Treatment Effects are Not Identified in the Presence of Curves or Letter Grades

Glen R. Waddell and Jenni Putz *

January 2020

Abstract

Individual-performance measures are routinely employed by applied researchers for evidence of policy efficacy. Yet, it is common for measures of student performance to be transformed—discretized through letter-grade transformations or subjected to a curve. We show that transformed grades preclude researchers from retrieving unbiased estimates of the effect of interventions. Moreover, typical curving mechanisms can introduce false patterns of treatment heterogeneity—some curves attenuate measurable responses to treatment among high-performing students, for example, or inflate responses among low-performing students. Even without explicit curving, the transformation to letter grade will itself perturb identifying variation in problematic ways and, in the presence of true heterogeneity, introduce yet another mechanism for bias to afflict estimated treatment effects.

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

JEL: I21, I26, C21

* Waddell (waddell@uoregon.edu) is a Professor of Economics at the University of Oregon and Research Fellow at IZA Bonn. Putz (jputz@uoregon.edu) is a graduate student at the University of Oregon. 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

Researchers often consider identifying systematic variation in student grades as the first fruits of policy intervention. However, it is rare for researchers to observe untransformed measures of performance, with explicit curves and/or letter-grade transformations falling between true performance and the record of performance. In this paper we demonstrate that "curves" and letter-grade transformations systematically distort our perception of treatment effects in otherwise-clean environments—that the ability of researchers to estimate unbiased treatment parameters using transformed performance measures is fundamentally jeopardized by this common practice.

There is a sense in which we follow papers that address concerns that ordinal measures of happiness and subjective notions of well-being can challenge identification (Bond and Lang, 2013; Schrödera and Yitzhakib, 2017). The question might occur to readers, then, whether the setting on which we focus is important enough to deserve added attention. To this we argue that the near-universal practice of curving, and the importance of grade-based performance measures in long and important literatures seems itself to warrant consideration. However, we will also uncover unique aspects within typical grading practices that are worth knowing—the suggestion of treatment heterogeneity where there is none, the sensitivity of treatment effects to a teacher's choice of curve, or generosity, or favour given to those at the bottom of the class. That said, letter-transformed grades (even when transforming raw scores) will itself present empirical challenges that are entirely new to the literature and should broadly challenge those who are in any way employing letter-grade-based outcomes.

We will remain agnostic as to what policy interventions the researcher is attempting to identify, though one could imagine student-level 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 even have in mind the manipulation of peer characteristics in the estimation of peer effects or social spillovers, as in Angrist

(2014).[1] Of the many student-level interventions one could have in mind—the interventions are not what is important here—that researchers look to transformed grades for evidence of treatment is our concern.[2]

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 plight of the researcher, 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 learn hampered by the effect of grade transformations.[3] Moreover, our particular concern that identifying student-level treatment is problematic should not be interpreted as suggesting that comparability is maintained in other group-level interactions—the fundamental non-comparability we describe here is of the sort that can also give rise to non-comparability at higher levels of aggregation.[4]

In our analysis, we consider several curving mechanisms—four fundamental approaches to curving, with shape parameters that change the degree to which they bind. In each, we simulate the random arrival of treatment at the student level and question the researcher's ability to retrieve an unbiased estimate of the effect of that treatment. We demonstrate a fundamental inability to know even the sign of the bias introduced by grade curving without knowledge of the curving mechanism. Moreover, even

[1] 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.)

[2] Adopting an environment in which there is randomly arriving treatment, we demonstrate the concern even in the absence of omitted-variable bias. Of course, one could re-cast the thought exercise entirely, where there is interest in assessing the role of any co-variate on grade-based student performance. For example, the operation of a curve or letter-grade transformations would compound the challenges to identifying the effect of disadvantage on student outcomes.

[3] For example, any student who has ever legitimately shared the sentiment of having "tried so hard this term," belies a certain experimentation with the degree to which their efforts are returned in 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), so to the extent curves obstruct the ability of students to update their own priors, we might also worry about the implications from that perspective.

[4] 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. 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." (The scaling in Lang (2010) is regarding attempts to make inference with respect to comparability or non-comparability across cohorts of students, and how such transformation can be misleading.) 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.

when the mechanism is known and the bias signable, we demonstrate an inability to know the degree to which there is bias introduced—though, under one commonly adopted curving mechanism, without knowledge of the pre-curve properties of the grade distribution and the instructor's preferences for post-curve properties, even the sign of the bias is unknowable. Typical curving mechanisms can also introduce false patterns of treatment heterogeneity—this, we believe, has significant implications for policy, which we discuss in our closing remarks. Even at this point, we paint a rather bleak picture of the researcher's ability to identify treatment in the presence of curved performance measures. However, curved grades often undergo letter-grade transformations, further interfering with the researcher's ability to learn from grade-based measures of student performance—in short, we will demonstrate that variation in the generosity of those letter-grade transformations introducing systematic variation into *which* students identify treatment.[5]

In Section 2 we define curves and describe what is known about the practice of grade curving—though we are all familiar with the practice, the literature is surprisingly light. It is our intention that these four span the range of curves adopted in practice. In Section 3 we quickly walk through a simulated environment in which we demonstrate the implications of curving on treatment-effect estimates—we do so across each curve environment and report result with and without the inclusion of classroom fixed effects. In Section 4 we separately discuss the implications associated with letter-grade transformations—the most-unavoidable takeaways may exist here. In Section 5 we offer some related thoughts and important implications for policy moving forward, concluding in Section 6.

2 Curves

2.1 Motivating framework

We are not proposing that curving itself is bad practice. We are, quite differently, proposing that due to this practice it is unjustified for researchers to anticipate that treatment effects can be cleanly identified in grade-based measures of student achievement. That said, in the construction of an argument we must of course define a space in which true treatment can exist. We will define "true" achievement in

[5] With respect to letter-grade transformations, note a shared intuition available in the consideration of subjective-happiness measurement (Bond and Lang, forthcoming). In both, the researcher is wanting to know about a latent object (i.e., true happiness, true performance) and is charged with drawing good inference from differences or changes in what is observable to the researcher. This will prove challenging in both environments, though in our case (as we will show) there is the additional complexity associated with competition for letter grades being zero-sum in nature.

what we will think of and describe as "raw" scores. This facilitates our comparisons across multiple transformations. To be clear, however, we do not imply a rightness to raw scores over curved scores, as we can imagine several motivations for the adoption of a curve, and therein for curved scores to fulfill some objective. For example, teachers may apply curves to keep students happy, or to comply with long-standing grade-distribution policies, for example, or to manage student expectations, or to game teaching evaluation. Grade inflation has also motivated norm-setting policies recently—controlling the fraction of As or Bs received in a class.[6]

In truth, we see any rule or practice that exists between the collection of raw measures of student performance and the recorded measures of performance as relevant to our question—more important given how little we know of their application in practice. (To that end, we are hopeful that we might motivate additional consideration of curve practices going forward.) The point is, it matters more that we setup a valid comparison across transformations than that we define treatment in a particular space. We therefore abstract away from heavily interpreting the objectives in the mechanism's design and lean more into the variety of pitfalls that result from researchers only observing curved or letter-transformed measures of student achievement.[7]

As long as classrooms adopt different curves (or parameterize common curves differently, or produce different inputs that then drive the endogenous non-linearities that some curves are shaped by), we are comfortable with this abstraction.[8]

[6] 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.)

[7] Ultimately, it could be argued that economists care about achievement measures because they are thought to measure skills that are valuable later in life. From that perspective, it is fair to ask whether raw or curved scores predict outcomes better. Due to the preponderance of transformed achievement measures, there is no answer to that question in the literature. We know that test scores matter—the long tradition of including the AFQT comes to mind—and grades matter, and class rank matters, but we know of no valid test of whether "raw" scores have greater predictive value in outcomes.

[8] While it seems uncontroversial to imagine that there be some space like "raw" scores in which true treatment is defined, we are sympathetic to the consideration of teacher incentives and what model might explain the advent of this or that curve in practice. For us to do otherwise here, though—to define a true treatment in any other "curved" space—would simply move the bias to all other such scales. To define the space for true achievement at the classroom level, as though teachers are themselves choosing the scale on which "true" treatment is to be measured, is also folly—we aggregate and compare across regimes routinely, and without knowing or modeling the selection of any such scale. That is what is important in motivating our efforts to inform best practices. Formally, in positioning true achievement in raw scores, one could argue that we are making two assumptions. First, that raw scores are themselves cardinal measures of achievement. Second, that teachers curve for reasons unrelated to the accurate (cardinal) measurement of achievement. We are fine with this interpretation, as our interest is less in the optimal mechanism design and more in knowing how to evaluate policy evaluations that rely on obscure transformations of measured performance.

2.2 Common transformations

In general, consider some function $g(x_{ic})$, where x_{ic} is student i's raw score in class c, and $g(x_{ic})$ is that student's curved score. In our context, $g(\cdot)$ should be at least weakly monotonic—if $x_A \geq x_B$ then $g(x_A) \geq g(x_B)$, 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.[9]

Throughout our analysis, we will consider four common curving mechanisms. In Figure 1 we demonstrate the transformation associated with each of these four families of curve, with the mapping of raw scores x_{ic} into curved scores $g(x_{ic})$. In each, we will also convey some of the typical movement available with various parameterizations within these families. In Panel A, for example, we depict slope-flattening curves—often 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.

[9] 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 are often 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 teachers inflating students who have "bad test day." To the extent systematic, this would again influence our ability to retrieve unbiased estimates of treatment.

Root curves, which we depict in Panel C, have the property that students with raw scores of 0 or 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 1.

2.3 Identifying variation in the presence of known transformations

It is often in curved environments such as these, in range spaces $g(x_{ic})$ given domain spaces x_{ic}, that researchers are interested in retrieving estimates of treatment. Below, we consider the effect of f, h, r, and t transformations on the researcher's ability to interpret $\hat{\beta}$ 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}. Even when $\mathbb{1}(\text{Treated}_i)$ is random with respect to ϵ_{ic}, β will be unidentified in the presence of curved scores.

As a general rule, signing the bias on the partial treatment will require diagnosing the gradient of the curve. Given that curved scores follow some general function $g(\cdot)$, anywhere in the domain space

that $g' > 1$, treatment-induced variation in x_{ic} will identify something greater than the true treatment effect. Likewise, anywhere that $g' < 1$, variation in x_{it} will identify in $g(x_{it})$ something smaller than that true underlying variation. In other words, where slopes in Figure 1 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_{it}, we are generally unable to recover treatment from data on curved scores, $g(x_{it})$.[10]

In Figure 2 we demonstrate the contributions to $\hat{\beta}$-identifying variation across student performance. 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 $\hat{\beta}$ without bias—"flattening" mechanisms attenuate treatment estimates, and "high-grade-to-100" mechanisms inflate treatment estimates.[11] By implication, only an upper bound on β can be identified in an $f(\cdot)$ environment, and only a lower bound on β 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 β-identifying biases. As was illustrated in Figure 1, 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 2. 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 inflating type-II errors.

In root curves, we also uncover something of a general implication—nonlinear curves can mimic

[10] 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.

[11] Note the special case of treatment in a high-grade-to-100 environment where $x_{ic} = 100$ —any such treatment must attenuate $\hat{\beta}$ 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 $x_{ic} = 100$.

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 1. 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, though 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 β) 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 performance, 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.[12]

In Figure 3 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 the *difference* between the treatment and control lines is identifying $\hat{\beta}$.) While random treatment implies that this weighted average is constant across x_{ic} for a given class, $\hat{\beta}$ is 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 performance of students. In general, the lower is the instructor's preferred mean (μ_c^{new}) relative to the raw mean, the lower will be $\hat{\beta}$. Though we do not show the potential for $\hat{\beta}$ to flip sign, to reduce the mean further would eventually produce *negative* treatment effects.[13] (We attempt to avoid confusion in the theoretical plots, but this phenomenon is evident in the simulation results that follow.)

In panels C and D of Figure 3 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

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

[13] Note that there is some bias in the retrieval of $\hat{\beta}$ 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.

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.

Even in environments with a known curve, 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.

3 Simulated curve transformations

We next take these diagnoses of individual contributions to the identification of treatment into forming something of an aggregate expectation of $\hat{\beta}$ (i.e., the weighted average of student contributions across classes). Before we do, however, note that letter-grade assignments and grade-point equivalents are themselves transformations, and treatment estimation in GPA-data will have its own nuanced implications, even in the absence of curving. We will consider this directly after focusing on an analysis of classroom-level applications of curving mechanisms.

To keep things simple, we will 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 1.

Where treatment is random 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 become 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 4 through 7 we show the aggregate effects of the curve-induced biases—biases entering at the individual level, as in figures 2 and 3, but here reflected in the aggregate identification of $\hat{\beta}$. 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.[14] 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 4 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 4 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 5 we find the opposite—treatment is biased upward throughout the distribution of students, so produces aggregate estimates that are unambiguously biased upward.

While root curves can, in theory, produce very large $\hat{\beta}$ relative to β, in practice we imagine this

[14] Specifically, in Panel A of each figure 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 Panel B we've accounted for the c-specific component estimating $g(Raw_{ic}) = \alpha + \beta \mathbb{1}(\text{Treated}_i) + \epsilon_i + \epsilon_c$.

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 2 that upward bias occurred among students below 25 out of 100. Thus. the plots of $\hat{\beta}$ estimates in Figure 6 are all biased downward. With identifying variation within classes, these biases are again robust to classroom fixed effects, as evidenced in Panel B.

In Figure 7 we produce two sets of plots—in Panel A we vary instructor preferences for a new mean (μ_c^{new}) while varying student performance, and in Panel B we vary student performance (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 3 that the higher is raw performance, the lower is the weighted contribution to the treatment estimate—Figure 7 makes evident that this can be sufficient to flip the sign of $\hat{\beta}$. (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.

Before considering letter-grade transformations, we 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 8, 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 partial identification of β—the problem is not escapable merely though absorbing "the curve" into unobserved classroom heterogeneity.

4 Letter-grade transformations

Thus far, we have not engaged in the question of letter-grade transformations or grade-point averages. Moreover, note that we have not reported simulation results for each individual curve, with and without the additional GPA transformations that likely follow in practice. In lieu of that discussion, we would like to put a wrapper around the above analysis, more generally, as transformation into letter grades

are likely where we see the biggest of issues and arguably where make the more-novel contribution.

4.1 The identification problem

In short, when student performance is transformed into letter grades, the effect of any treatment of size β can only ever be evidenced among individuals already within distance β of a letter-grade cutoff. Even though treatment presumably increases x_{ic} by β for all treated individuals, the letter-grade can only change for those within β of a letter-grade threshold. Clearly, the use of letter-grade rules implies that treatment estimates do not reflect the entire class as a whole, but only those at the margin of a letter distinction, and we should think of $\hat{\beta}$ being identified only by variation within $[-\beta, \beta]$ bandwidths around letter-grade cutoffs.[15] In this way, *who* is at the margin matters.

However, even within these endogenous (to β) bandwidths treatment is not well identified when the letter-grade assignments are themselves introducing zero-sum tradeoffs within their design.[16] Namely, in any such letter-assignment, for any treated individual within β of a letter-grade threshold there is an individual in the control group who is at the margin of being displaced—a clear violation of the stable unit treatment value assumption (Rubin, 1980, 1986). Though any zero-sum environment would face this challenge, in our context it is specifically those in the control group who are within $-\beta$ of the same cutoff (i.e., on the other side) who are contaminated.[17] As treated individuals on the left-hand side of any threshold experience β-improvements, for example, they displace one of the "controls" on the right-hand side of the same threshold who did not experience the improvement. Measurably, they then suffer an equal but opposite change in grade-point. In the end, contributions to identifying $\hat{\beta}$

[15] In the presence of curved scores, these distances would be characterized in $g(\beta)$-denominated spaces around cutoffs, which would vary depending on the curve applied. Given any non-linearity in $g(\cdot)$, for example, they need not be symmetric around a given threshold or common across thresholds.

[16] While the generosity may differ widely, many letter-grade transformations can be articulated as, for example, the top-30 percent of the class sharing in the available As, with Bs assigned to the next-highest 30 percent, and so on. In that way, it set's up as zero-sum around grading margins—around β-sized margins. It is these to which we refer here, as simple rules that map x_{it} or $g(x_{it}$ scores into letters according to absolute cutoffs (90 percent is an A, 80 percent is a B, etc), which are typically coincident with movements in the mean score such that the "straight" curve makes sense. We discuss this mechanism as a "two-point transformation," in Section 2.2.

[17] This bias has been pointed out 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.

within each $[-\beta, \beta]$ bandwidth of a letter-grade distinction will be biased upward.[18]

Thus, we find opposing forces. To the extent that the mass of treated individuals is not within β of a letter threshold, the estimated average treatment will necessarily attenuate—after letter-grade transformations, responses in x_{ic} (or in $g(x_{ic})$) are partially unobserved to the econometrician, and any treatment occurring to individuals outside of these small bandwidths puts weight on $\beta = 0$ in proportion to the mass of treated individuals falling outside of these same bandwidths. However, within bandwidths, the partial identification of treatment is biased upward. While raising the possibility that double counting treatment within these bandwidths offsets the weight on $\beta = 0$ outside of the relevant bandwidths, retrieving something like the true treatment would only be by chance.

4.2 Other implications

To illustrate several ways in which "who is at the margin" might also matter, in Figure 9 we've plotted the same $N(70, 10)$ that was the baseline in the simulated environments against two different letter-grade mappings. (We have not added the complication of a *per se* curve, as the concern over GPA transformations will remain even in the absence of curve-like transformations.) In particular, in Panel A we plot the CDFs of raw scores mapped into letter grade 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. 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 x_{ic} will have even the potential to identify $\hat{\beta}$. This highlights the first of several relevant dimensions that have implications for our ability to make inference from GPA measures.

Among other noteworthy implications:

There is a likely 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. Our intuition has us anticipating that, if anything, incentives to perform are heightened around letter-grade distinctions and, if those margins are known to students in a way that allows them to best respond to that

[18] This is by a factor of two in a strict zero-sum environment without curving, which suggests that were one to be able to identify *who* in a class was indeed marginal (in letter grade), one could estimate $\hat{\beta}$ on subsamples of those marginal students, and retrieve an unbiased estimate of β by the division of $\hat{\beta}$ by two. While true in this special case, we anticipate data limitations making even this impractical. Moreover, $g(\cdot)$ can lead to $\hat{\beta}$ being inflated by an unknown factor, in which case division by two is of no solution. (Recall also that this determination is further complicated by the relevant bandwidth itself being proportional to β.)

proximity, then the confoundedness would bias $\hat{\beta}$ upward relative to what would be measured were it also to be contributed to by students not at the margin of letter-grade distinctions.

The more generous is the mapping of scores into letter grades, the more likely treatment will be identified off of lower-performing students. This is evident in comparing panels A and B of Figure 9—in x_{ic} space, Panel-A bandwidths are strictly to the left of Panel-B bandwidths.[19] 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 then itself drive changes in the type of student ending up within β-bandwidths of the identifying thresholds, and the type that ends up away from them (i.e., adding weight to $\hat{\beta} = 0$).

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—the less attenuation bias there will be in $\hat{\beta}$ and the more "double" counting there will be. While the bias is generally unsignable, we should clearly fear drawing 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 but from more double-counting—they should clearly be higher, for reasons entirely other than the efficacy of treatment. Identical policy initiatives should look better merely because an institution allows for plus/minus distinctions in their grading rubrics.[20] We illustrate this in Figure 10.

The larger is β, the larger is the proportion of students who fall into a $[-\beta, \beta]$ range around a letter-grade cutoff. This itself introduces an awkward association between the magnitude of treatment and the ability to identify treatment without bias. For larger β, there is less attenuation in $\hat{\beta}$ (that was due to treatment falling where it couldn't be seen) but also more "double" counting (where treatment was seen)—mechanically this is so. Thus, the bias in $\hat{\beta}$ is itself increasing in β.

[19] In our simulated 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.

[20] The 2009 High School Transcript Study (HSTS) reports that the most common GPA scale among US public schools is the four-point grade scale (HSTS, US Department of Education). In this scale, the A equals four points, B equals three points, C equals two points, D equals one point, and F equals zero points. (HSTS uses this four-point grade scale to standardize each student's GPA, regardless of original scale.) It is our understanding that colleges are more likely to allow 13 grade distinctions (A+, A, A-, B+, etc.) though large university systems do continue to employ the traditional five-point scale. If anything, our read of the available data suggests that there is a move toward rubrics that allow plus/minus grades. (This suggests that similar policy may appear to perform better over time, simply due to the change in rubric.)

4.3 Discussion

Letter-grade transformations clearly interfere with our ability to retrieve the causal parameter of interest. Thus, even without direct scaling mechanisms, that academic performance is often transformed into letter grade should itself be a concern. In the "happiness" literature, Bond and Lang (forthcoming) argue that we require strong identifying assumptions to draw inference from subjective measures of happiness, and that results based on subjective-happiness scales should be heavily conditioned. In both environments the researcher is wanting to know about a latent object (i.e., true happiness, true performance) and is charged with drawing good inference from differences or changes in what is observable to the researcher.[21] In educative environments, however, it is as though we have a measure of performance that first undergoes transformation that renders identification challenging (as in figures 1 through 8, for example) and *then* we add a letter-grade transformation to further frustrate policy evaluation.[22] In any case, we caution strongly against anticipating that researchers have the ability to retrieve the causal parameter of interests from such representations of student performance.

As some readers have imagined that percentile rankings would offer a quick fix, we note that such practice exhibits much the same set of tradeoffs that exist with letter-grade rules. For example, the "β-bandwidths" we discuss above remain active—for treated students more than β away from their neighbour, there is insufficient movement in raw score for treatment to trigger a change in percentile rank. For this reason, $\hat{\beta}$ is attenuated in a percentile-rank regime. Moreover, where β is large enough for treatment to be evidenced in percentile rank, there is an equal and offsetting reduction in the percentile rank of that neighbour, which then inflates the measured effect of treatment—where it matters enough to be measurable, it is double counted. (Changes in percentile rank are more likely to be triggered by a given β than are changes in letter grade, but the mechanics are largely the same as letter-grade rules.)

[21] See Bond and Lang (2018) for attempts to replicate key results from the happiness literature, showing that none of the findings can be obtained relying only on nonparametric identification.

[22] If we add back any transformation $g(\cdot)$ to the mapping—this transformation would live between the determination of x_{ic} and the application of letter-grade transformation itself—the slope g' then acts on raw scores by changing the range of contributors around each letter-grade threshold in Figure 9. Specifically, where $g' > 1$ the range around letter-grade thresholds is wider (i.e., more students fall within $|g(\beta)|$ of a threshold than would fall within $|\beta|$ of a threshold), and where $g' < 1$ the range is more narrow (i.e., fewer students fall within $g(\beta)$ of a threshold).

5 Policy relevance

The clearest of policy prescriptions is somewhat discouraging—as a general rule, we've lost confidence in the ability to reliably estimate the effects of policy intervention in student grades, tempting though it is given it's a first-order measure of academic performance. Even without direct scaling mechanisms, however, that academic performance is so-often transformed into letter grade should itself be a concern, given our reliance on these measures for policy evaluation.

In place of grade-based outcomes for policy evaluation, we implicitly elevate 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). Likewise, we can imagine scenarios where treatment effects could be re-characterized as lower bounds, given that outcomes were GPA-based.[23] 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 other wise 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.[24] Even prior to equating, however, there are often transformations implicit in standardized tests. For example, in Panel A of Figure 11 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

[23] 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.

[24] 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`).

even standardized-testing might suffer similarly, and induced heterogeneity depending on where in the 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).[25]

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 might fear the fallout 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.[26]

In Figure 12 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,

[25] 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).

[26] 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.

of course, but also highlights that treatment estimates can be influenced by teachers through their 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.

6 Conclusion

As a matter of course, we anticipate that 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 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, grades are typically transformed by curves and letter-grade assignments.

We document several curve-type environments that capture the essence of what can at times be less formal in practice, and several letter-grade transformations representing various levels of generosity. Collectively, we demonstrate their implications on the researcher's ability to retrieve unbiassed estimates of treatment. In the end, these transformations stand in the way of the researcher's ability to retrieve unbiased estimates of treatment. Biases can be large in magnitude, and estimates of treatment can be of the wrong sign. In common transformations—moving the mean of a distribution, for example—the direction of bias is often unsignable.

Moreover, typical curve mechanisms introduce false patterns of treatment heterogeneity—some curves attenuate apparent responses to treatment among high-performing students, while exaggerating apparent responses among low-performing students. 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.

With respect to letter-grade transformations we highlight two concerns, in particular, even when applied to un-curved measures of student performance. First, letter-grade transformations induce attenuation bias to the extent treated individuals are not close enough to letter distinctions (or treatment is not large enough) for treatment to induce a change in letter grade. Second, this attenuation is offset by the contamination of likewise-marginal individuals in the control group who are displaced (in letter grade, that is) by the treatment of those in close proximity. In the presence of true treatment heterogeneity across the distribution of student ability, letter-grade transformations yield average-treatment effects that are a function of the generosity of the letter-grade assignment and how many letter-grade distinctions are adopted. In any case, unbiased estimates of treatment are unrecoverable.

In light of this analysis, we question inference statements based on treatment estimates that are retrieved from environments in which performance was either curved or transformed by letter grade.

References

Angrist, Joshua D, "The Perils of Peer Effects," *Labour Economics*, 2014, *30* (C), 98–108.

_ , Daniel Lang, and Philip Oreopoulos, "Incentives and Services for College Achievement: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, January 2009, *1* (1), 136–63.

_ , Philip Oreopoulos, and Tyler Williams, "When Opportunity Knocks, Who Answers? New Evidence on College Achievement Awards," *Journal of Human Resources*, 2014, *49* (3), 572–610.

Arcidiacono, Peter, Esteban Aucejo, Arnaud Maurel, and Tyler Ransom, "College Attrition and the Dynamics of Information Revelation," Working Paper 22325, National Bureau of Economic Research 2016.

Barrow, Lisa, Lashawn Richburg-Hayes, Cecilia Elena Rouse, and Thomas Brock, "Paying for Performance: The Education Impacts of a Community College Scholarship Program for Low-Income Adults," *Journal of Labor Economics*, 2014, *32* (3), 563–599.

Betts, Julian R and Jeffrey Grogger, "The Impact of Grading Standards on Student Achievement, Educational Attainment, and Entry-Level Earnings," *Economics of Education Review*, 2003, *22* (4), 343–352.

Bond, Timothy N and Kevin Lang, "The Evolution of the Black-White Test Score Gap in Grades K–3: The Fragility of Results," *The Review of Economics and Statistics*, 2013, *95* (5), 1468–1479.

_ and _ , "The Sad Truth About Happiness Scales: Empirical Results," Working Paper 24853, National Bureau of Economic Research July 2018.

_ and _ , "The Sad Truth About Happiness Scales," *Journal of Political Economy*, forthcoming.

Calsamiglia, Caterina and Annalisa Loviglio, "Grading On A Curve: When Having Good Peers Is Not Good," *Economics of Education Review*, 2019, *73.*

Carrell, Scott E and Bruce I Sacerdote, "Why Do College-Going Interventions Work?," *American Economic Journal: Applied Economics*, July 2017, *9* (3), 124–51.

Cook, Philip, Kenneth Dodge, George Farkas, Roland Fryer, Jonathan Guryan, Jens Ludwig, Susan Mayer, Harold Pollack, and Laurence Steinberg, "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 2014.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora, "Do Labor Market Policies have Displacement Effects? Evidence from a Clustered Randomized Experiment," *The Quarterly Journal of Economics*, 2013, *128*, 531–580.

Cuellar, Alison and Dhaval M Dave, "Causal Effects of Mental Health Treatment on Education Outcomes for Youth in the Justice System," *Economics of Education Review*, 2016, *54*, 321–339.

Davis, Jonathan M V, Jonathan Guryan, Kelly Hallberg, and Jens Ludwig, "The Economics of Scale-Up," Working Paper 23925, National Bureau of Economic Research 2017.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks, "School Accountability, Postsecondary Attainment and Earnings," *Review of Economics and Statistics*, 2016, *98* (5), 848–862.

Diamond, Rebecca and Petra Persson, "The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests," Working Paper 22207, National Bureau of Economic Research 2016.

Dynarski, Susan, Joshua Hyman, and Diane Whitmore Schanzenbach, "Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion," *Journal of Policy Analysis and Management*, 2013, *32* (4), 692–717.

Figlio, David and Maurice E Lucas, "Do High Grading Standards Affect Student Performance?," *Journal of Public Economics*, 2004, *88* (8), 1815–1834.

Lang, Kevin, "Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member," *Journal of Economic Perspectives*, September 2010, *24* (3), 167–82.

Levitt, Steven D, John A List, and Sally Sadoff, "The Effect of Performance-Based Incentives on Educational Achievement: Evidence From a Randomized Experiment," Working Paper 22107, National Bureau of Economic Research 2016.

Lindo, Jason M, Nicholas J Sanders, and Philip Oreopoulos, "Ability, Gender, and Performance Standards: Evidence from Academic Probation," *American Economic Journal: Applied Economics*, April 2010, *2* (2), 95–117.

Nielsen, Eric, "How Sensitive are Standard Statistics to the Choice of Scale?," Working Paper 2017.

_ , "The Income-Achievement Gap and Adult Outcome Inequality," Working Paper 2019.

Rodríguez-Planas, Núria, "Longer-Term Impacts of Mentoring, Educational Services, and Learning Incentives: Evidence from a Randomized Trial in the United States," *American Economic Journal: Applied Economics*, July 2012, *4* (4), 121–39.

Rubin, Donald B., "Comment on: 'Randomization Analysis of Experimental Data in the Fisher Randomization Test' by D. Basu," *Journal of the American Statistical Association*, 1980, *75*, 591–593.

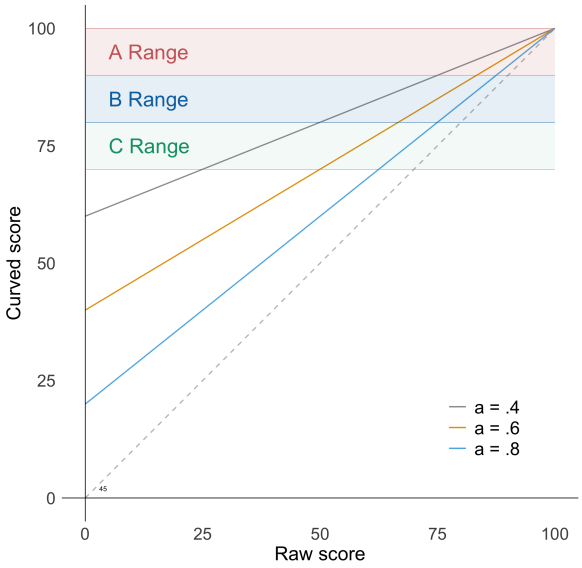
_ , "Which ifs have Causal Answers? Comment on: 'Statistics and Causal Inference' P. Holland," *Journal of the American Statistical Association*, 1986, *81*, 961–962.

Schrödera, Carsten and Shlomo Yitzhakib, "Revisiting the Evidence for Cardinal Treatment of Ordinal Variables," *European Economic Review*, 2017, *92*, 337–358.

Figure 1: Various transformations of raw scores into curved scores

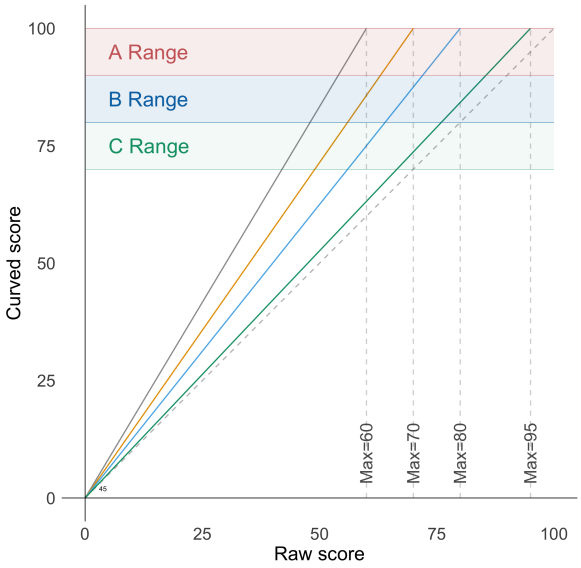
Panel A: Flattening curves

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



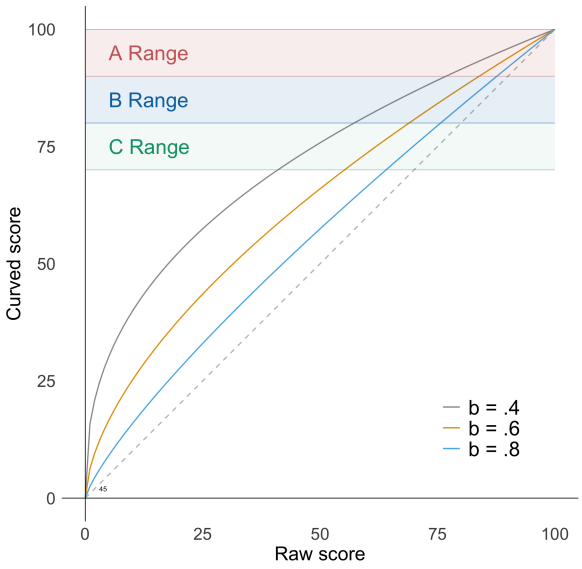
Panel B: High-grade-to-100 curves

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



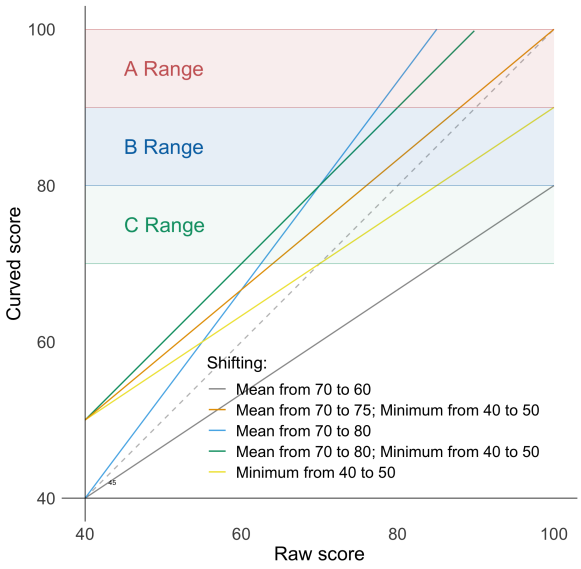
Panel C: Root curves

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



Panel D: Two-point transformations

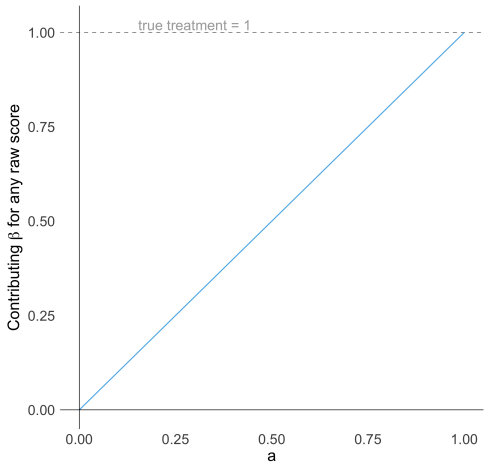
$t(x_{ic}, \min_c^{new}, \mu_c^{new}) = \min_c^{new} + \frac{\mu_c^{new} - \min_c^{new}}{\mu_c^{raw} - \min_c^{raw}}(x_{ic} - \min_c^{raw})$



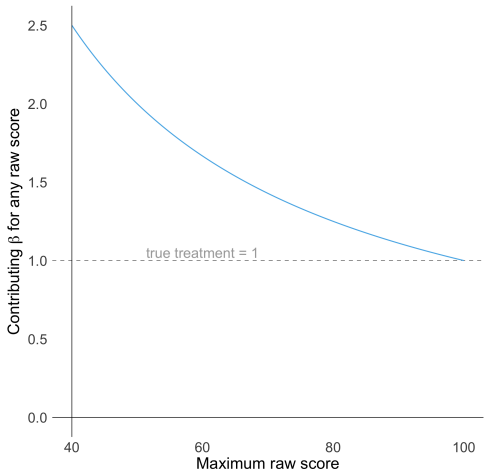
Notes: See Section 2.2 for related discussion. (In Panel D we assume a minimum raw score of 40.)

Figure 2: Partial contributions to identifying treatment in $f(\cdot)$, $h(\cdot)$, and $r(\cdot)$ environments

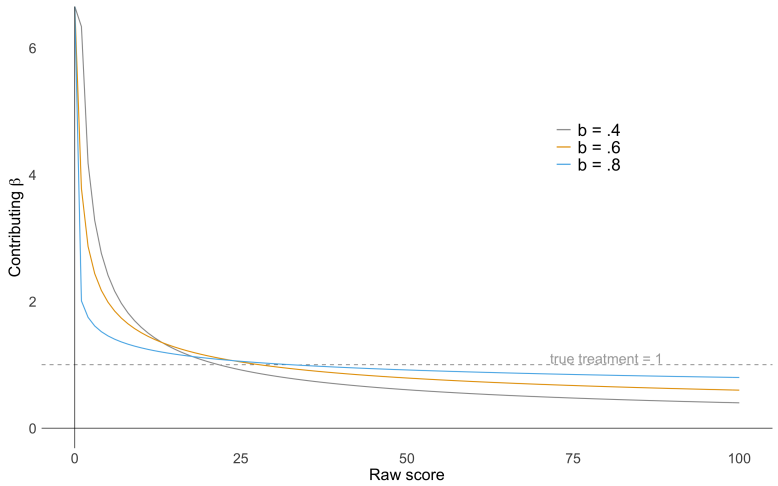
Panel A: Flattening curves attenuate estimates of treatment



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

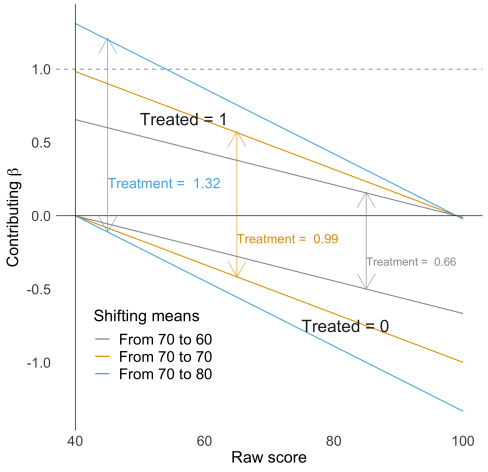


Notes: As in Figure 1, flattening curves are captured in $f(x_{ic}, a) = ax_{ic} + (1-a)100$, high-grade-to-100 curves are captured in $h(x_{ic}) = \frac{100x_{ic}}{\max_c(x_{ic})}$, and root curves are captured in $r(x_{ic}, b) = 100^{1-b}x_{ic}^{b}$.

Figure 3: Partial contributions to identifying treatment for two-point transformations, $t(\cdot)$, and the contamination of controls

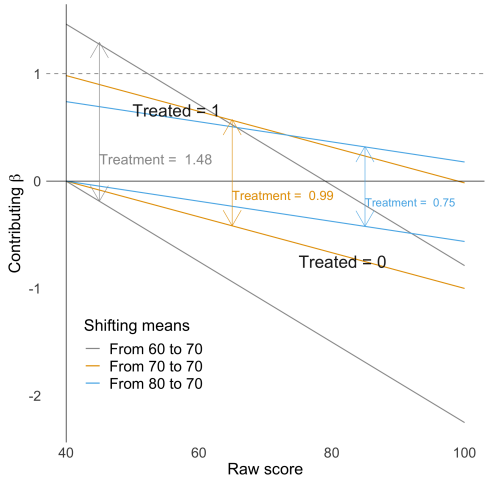
Panel A: Varying instructor preference for mean

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



Panel B: Varying student performance

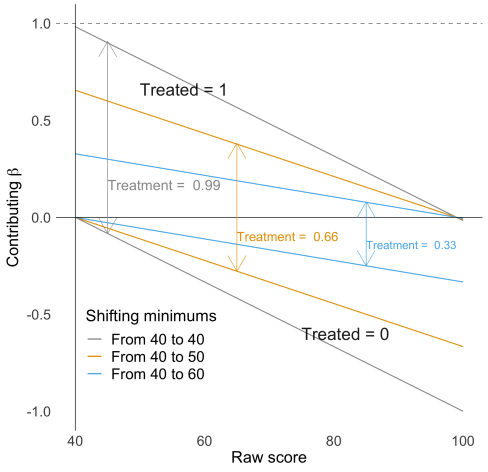
(For a given instructor preference, treatment estimates are larger the lower was student performance)



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

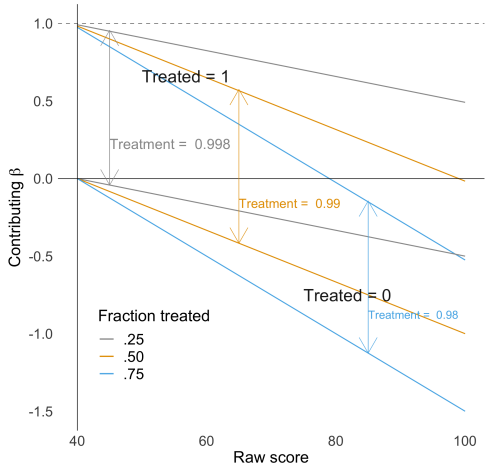
Panel C: Varying instructor preference for minimum

(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

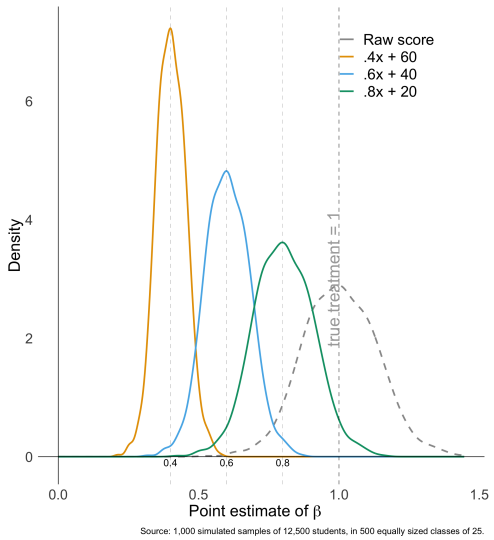
(Treatment estimates are smaller, the larger is the fraction of students treated)



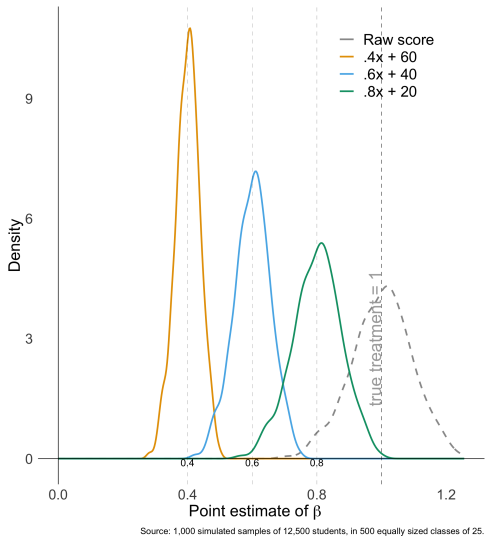
Notes: In all panels, the estimated treatment is the average of treatment- and control-unit contributions, from equations (9) and (9). In reality, this is weighted by the density of students in x_{ic}. As in Figure 1, two-point transformations are captured in $t(x_{ic}, \min_c^{new}, \mu_c^{new}) = \min_c^{new} + \frac{\mu_c^{new} - \min_c^{new}}{\mu_c^{raw} - \min_c^{raw}}(x_{ic} - \min_c^{raw})$.

Figure 4: "Flattening" curves attenuate treatment-effect estimates

Panel A: Without classroom FE



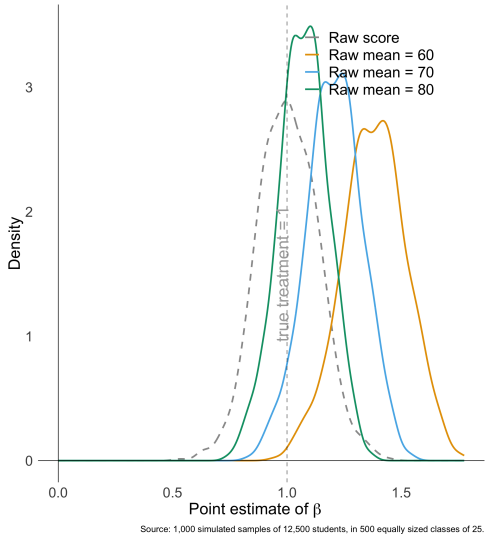
Panel B: With classroom FE



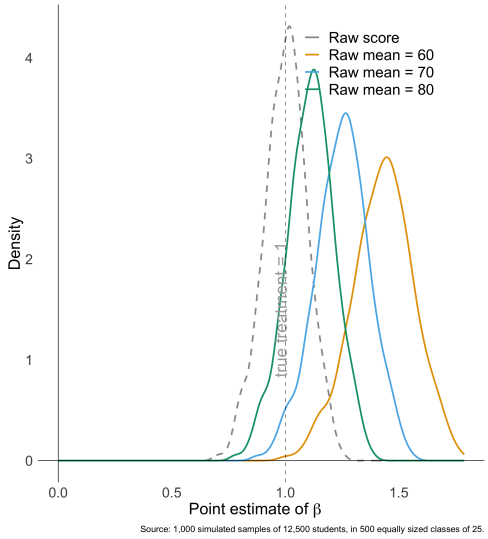
Notes: In all cases, we plot the distribution of estimated treatment effects ($\hat{\beta}$) from models of f-transformed raw scores of individuals i in classrooms c, as in $f(\text{Raw score}_{ic}) = \alpha + \beta \mathbb{1}(\text{Treatment}_i) + \epsilon_{ic}$, where $f(x_{ic}, a) = ax_{ic} + (1-a)100$.

Figure 5: "High-grade-to-100" curves amplify treatment-effect estimates

Panel A: Without classroom FE

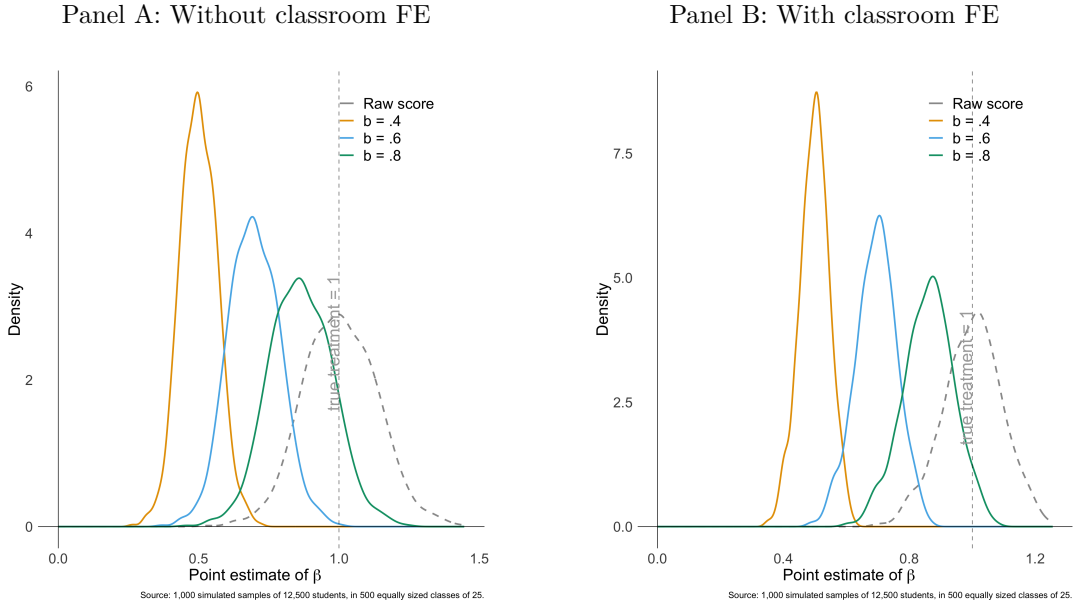


Panel B: With classroom FE



Notes: In all cases, we plot the distribution of estimated treatment effects ($\hat{\beta}$) 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 $h(x_{ic}) = \frac{100 x_{ic}}{\max_c(x_{ic})}$.

Figure 6: In practice, root curves attenuate treatment-effect estimates

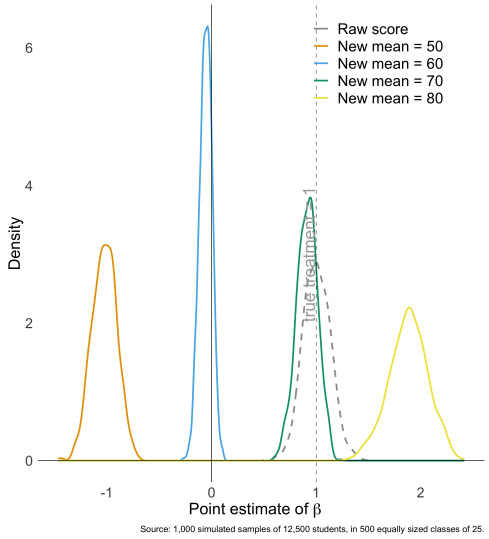


Notes: In all cases, we plot the distribution of estimated treatment effects ($\hat{\beta}$) from models of r-transformed raw scores of individuals i in classrooms c, as in $r(\text{Raw score}_{ic}) = \alpha + \beta \mathbb{1}(\text{Treatment}_i) + \epsilon_{ic}$, where $r(x_{ic}, b) = 100^{1-b} x_{ic}^{b}$.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

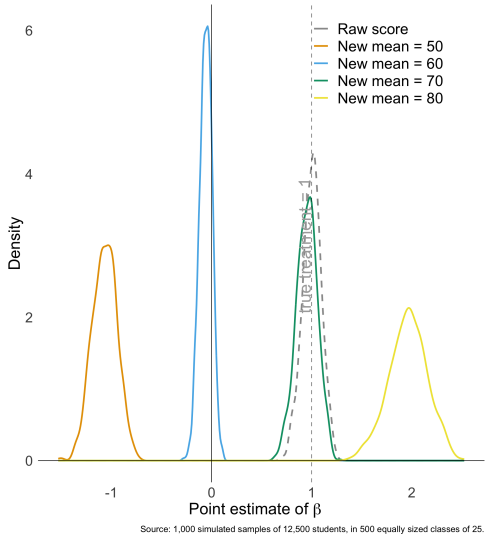
Figure 7: Two-point transformations lead to unsignable biases in treatment-effect estimates

Panel A: Variation in instructor preference around student performance

(i) Without classroom FE

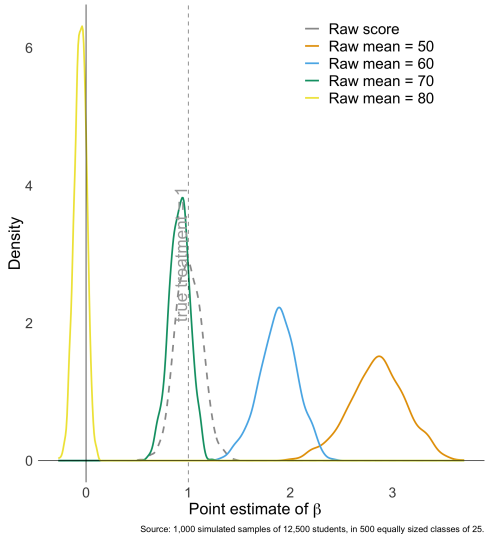


(ii) With classroom FE

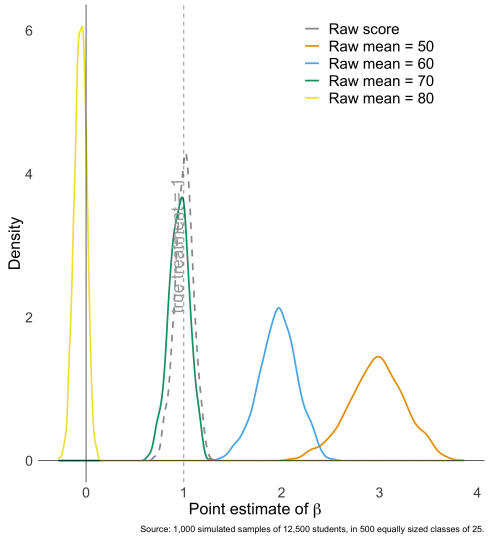


Panel B: Variation in student performance around instructor preference

(i) Without classroom FE

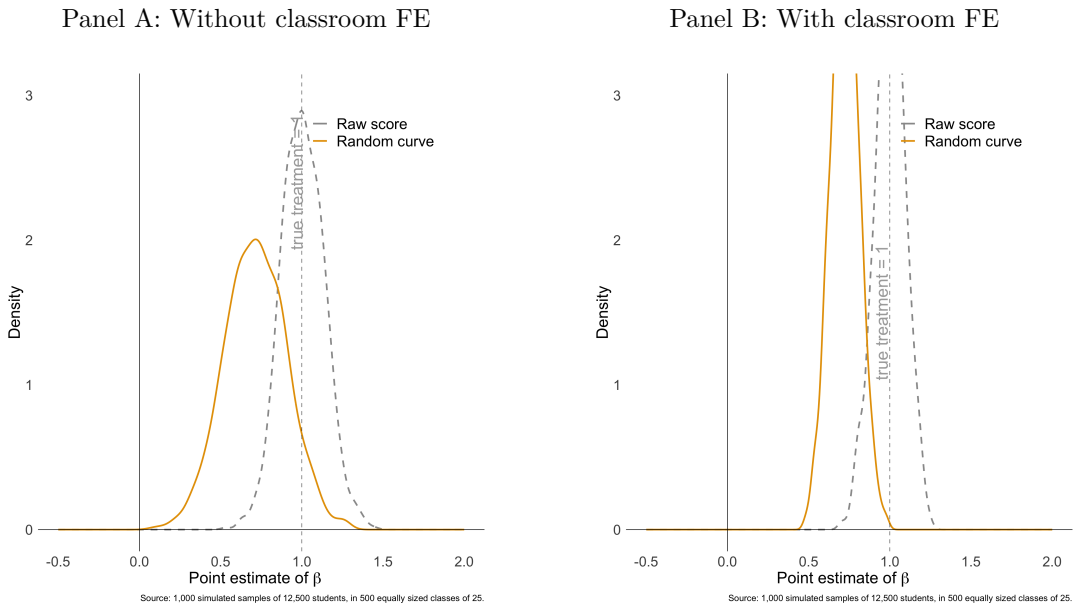


(ii) With classroom FE



Notes: In all cases, we plot the distribution of estimated treatment effects ($\hat{\beta}$) 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 $t(x_{ic}, \min_c^{new}, \mu_c^{new}) = \min_c^{new} + \frac{\mu_c^{new} - \min_c^{new}}{\mu_c^{raw}(x_{ic}) - \min_c^{raw}(x_{ic})}(x_{ic} - \min_c^{raw}(x_{ic}))$.

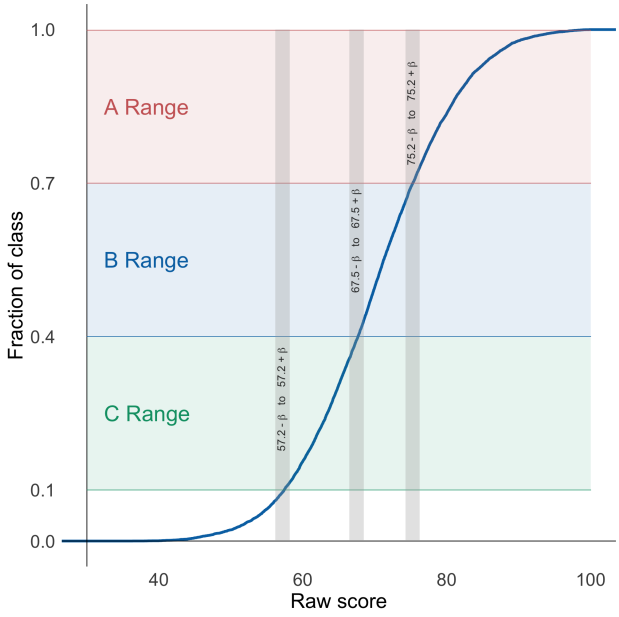
Figure 8: Random curves, and the bias in treatment-effect estimates



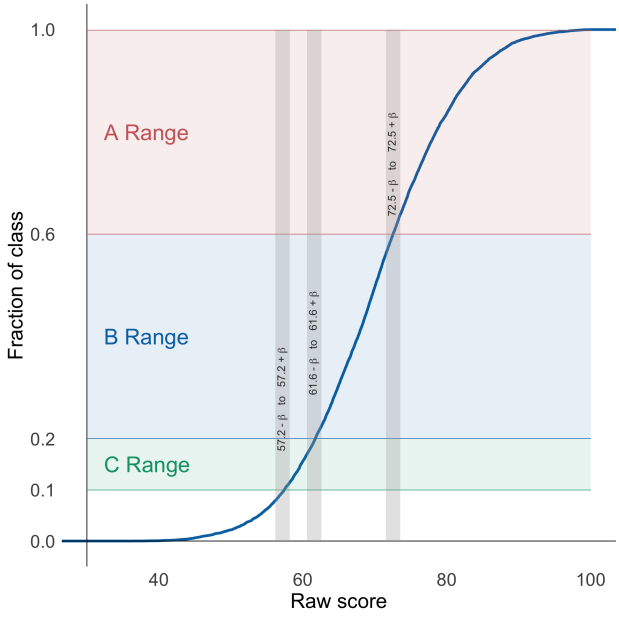
Notes: In all cases, we plot the distribution of estimated treatment effects ($\hat{\beta}$) from models where a curve was assigned randomly to classes (with an equal weight given to "no curve" being assigned).

Figure 9: The bandwidths of domain space that identify treatment effects in letter-grade transformations

Panel A: CDF, GPA-transformation with top 60% splitting As and Bs



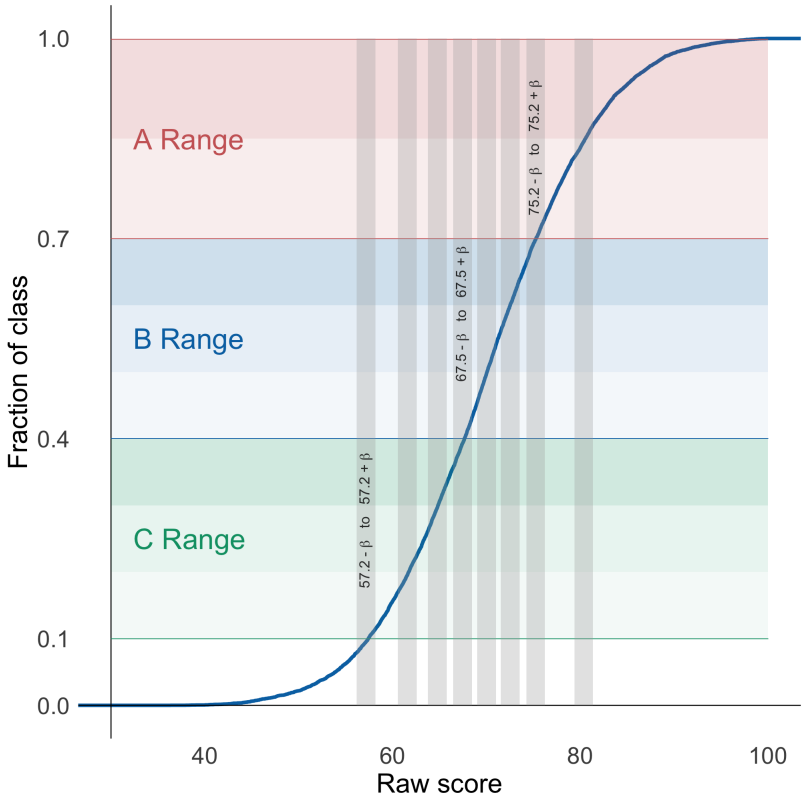
Panel B: CDF, GPA-transformation with top 80% splitting As and Bs



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 a bandwidth $\pm\beta$ (shaded in gray) within which treatment is measurable.

Figure 10: Allowing for plus/minus letter grades increases the domain space that identifies treatment

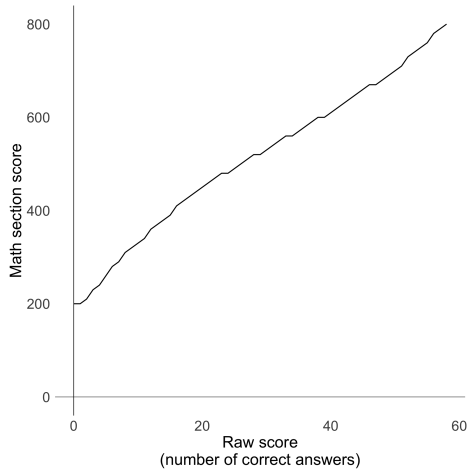
CDF, GPA-transformation with top 60% splitting As and Bs, with plus/minus



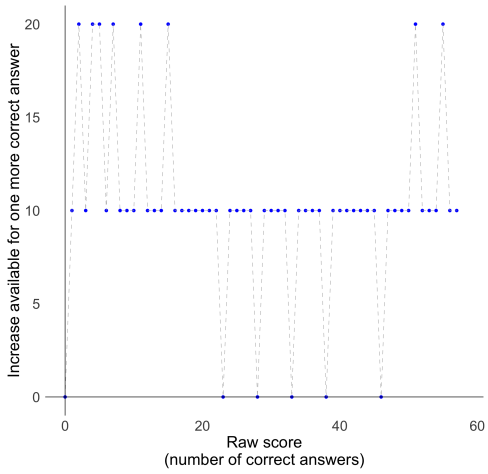
Notes: We plot the CDF of $x_{ic} \sim N(70, 10)$, with letter-grade cutoffs and the students that contribute to treatment (i.e., those within a treatment effect of letter-grade cutoffs) indicated by the shaded regions. The top-30 percent of the class receive As, the next-30 percent receive Bs, and those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

Figure 11: How raw scores on the SAT Practice Test are converted into scaled scores

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

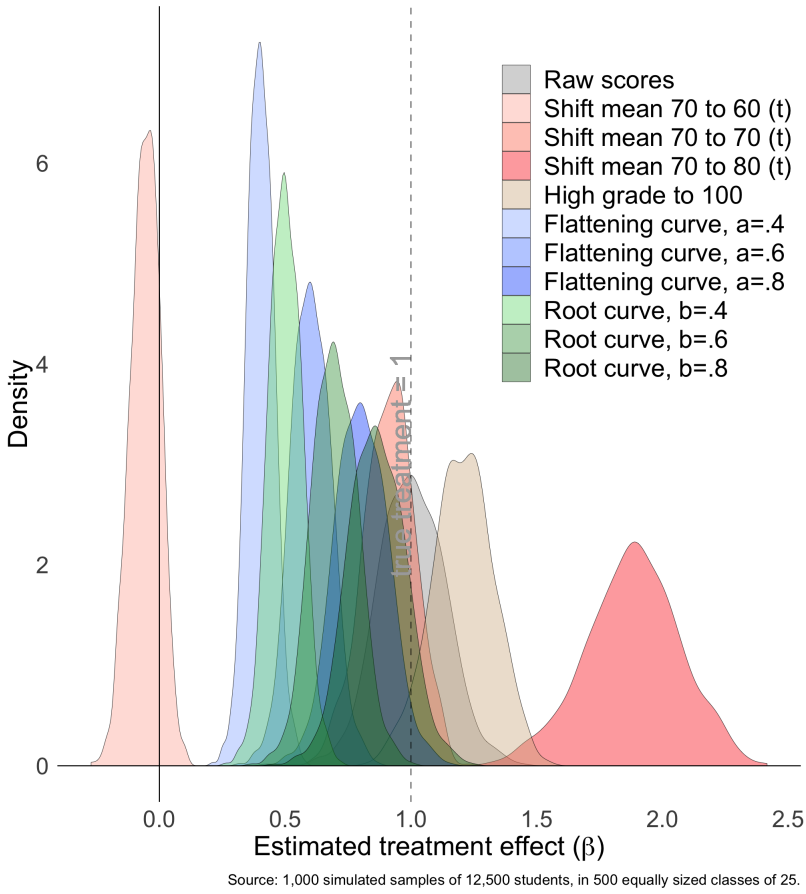


Panel B: The change in scaled SAT math score with a one-question improvement in raw score



Notes: See `https://collegereadiness.collegeboard.org/pdf/scoring-sat-practice-test-1.pdf`, "Raw Score Conversion Table".

Figure 12: 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 2.2.

A Appendix

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 $a\in(0,1]$.[27] 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 2C above—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.)

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

[27] 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 $\hat{\beta}=a\beta$.

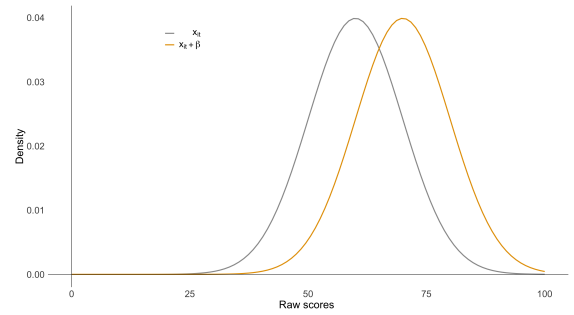
density reflects g, the *change* in the density of the treated group reflects a g-transformed β—it is these perturbations in particular that challenge identification, and that they potentially vary *across* raw scores.

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.

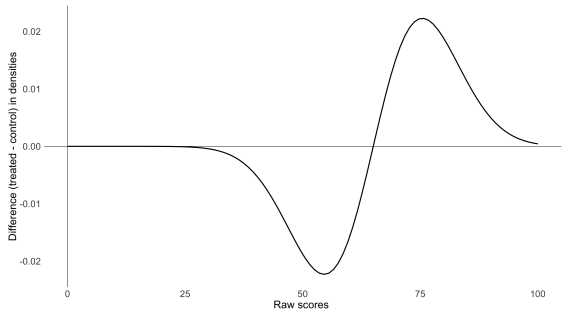
Figure A1: PDFs by treatment classification, and their (treated – control) differences

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

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

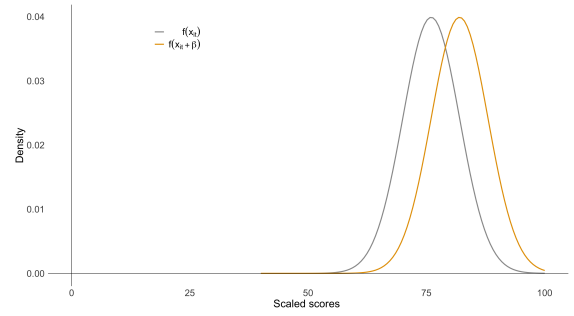


The "$x_{ic} + 1$" less "x_{ic}" difference

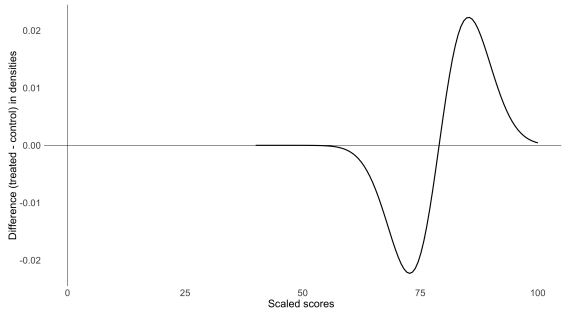


Panel B: A flattening curve ($a = .6$)

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

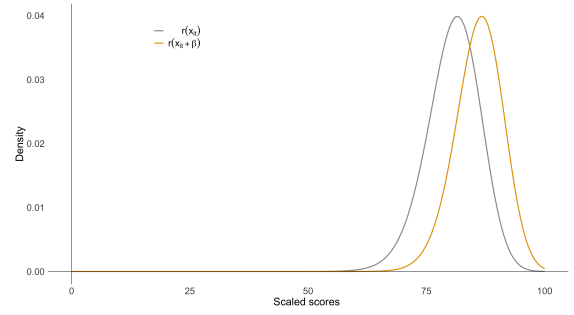


The $f(x_{ic} + 1) - f(x_{ic})$ difference

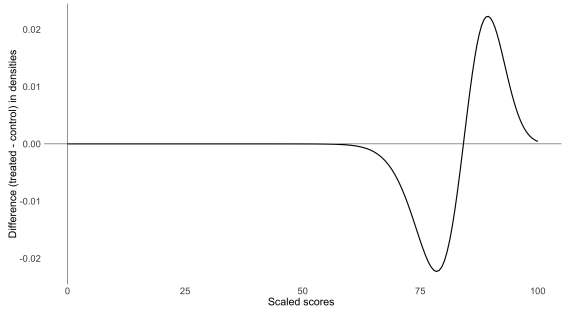


Panel C: A root curve ($b = .4$)

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 2.2 for related discussion of curves.