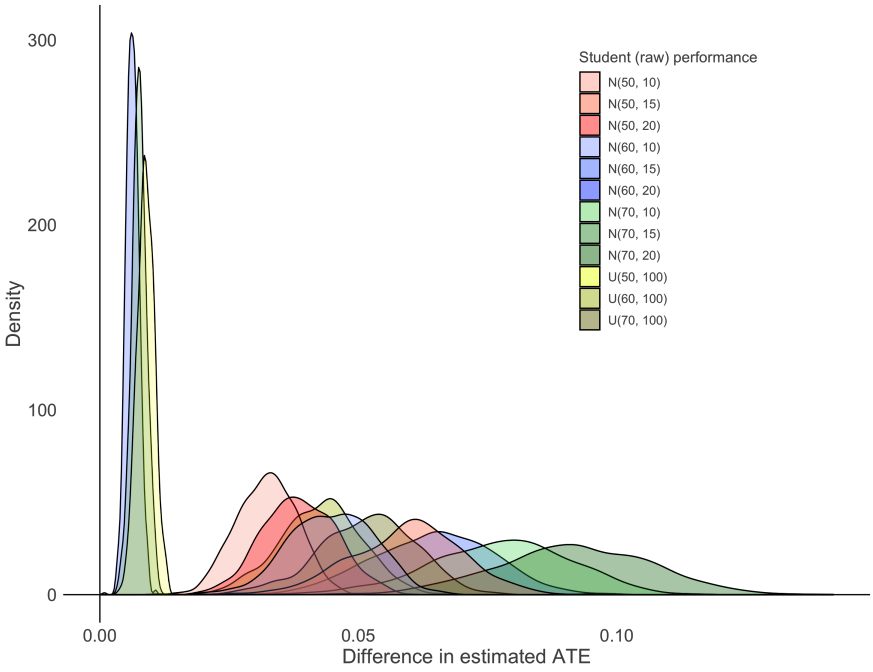
$$\int_{0}^{Dm-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Dm-\beta}^{Dm} f(x_i)\mathbb{1}(T_i=1)(D-F)dx - \int_{Dm}^{Dm+\beta} f(x_i)\mathbb{1}(T_i=0)(F-D)dx + \int_{Dm+\beta}^{Cm-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Cm-\beta}^{Cm} f(x_i)\mathbb{1}(T_i=1)(C-D)dx - \int_{Cm}^{Cm+\beta} f(x_i)\mathbb{1}(T_i=0)(D-C)dx + \int_{Cm+\beta}^{Bm-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Bm-\beta}^{Bm} f(x_i)\mathbb{1}(T_i=1)(B-C)dx - \int_{Bm}^{Bm+\beta} f(x_i)\mathbb{1}(T_i=0)(C-B)dx + \int_{Bm+\beta}^{Am-\beta} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx + \int_{Am-\beta}^{Am} f(x_i)\mathbb{1}(T_i=1)(A-B)dx - \int_{Am}^{Am+\beta} f(x_i)\mathbb{1}(T_i=0)(B-A)dx + \int_{Am+\beta}^{100} f(x_i)\mathbb{1}(T_i=1)\cdot 0\ dx\ . \tag{A5.6}$$

Thus, the difference (A5.5) – (A5.6), which expresses the difference in the estimated average treatment effect in switching to a plus/minus regime, is therefore

$$\begin{aligned}
&\int_{Dm-\beta}^{Dm} f(x_i)\mathbb{1}(T_i=1)(Dm-F-D+F)dx - \int_{Dm}^{Dm+\beta} f(x_i)\mathbb{1}(T_i=0)(F-Dm-F+D)dx + \\
&\quad \int_{D-\beta}^{D} f(x_i)\mathbb{1}(T_i=1)(D-Dm)dx - \int_{D}^{D+\beta} f(x_i)\mathbb{1}(T_i=0)(Dm-D)dx + \\
&\quad \int_{Dp-\beta}^{Dp} f(x_i)\mathbb{1}(T_i=1)(Dp-D)dx - \int_{Dp}^{Dp+\beta} f(x_i)\mathbb{1}(T_i=0)(D-Dp)dx + \\
&\int_{Cm-\beta}^{Cm} f(x_i)\mathbb{1}(T_i=1)(Cm-Dp-C+D)dx - \int_{Cm}^{Cm+\beta} f(x_i)\mathbb{1}(T_i=0)(Dp-Cm-D+C)dx + \\
&\quad \int_{C-\beta}^{C} f(x_i)\mathbb{1}(T_i=1)(C-Cm)dx - \int_{C}^{C+\beta} f(x_i)\mathbb{1}(T_i=0)(Cm-C)dx + \\
&\quad \int_{Cp-\beta}^{Cp} f(x_i)\mathbb{1}(T_i=1)(Cp-C)dx - \int_{Cp}^{Cp+\beta} f(x_i)\mathbb{1}(T_i=0)(C-Cp)dx + \\
&\int_{Bm-\beta}^{Bm} f(x_i)\mathbb{1}(T_i=1)(Bm-Cp-B+C)dx - \int_{Bm}^{Bm+\beta} f(x_i)\mathbb{1}(T_i=0)(Cp-Bm-C+B)dx + \\
&\quad \int_{B-\beta}^{B} f(x_i)\mathbb{1}(T_i=1)(B-Bm)dx - \int_{B}^{B+\beta} f(x_i)\mathbb{1}(T_i=0)(Bm-B)dx + \\
&\quad \int_{Bp-\beta}^{Bp} f(x_i)\mathbb{1}(T_i=1)(Bp-B)dx - \int_{Bp}^{Bp+\beta} f(x_i)\mathbb{1}(T_i=0)(B-Bp)dx + \\
&\int_{Am-\beta}^{Am} f(x_i)\mathbb{1}(T_i=1)(Am-Bp-A+B)dx - \int_{Am}^{Am+\beta} f(x_i)\mathbb{1}(T_i=0)(Bp-Am-B+A)dx + \\
&\quad \int_{A-\beta}^{A} f(x_i)\mathbb{1}(T_i=1)(A-Am)dx - \int_{A}^{A+\beta} f(x_i)\mathbb{1}(T_i=0)(Am-A)dx + \\
&\quad \int_{Ap-\beta}^{Ap} f(x_i)\mathbb{1}(T_i=1)(Ap-A)dx - \int_{Ap}^{Ap+\beta} f(x_i)\mathbb{1}(T_i=0)(A-Ap)dx\ ,
\end{aligned} \tag{A5.7}$$

To evaluate (A5.7), however, one must assume an underlying distribution of raw performance, as the value of (A5.7) is sensitive to where the mass of students is. In particular, note that this is especially sensitive to the mass around failing grades. (Given the non-existence of F+ grades, dropping from a D- (i.e., 0.7 grade points) to an F (i.e., zero grade points instead of 0.3) breaks the symmetric pattern that exists elsewhere in the distribution. Thus, the students around the D-/F letter-grade cutoff

Figure B2: What is the change in estimated treatment effect when switching to a plus/minus regime?



Notes: For each assumed distribution of underlying raw performance, we show the distribution of *increases* in the estimated treatment effect (i.e., the evaluation of Equation A5.7) from 1,000 simulated samples.

become particularly important and the value of (A5.7) will depend on where students are located in the distribution.)

In Figure B2 we show the results of simulating samples to apply to Equation (A5.7). The grading scheme used in each simulation is a zero-sum, with plus/minus grades such that the top-30 percent of the class splits the As evenly, the next-highest 30 percent splits the Bs evenly, the next-highest 20 percent splits the Cs evenly, the next-highest five percent splits the Ds evenly, and the rest of the class receives an F. We assume that treatment increases the performance of treated students by ten percent of the mean performance. We assign grade points to letter grades using the traditional grade-point scale (i.e., A+=4.3, A=4.0, A-=3.7), and simulate (A5.7) 1,000 times for each of twelve different distributions. For Normal distributions, the estimated treatment effect increases when moving to a plus/minus regime in all simulations—on average, (A5.7) evaluates to an increase of between 0.032 to 0.098. Equation (A5.7) is also everywhere positive where raw grades are assumed to be uniform—on average, the estimated average treatment effect is between 0.006 and 0.009 higher in a plus/minus regime. In 12,000 of the 12,000 simulated samples across these various distributions, (A5.7) is positive—we therefore anticipate that the estimated average treatment effect is higher in plus/minus regimes than in grading regimes that use only five letter grades.