Where and For Whom Can a Brief, Scalable Mindset Intervention Improve Adolescents' Educational Trajectories?

Authors:

David S. Yeager[1]*, Paul Hanselman[1]*, Gregory M. Walton[3], Robert Crosnoe[1], Chandra Muller[1], Elizabeth Tipton[7], Barbara Schneider[5], Christopher Hulleman[6], Cintia Hinojosa[7], Dave Paunesku[8], Carissa Romero[9], Kate Flint[10], Alice Roberts[10], Jill Trott[10], Ronaldo Iachan[10], Jenny Buontempo[1], Sophia Yang Hooper[1], Jared Murray[1], Carlos Carvalho[1], Richard Hahn[11], Ronald Ferguson[12], Angela L. Duckworth[13], Carol S. Dweck[3].

* Address correspondence to David S. Yeager, dyeager@utexas.edu, or Paul Hanselman, paul.hanselman@uci.edu.

Affiliations:

[1] University of Texas at Austin, [2] University of California, Irvine, [3] Stanford University, [4] Columbia University Teacher's College, [5] Michigan State University, [6] University of Virginia, [7] University of Chicago Booth School of Business, [8] Project for Education Research that Scales, [9] Paradigm Strategy Inc., [10] ICF International, [11] Arizona State University, [12] Harvard University, [13] University of Pennsylvania.

Author Note:

This manuscript uses data from the *National Study of Learning Mindsets* (PI: D. Yeager; Co-Is: R. Crosnoe, C. Dweck, C. Muller, B. Schneider, & G. Walton). The program and surveys were administered using systems and processes developed by the Project for Education Research That Scales (PERTS, www.perts.net). Data collection was carried out by ICF International (Project directors: Kate Flint and Alice Roberts). Planning meetings were hosted by the Mindset Scholars Network at the Center for Advanced Study in the Behavioral Sciences (CASBS), and the study received assistance from M. Levi, M. Shankar, T. Brock, C. Macrander, T. Wilson, E. Konar, E. Horng, H. Bloom, and M. Weiss. Funding was provided by the Raikes Foundation, the William T. Grant Foundation, the Spencer Foundation, the Bezos Family Foundation, the Character Lab, the Houston Endowment, and the President and Dean of Humanities and Social Sciences at Stanford University. Preparation of the manuscript was supported by National Institute of Child Health and Human Development (Grant No. 10.13039/100000071 R01HD084772-01), P2C-HD042849 (to the Population Research Center at The University of Texas at Austin). This study was a part of the "pre-registration challenge" (https://osf.io/r82dw/).

Pre-print Disclaimer and Media Embargo

This pre-print is being shared to promote timely dissemination of research findings and to solicit feedback from the scholarly community about this research. The findings and manuscript will almost certainly change before publication. The investigative team welcomes suggestions for clarifying or improving the research or the documentation of the findings. The findings are embargoed from discussion with journalists or other members of the media until publication in a journal.

Abstract

A pressing global challenge is to identify interventions that improve adolescents' developmental trajectories. But no intervention will work for all young people everywhere. It is critical then to study the *heterogeneity* of intervention effects in a way that is generalizable and replicable. In the *National Study of Learning Mindsets* (*N* = 12,542) researchers randomly assigned 9th grade students in a representative sample of 65 U.S. public schools to a growth mindset intervention, which conveyed that intellectual abilities are not fixed but can be developed. The brief (~50-minute), scalable and low-cost intervention reduced by 3 percentage points the rate at which adolescents in the U.S. were off-track for graduation at the end of the year, corresponding to an estimated benefit of approximately 100,000 adolescents per year. This is the first experimental evidence that an intervention can improve adolescents' educational trajectories in a national probability sample. Yet the growth mindset intervention effect was heterogeneous in predictable ways. Some sub-groups of students (lower-achievers) and schools (those with supportive behavioral norms) showed appreciably larger increases in grades. Heterogeneity findings were reproduced in a conservative Bayesian "sum-of-regression-trees" analysis, which guards against false discoveries. These findings lead to novel hypotheses about ways to enhance intervention effects and target public policies. Findings also illustrate the power of even slight adjustments in motivational priorities to create enduring change during adolescence.

Where and For Whom Can a Brief, Scalable Mindset Intervention Improve Adolescents' Educational Trajectories?

For many young people, adolescence is a fork in the road. While some adolescents continue to acquire the foundational skills that will help them thrive in adulthood, others do not [1–3]. Adolescents' trajectories can be redirected, however. Neurodevelopmental and hormonal changes during pubertal maturation create a readiness to pivot rapidly in response to even small changes in motivational priorities, initiating self-reinforcing cycles that alter long-term trajectories [3–6]. A global challenge of great significance, then, is to discover, evaluate, and scale effective interventions to redirect adolescent motivation and behavior in positive directions [1,3].

Unfortunately, interventions to influence adolescents have, on average, produced very weak or null effects, even though interventions have implemented substantial changes to school staffing and curriculum and required many months of classroom time [6]. In part this is because many interventions are simply a poor match with adolescents' developmental values and needs [6–8]. But even interventions that seem developmentally appropriate can show small effects. This has led some observers to ask whether adolescent behavior change at a national or global scale—change that is internalized and lasting—is even an achievable goal for science [7].

We argue that scientific progress toward the discovery of scalable adolescent behavior-change interventions has been limited in part by the search for *average* effects, with insufficient scientific resources allocated to understanding *heterogeneity* of effects across individuals and contexts. An intervention to redirect an adolescent's motivational priorities should improve developmental trajectories only when trajectories are at risk, and only when the context allows the new behavior to become self-reinforcing, but not otherwise [9–11]. Researchers only interested in average effects might therefore prematurely abandon interventions with small average effects that are concentrated among individuals at greatest risk [12]. Furthermore, if scientists could identify factors that enhance or diminish intervention effects, this could yield insights into intervention mechanisms, inform precision public policy, and feed into innovations to make interventions more effective [13,14].

Three barriers have often stood in the way of studying treatment effect heterogeneity: non-experimental methods, non-generalizable samples, and analysis methods that could produce chance findings [15,16]. We overcome these barriers by using the preferred methods for internal validity (double-blind *random assignment* to treatment or control [17,18]), for generalizing to diverse groups (*random sampling* of sites from a national sampling frame [19]), and for increasing the likelihood that findings are reproducible (data collection by independent researchers, developing a limited number of moderation hypotheses and stating them prior to analysis [20], and analysis of a blinded dataset by independent statisticians using conservative models designed to prevent false discoveries [15]). This study is the first to have all three of these features.

We carried out our investigation in the context of a social problem with immense importance for the global economy: academic success during high school. Advanced skills and educational attainment are now more critical than ever for determining an individual's future earnings, health, and life expectancy [21,22]. Yet, in the U.S. alone, one in five of the roughly 3.5 million adolescents who enter 9th grade each year will fail to graduate on time [23,24]. Even many of those who do graduate are not on track for acquiring the cognitive and non-cognitive skills they need for the global labor market [25,26]. Unfortunately, the last few decades have seen little improvement in high school success rates [24,25]. As a result, U.S. public high schools have not yet been able to stem the large and growing tide of inequalities in educational attainment.

The substantive question we ask is: *for whom and under what conditions*[27,28] will a brief, online, low-cost intervention improve the achievement of underperforming 9th grade adolescents in U.S. public schools? This question is critical to answer because a misstep in the first year of high school, such as a failed math or science class, could put a young person on a different educational trajectory, due to the highly structured and cumulative nature of educational systems [29,30].

The intervention we evaluate is a *growth mindset of intelligence* intervention [31–34]. The growth mindset intervention builds on a basic insight from a long tradition of psychological research, namely that people have latitude in how they view themselves and their circumstances, and these views can affect motivation and behavior [35–38]. Adolescents given the same learning opportunities may thus differ in how they make sense of and engage with them, and therefore how well they fare in school.

Specifically, adolescents may hold what has been called a *fixed mindset*, which refers to the belief that intelligence is fixed and cannot change [31]. Looking through this fixed mindset lens, adolescents are more likely to view failures or mistakes as signs that they lack ability, and may come to expect that they cannot succeed [32,39]. A fixed mindset therefore poses a barrier to motivation among adolescents undergoing challenging learning experiences.

Importantly, however, *growth mindset* interventions can realign motivation by redirecting adolescents away from a fixed mindset. The growth mindset intervention conveys the idea that intellectual abilities can be developed through sustained effort, good strategies and appropriate help and support from others. The growth mindset intervention conveys this in part through a memorable metaphor: the brain is like a muscle that grows stronger and smarter when it undergoes rigorous learning experiences [33]. Adolescents hear about the metaphor, reflect on ways to strengthen their brains by persisting on learning challenges, and they write about how they could use a stronger brain to achieve meaningful goals [40]. Adolescents in past studies who received a growth mindset intervention were more likely than a control group to view present difficulties as signs that they were growing their skills, causing them to stay more motivated in the face of challenges and demonstrate greater engagement [34,40]. Even relatively brief exposures to growth mindset intervention materials have improved the grades of adolescents who were at risk for poor academic trajectories (e.g. lower-achieving adolescents) relative to counterparts in the control group [34,40] (also see [41]).

The growth mindset of intelligence intervention was a good case study for the first examination of heterogeneous effects of an adolescent behavior-change intervention in a nationally representative sample, for several reasons. First, the growth mindset intervention grows out of decades of basic laboratory science [42], so the psychological mechanisms are well-documented [39]. Second, the growth mindset intervention has been subjected to multiple replications (including a previous large-sample, pre-registered replication study [40]), so average effects to date are not likely to be false positive. Third, fixed mindset beliefs are highly prevalent in society, so a growth mindset intervention is likely to be useful in schools across the U.S. Fourth, the growth mindset intervention can be delivered briefly and at very low per-person cost via a self-administered computerized module. This made it ideal for random assignment at the person level so we could study heterogeneity across groups of students and schools. Fifth, there is nevertheless reason to expect that intervention effects will be heterogeneous.

At the individual level, we expected heterogeneity due to students' prior performance. Lower-achieving students were most likely to be on a negative trajectory that could be redirected by a growth mindset, and this is what past research found. Higher-achieving adolescents' grades were unlikely to be increased because of a range restriction on their grades and because they may already have habits (e.g. turning in work on time) or beliefs (e.g., a growth mindset) that lead to higher grades [40].

At the context level, we generated moderation predictions by integrating theories of psychological interventions with theories in the sociology of education. The first moderator we test is *school achievement level*—the school's typical test scores, proportion of adolescents who are college ready or who earn college credit, and so on—which can reflect the learning opportunities available to adolescents, not just in high school but in their lives overall [43]. When learning opportunities are not available, a motivational intervention should not be of much benefit [44,45]. But when schools already provide abundant resources to prevent students from failure, a motivational intervention may not be necessary. Therefore we test whether school achievement moderates the mindset treatment in a positive or negative direction.

The second moderator we test is *school growth-mindset behavioral norms*—whether the typical or desirable behavior in the school is consistent with a growth mindset, meaning that it supports greater challenge-seeking or academic effort. Research in neuroscience[46,47], sociology [30], cultural anthropology[48], social psychology[49] and developmental psychology[50,51] has shown that adolescence in particular is a stage during which personal beliefs and behavior can be powerfully shaped by social norms and the desire to conform to them. This fact leads to two competing predictions. Either the growth mindset intervention will be more effective in schools with supportive norms—because the peer climate sustains the initial treatment effect—or the intervention will be less effective in schools with supportive norms—because the supportive peer climate would be "treating" the control group already.

In sum, we use a representative sample to identify the student sub-groups and school contexts where a seemingly-remarkable result is most likely to appear—namely that a brief, online, scalable, and low-cost intervention could redirect adolescents' educational trajectories. In doing so we provide a "case study" for how to attend to treatment effect heterogeneity in the design and execution of an intervention at scale.

Methods

Data come from the *National Study of Learning Mindsets*, which included $N = 12{,}541$ 9th grade adolescents whose administrative data could be obtained from a national probability sample of 65 regular public schools in the U.S. To achieve arms-length independence, a research firm not involved in designing the materials or study hypotheses drew the sample, recruited schools, facilitated treatment delivery, obtained administrative data, and cleaned and merged data. Data were processed blind to treatment status.

A random sample of schools, rather than a convenience sample, meant that it represented the full array of the U.S. public educational contexts [52]. Even a well-constructed probability sample, however, would not yield sufficient numbers of rare types of schools, and poor coverage of rare groups can undermine inferences about moderators, especially when there is confounding in the moderators [15,52]. Therefore, the study constructed a sampling plan using hypotheses about school-level moderators to inform probabilities of selection [52,53]. Sampling probabilities were stratified by the school's achievement level and a potential confound for school achievement: the racial minority composition of the school (in the sampling frame the percent of adolescents in the school who identify as black/African-American or Hispanic/Latino/a was highly correlated with achievement in the U.S., $r = -.66$, $p < .001$).

A total of 139 schools were randomly selected from a sampling frame of over 12,000 regular U.S. public high schools and invited to administer the intervention and provide student records; 65 schools agreed, participated, and provided student records. The schools that agreed to participate were not different from the sampling frame in terms of school achievement level, suggesting that non-response was not biased in terms of a key moderator. Yet schools that agreed were more likely to be rural and were on average smaller than those in the sampling frame, and so survey weights adjusted for these non-response biases.

Within schools, the median student response rate was 98%, which is very high and means that lower-achieving students (who are often less likely to complete surveys) were sufficiently represented. Participants were diverse: 11% Black/African-American, 4% Asian-American, 24% Hispanic/Latino/a, 43% White, and 18% another race or ethnicity; 29% reported that their mother had a bachelor's degree or higher.

We revised past growth mindset interventions so that the materials focused on the perspectives, concerns, and reading levels of 9th grade students in the U.S via a research and development process carried out over several years with thousands of students (described in a paper about this process [40]). The active control condition, focusing on brain localization, was similar in form to the growth mindset intervention and was rated by adolescents as both interesting and engaging,[40] but it did not address the key concept of growth mindset.

The intervention (or control) was delivered as early in the school year as possible (80% of students received it in the fall semester) via two self-administered online sessions that lasted approximately 25 minutes each and occurred roughly 20 days apart during regular school hours. The computer software randomly assigned adolescents to intervention or control materials, which adolescents completed along with various survey questions. Students, teachers, and researchers were kept blind to condition assignment. To students and teachers, the intervention experience was simply a “survey” (to prevent expectancy effects).

The primary outcome variable was adolescents’ post-intervention grade point average (GPA) in core classes (English, math, science, and social studies), obtained from administrative data sources. GPA in 9th grade is a much better predictor than standardized achievement test scores of adolescents’ educational trajectories—their high school graduation, college enrollment, college retention, and wages in adulthood [54]. Like these long-term outcomes, a higher 9th grade GPA results in part from an adolescent’s sustained motivation. Therefore 9th grade GPA is a theoretically- and practically-relevant outcome.

Data were analyzed following a pre-registered analysis plan (the so-called “pre-registration challenge,” osf.io/afmb6/) that was developed by an interdisciplinary team, including one external evaluator. All analyses were “intent to treat” (ITT); data were analyzed as long as students saw the first page of the randomized materials. Average treatment effects were estimated in a cluster-robust fixed-effects linear regression model that used weights provided by the research firm to make coefficients generalizable (the same substantive conclusions emerged without weights; see the online supplement for alternative model specifications). Multilevel models were used to understand cross-site variability. Equations are presented in the pre-analysis plan.

Results

Based on administrative records, 9th grade adolescents assigned to the growth mindset intervention, as compared to the control activity, earned slightly higher GPAs in core classes at the end of 9th grade. On a 4-point grade metric ("A" = 4.0, "B" = 3.0, etc.), the average treatment effect was 0.03 grade points, *SE* = .01, *N* = 12,542 students, *k* = 65 schools, *t* = 3.09, *P* = .003.

The intervention also decreased adolescents' overall *poor performance rates* (the percent of adolescents earning a D or F average, i.e., GPA below a 2.0), by 3 percentage points, *SE* = 0.01, *t* = -3.06, *P* = 0.004, replicating past research [40]. The poor performance threshold is critical because 9th grade adolescents with GPAs in the D or F range rarely graduate on time, making the poor performance rate an excellent indicator of whether adolescents were put on a new trajectory [54,55].

The national probability sample generalizes to roughly 3.5 million 9th grade adolescents per year, and so the model estimates that the brief, online growth mindset intervention, if given to all public schools in the U.S., would prevent an expected 98,000 adolescents each year from finishing 9th grade with D or F GPAs (2.8% × 3.5 million), keeping them on track to graduate (assuming logistical barriers could be overcome).

We next examined the hypothesis that effects would be larger for the sub-group of *lower-achieving* adolescents, defined as adolescents who were earning GPAs at or below the school-specific median in the term prior to random assignment (osf.io/afmb6/). As expected, among lower-achieving adolescents, the estimated growth mindset impact on GPA was a positive and significant .08 grade points, *SE* = .03, *N* = 6,219, *k* = 65, *t* = 3.14, *P* = .003. The impact on the poor performance rate was a negative and significant 6 percentage points, t = -3.44, P = 0.001, from a base rate of 46% among control adolescents. For both GPA and poor-performance rates, the Intervention × Lower-achiever interactions were significant, Ps < .001, replicating prior research. As predicted, among higher-achieving adolescents there were no significant effects on core course GPA or poor-performance rates (see online supplement).[1]

How does the average effect of a psychological intervention compare to effect sizes of school reforms typically seen in the literature? Empirical benchmarks can provide an anchor for what to expect [57]. A search for reforms listed on the U.S. federal government's What Works Clearinghouse (a database of rigorously-evaluated programs) found that nearly all past high school programs—tutoring programs, school redesigns, and more—showed no significant benefits on objective outcomes. A small handful of best-in-class program evaluations documented modest benefits. One large, prominent study showed effects for lower-achieving students of 0.06 grade points at a cost of $2,000 per adolescent [58]. Another reduced 9th grade poor performance rates by 3.7 percentage points overall, at a cost of $4,000 to $20,000 per adolescent [59]. Growth mindset average effects are therefore about as effective on average but far less costly.

Our main interest, however, was not in *average* effects but in *heterogeneity* of effects. In a mixed effects model with a fixed intercept and a random slope [12], we estimated a significant standard deviation of treatment impact among lower-achieving students across schools (i.e.

[1] A mediation analysis was consistent with theory. The growth mindset intervention reduced the prevalence of fixed mindset beliefs reported at the end of the second treatment session (i.e. how much adolescents agreed that "You have a certain amount of intelligence, and you really can't do much to change it", 1 = *Strongly disagree,* 6 = *Strongly agree*), B = -.42, SE = .02, t = 19.15, P < .001. Post-treatment reports of fixed mindsets predicted post-treatment GPAs, r = -.22, P < .001, replicating past research [32,40]. Finally, the mediation analysis[29] found a significant indirect effect of the intervention on low-achieving adolescents' GPAs via reports of growth mindset, B = .012 [95% CI .004, .019], P < .001 (this model addressed sequential ignorability[56] by controlling for pre-treatment mindset).

heterogeneous treatment effects), unstandardized $\tau = 0.08$, *Q*-statistic $P = .02$, permutation test $P = .05$ [12]. This cross-site heterogeneity in the lower-achieving student effect was not explained by differences in how much adolescents accepted the intervention message. That is, there was no heterogeneity in intervention effects on self-reported fixed mindsets across schools, $\tau = 0.16$, *Q*-statistic $P = .44$, permutation test $P = .25$. In all schools, it seems, adolescents successfully received the initial mindset "push." Schools were heterogeneous, however, in whether that push started a self-sustaining process that resulted in higher GPAs.

Was cross-site heterogeneity in growth mindset effects on GPA explained by *school achievement level* and *school mindset norms*? In answering this, the mixed effects model controlled for the racial/ethnic composition of the school, and its interaction with treatment status, to account for confounding between school race/ethnicity composition and the focal moderators. Interactions with race/ethnicity were never significant.

Treatment effects among lower-achieving students were larger in medium and lower-achieving schools, relative to the highest-achieving schools, Intervention × School achievement level interaction predicting 9th grade GPA, $B = -.07$, $SE = .03$, $z = -2.25$, $N = 6{,}219$, $k = 65$, $P = .024$, in the linear mixed effects model. Post-hoc tests in the mixed effects model found the contrast between the bottom 25% of schools and the middle 50% of schools was not significant, $z = -0.36$, $N = 6{,}219$, $k = 65$, $P = .721$, but the middle 50% was different from the top 25%, $z = -2.04$, $N = 6{,}219$, $k = 65$, $P = .041$. Intervention effects on GPA were null in the highest-achieving 25% of schools and significant in the lower 75%. See Figure 1.

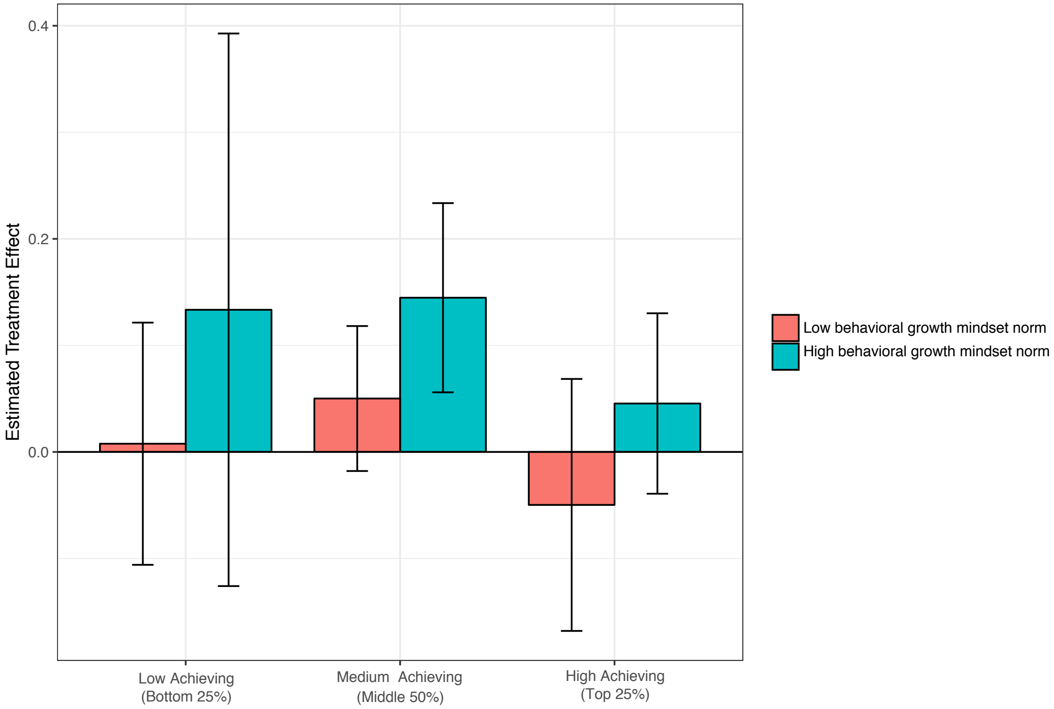
At first glance, negative interaction effect could be surprising. It might seem as though lower-achieving schools should not show benefits, because students there face many other challenges beyond mindset that undermine performance. The finding that *any* improvement was possible in lower-achieving schools is evidence that, sometimes, students in those contexts had more latent potential to succeed than previously thought.

Why were there null effects on GPA in the highest-achieving schools? As noted, it was not because the treatment failed to change mindsets in those schools. Instead, one explanation is that in the highest-achieving schools, lower-achieving students were simply less likely to be on a trajectory for failure. National data[60] shows that students in the top 25% of high schools were already nearly a full grade level ahead of their peers nationally (+0.91 grade levels) before they started high school. Moreover, using our data's control group only, in the highest-achieving schools the 9th grade poor performance rate (i.e. GPA in the D or F range) for lower-achieving students was 16%, compared to 34% in the medium and low-achieving schools. Thus, in the highest-achieving schools, resources to prevent failure may have been plentiful.

A second explanation is that adolescents in higher-achieving schools showed their growth mindsets in a different way than their GPAs. For a subset of schools ($k = 46$), we found out whether adolescents took Algebra II or an even higher class in 10th grade—the dividing line for staying on track for Calculus by senior year. Taking a more challenging math class—rather than dropping down to classes where they can get an easy "A"—is an indicator of growth mindset-relevant behavior [61]. In the highest-achieving schools (the top 25%), the growth mindset treatment increased adolescents' likelihood of taking Algebra II or a higher math course in 10th grade by 4 percentage points, $SE = .017$, $t = 2.06$, $P = .039$ (this analysis included both lower- and higher-achieving adolescents, who did not differ for this outcome).

The second planned moderation analysis concerned the school's mindset-related norm. The school norm was assessed via student behavior with a validated measure called the *make-a-math-worksheet* task [40], in which students created worksheets (to complete later) with challenging problems (from which they would learn a lot) or easy problems (that they would get right but not learn much from). For each school we calculated the average number of challenging math problems that adolescents in the control condition assigned themselves to work on. We presumed that challenge-seeking *behaviors* would be more visible to peers (and therefore more likely to induce conformity, i.e. moderate treatment effects) than *beliefs* (students' self-reported mindsets) [62]. Models tested for moderation by both behavioral norms (the main measure) and self-reported mindsets (the secondary measure).

Figure 1. Conditional average treatment effects and standard errors for the growth mindset intervention on 9th grade GPA in core courses, by school achievement and norms, among lower-achieving adolescents. Note: Unstandardized effect sizes. Effect sizes estimated in a linear mixed effects model without survey weights. See online supplemental material for model specifications. Low achievement = Bottom 25%; Medium Achievement = Middle 50%; High Achievement = top 25%; Low behavioral growth mindset norm = Below-median levels of challenge-seeking; High behavioral growth mindset norm = Above-median levels of challenge-seeking.



Analyses uncovered a positive and significant Intervention × Behavioral Growth Mindset Norms interaction when predicting lower-achieving adolescents' GPAs, $B = .127$, $SE = .054$, $z = 2.33$, $N = 6{,}219$, $k = 65$, $P = .02$. The growth mindset intervention produced a greater increase in end-of-year GPA when norms (as measured by peer behaviors on the make-a-math-worksheet task) were in line with the intervention's message. See Figure 1. As we suspected, there was no Intervention × Self-Reported Mindset interaction, P>.80, confirming that it was the average behavior of peers in the school—not the average of peers' beliefs—that moderated treatment effects [63].

In some respects, this moderation finding is surprising as well because the opposite pattern would have been reasonable. One might have expected that the growth mindset intervention would be most effective in schools with unsupportive norms, because that is where the intervention is most needed. This is not the finding that emerged, however.

Instead it seems that, like a stone pushed down a hill that gains momentum from each revolution, a brief growth mindset intervention delivered in a school where students tend to value hard work accumulated sustained effects on GPA. But like a stone on a flat surface given a hefty shove but eventually halted by friction, the benefits of a growth mindset intervention may fade away when delivered in a school where students pay a social price for working hard.

This finding leads to the exciting possibility that systematic interventions to improve the normative mindset *environment*. Changing the environment, above and beyond students' own individual mindsets, might yield transformative effects on developmental trajectories [64,65].

Finally, independent statisticians reproduced the key moderation findings by estimating a hierarchical, nonlinear Bayesian model [15] using a blinded dataset that masked the identities of the variables, to further reduce the possibility of chance findings [20]. The Bayesian model is useful because it a) uses machine learning tools to discover non-linearities and interactions even if they are not pre-specified, while delivering a better fit to the observed data, b) provides meaningful estimates of uncertainty for quantities of scientific interest, and c) uses regularization to guard against false discoveries and reduce the statistical error in effect size estimates [15] (for an applied example, see [66]). The Bayesian model reproduced the pre-registered linear mixed effects results: it found that school achievement and growth mindset norms moderated treatment effects, with strongest treatment effects for middle and lower-achieving schools with higher growth-mindset norms. The model did *not* find that potential confounds or alternative moderators—racial composition or self-reported mindset norms—were moderators, confirming the linear mixed effect results. Figures for the Bayesian model appear in the online supplemental material.

Discussion

What have we learned from analysis of data from the *National Study of Learning Mindsets* so far? As expected, an online growth-mindset intervention lasting approximately 50 minutes increased adolescents' motivation to learn, as assessed by GPA taken from administrative records at the end of the year, in the gateway period of 9th grade, and with a national probability sample of high schools. Also as expected, average effects were small because many students are already doing well, do not have motivational issues, or are not in environments that encourage or support growth-mindset behaviors. When we take account of such factors, more noteworthy effects emerge. The improvements in the gateway outcome of 9th grade GPA were concentrated among adolescents who are at significant risk for compromised well-being and economic welfare: those with lower levels of prior achievement attending relatively lower-achieving schools. The finding that an intervention can redirect this adolescent outcome in this sub-group, in under an hour, without training of teachers, and at scale (i.e. in a random sample of nation's schools), represents a significant advance.

Even so, the present results raise questions that it will be important to examine in the future. One exciting area to pursue will involve from data collected on math teachers in the schools included in the *National Study of Learning Mindsets*. Teachers completed measures relevant both to the learning opportunities they provide students, and the growth mindset climate they create, and so it will be possible to conduct finer-grained analyses than the present study's school-level moderation analyses. We will make the present dataset open to researchers, and so we expect many insights to come from secondary data analysis.

More generally, the present study is to present a case for how to plan in advance to learn from heterogeneous treatment effects. This case may prove useful in the coming years. There has recently been an explosion in interest in light-touch interventions (i.e. "nudges" [67] and "wise interventions" [36]), as researchers and governments have exponentially grown the number and scope of field experiments testing solutions to social problems. Even with enormous amounts of data, however, investigators have rarely understood heterogeneity. Those that have examined heterogeneity systematically have often discovered that non-probability samples can yield badly biased and misleading conclusions about effect sizes even with massive samples [68]. We hope that our systematic, *a priori* approach to probing heterogeneity of effects with representative samples will serve as a useful model for accumulating knowledge of the conditions under which behavioral science interventions improve human welfare in general and adolescent development in particular.

References

1. Patton, G. C. *et al.* Our future: A Lancet commission on adolescent health and wellbeing. *The Lancet* **387,** 2423–2478 (2016).
2. Patton, G. C. *et al.* Adolescence and the next generation. *Nature* **554,** 458–466 (2018).
3. Dahl, R. E., Allen, N. B., Wilbrecht, L. & Suleiman, A. B. Importance of investing in adolescence from a developmental science perspective. *Nature* **554,** 441–450 (2018).
4. Telzer, E. H. Dopaminergic reward sensitivity can promote adolescent health: A new perspective on the mechanism of ventral striatum activation. *Dev. Cogn. Neurosci.* **17,** 57–67 (2016).
5. Crone, E. A. & Dahl, R. E. Understanding adolescence as a period of social-affective engagement and goal flexibility. *Nat. Rev. Neurosci.* **13,** 636–650 (2012).
6. Yeager, D. S., Dahl, R. E. & Dweck, C. S. Why interventions to influence adolescent behavior often fail but could succeed. *Perspect. Psychol. Sci.* **13,** 101–122 (2018).
7. Steinberg, L. How to improve the health of American adolescents. *Perspect. Psychol. Sci.* **10,** 711–715 (2015).
8. Eccles, J. S., Lord, S. & Midgley, C. What are we doing to early adolescents? The impact of educational contexts on early adolescents. *Am. J. Educ.* **99,** 521–521 (1991).
9. Cohen, G. L., Garcia, J. & Goyer, J. P. Turning point: Targeted, tailored, and timely psychological intervention. in *Handbook of Competence and Motivation (2nd Edition): Theory and Applicaiton* (Guilford Press, 2017).
10. Goyer, J. P. *et al.* Self-affirmation facilitates minority middle schoolers' progress along college trajectories. *Proc. Natl. Acad. Sci.* **114,** 7594–7599 (2017).

11. Yeager, D. S. & Walton, G. M. Social-psychological interventions in education: They're not magic. *Rev. Educ. Res.* **81,** 267–301 (2011).

12. Bloom, H. S., Raudenbush, S. W., Weiss, M. J. & Porter, K. Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *J. Res. Educ. Eff.* **10,** 817–842 (2017).

13. Weiss, M. J., Bloom, H. S. & Brock, T. A conceptual framework for studying the sources of variation in program effects. *J. Policy Anal. Manage.* **33,** 778–808 (2014).

14. Rothman, K. & Greenland, S. Causation and causal inference in epidemiology. *Am. J. Public Health* **95,** S144–S150 (2005).

15. Hill, J. L. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* **20,** 217–240 (2011).

16. Bloom, H. S. & Michalopoulos, C. When is the story in the subgroups?: Strategies for interpreting and reporting intervention effects for subgroups. *Prev. Sci.* **14,** 179–188 (2013).

17. Morgan, S. L. & Winship, C. *Counterfactuals and causal inference*. (Cambridge University Press, 2014).

18. Shadish, W. R., Cook, T. D. & Campbell, T. D. *Experimental and quasi-experimental designs for generalized causal inference*. (Wadsworth Publishing, 2001).

19. Kish, L. *Statistical design for research*. (John Wiley & Sons, 2004).

20. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1,** 0021 (2017).

21. Pleis, J. R., Ward, B. W. & Lucas, J. W. *Vital and health statistics: Summary health statistics for U.S. adults: National health interview survey, 2009.* (U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics, 2010).

22. Lynch, S. M. Cohort and life-course patterns in the relationship between education and health: A hierarchical approach. *Demography* **40,** 309–331 (2003).

23. McFarland, J., Stark, P. & Cui, J. *Trends in high school dropout and completion rates in the United States: 2013*. (U.S. Department of Education, 2016).

24. Loveless, T. The 2017 Brown Center report on American education. *Brookings* (2017).

25. Heckman, J. J. & LaFontaine, P. A. The American high school graduation rate: Trends and levels. *Rev. Econ. Stat.* **92,** 244–262 (2010).

26. Goldin, C. D. & Katz, L. F. *The race between education and technology*. (Harvard University Press, 2009).

27. Bryk, A. S. Support a science of performance improvement. *Phi Delta Kappan* **90,** 597–600 (2009).

28. Bryk, A. S. 2014 AERA Distinguished Lecture. *Educ. Res.* (2015). doi:10.3102/0013189X15621543

29. Carroll, J. M., Muller, C., Grodsky, E. & Warren, J. R. Tracking health inequalities from high school to midlife. *Soc. Forces* **96,** 591–628 (2017).

30. Crosnoe, R. *Fitting in, standing out: Navigating the social challenges of high school to get an education*. (Cambridge University Press, 2011).

31. Yeager, D. S. & Dweck, C. S. Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educ. Psychol.* **47,** 302–314 (2012).

32. Blackwell, L. S., Trzesniewski, K. H. & Dweck, C. S. Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Dev.* **78,** 246–263 (2007).

33. Aronson, J. M., Fried, C. B. & Good, C. Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. *J. Exp. Soc. Psychol.* **38,** 113–125 (2002).

34. Paunesku, D. *et al.* Mindset interventions are a scalable treatment for academic underachievement. *Psychol. Sci.* **26,** 284–293 (2015).

35. Wilson, T. D. & Juarez, L. P. Intuition is not evidence: Prescriptions for behavioral interventions from social psychology. *Behav. Sci. Policy* **1,** 13–20 (2015).

36. Walton, G. M. The new science of wise psychological interventions. *Curr. Dir. Psychol. Sci.* **23,** 73–82 (2014).

37. Ross, L. & Nisbett, R. E. *The person and the situation: Perspectives of social psychology*. (Pinter & Martin Ltd, 1991).

38. Asch, S. E. Studies in the principles of judgments and attitudes: II. Determination of judgments by group and by ego standards. *J. Soc. Psychol.* **12,** 433–465 (1940).

39. Burnette, J. L., O'Boyle, E. H., VanEpps, E. M., Pollack, J. M. & Finkel, E. J. Mind-sets matter: A meta-analytic review of implicit theories and self-regulation. *Psychol. Bull.* **139,** 655–701 (2013).

40. Yeager, D. S. *et al.* Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *J. Educ. Psychol.* **108,** 374–391 (2016).

41. Wilson, T. D. & Linville, P. W. Improving the academic performance of college freshmen: Attribution therapy revisited. *J. Pers. Soc. Psychol.* **42,** 367 (1982).

42. Dweck, C. S. & Leggett, E. L. A social-cognitive approach to motivation and personality. *Psychol. Rev.* **95,** 256–273 (1988).

43. Owens, A., Reardon, S. F. & Jencks, C. Income segregation between schools and school districts. *Am. Educ. Res. J.* **53,** 1159–1197 (2016).

44. Crosnoe, R. Low-Income Students and the Socioeconomic Composition of Public High Schools. *Am. Sociol. Rev.* **74,** 709–730 (2009).

45. Schiller, K. S., Schmidt, W. H., Muller, C. & Houang, R. T. Hidden disparities: How courses and curricula shape opportunities in mathematics during high school. *Equity Excell. Educ.* **43,** 414–433 (2010).

46. Chein, J., Albert, D., O'Brien, L., Uckert, K. & Steinberg, L. Peers increase adolescent risk taking by enhancing activity in the brain's reward circuitry. *Dev. Sci.* **14,** 1–10 (2011).

47. Sherman, L. E., Payton, A. A., Hernandez, L. M., Greenfield, P. M. & Dapretto, M. The power of the like in adolescence: Effects of peer influence on neural and behavioral responses to social media. *Psychol. Sci.* **27,** 1027–1035 (2016).

48. Schlegel, A. The adolescent socialization of the Hopi girl. *Ethnology* **12,** 449–462 (1973).

49. Paluck, E. L., Shepherd, H. & Aronow, P. M. Changing climates of conflict: A social network experiment in 56 schools. *Proc. Natl. Acad. Sci.* **113,** 566–571 (2016).

50. Cohen, G. L. & Prinstein, M. J. Peer contagion of aggression and health risk behavior among adolescent males: An experimental investigation of effects on public conduct and private attitudes. *Child Dev.* **77,** 967–983 (2006).

51. Helms, S. W. *et al.* Adolescents misperceive and are influenced by high-status peers' health risk, deviant, and adaptive behavior. *Dev. Psychol.* **50,** 2697–2714 (2014).

52. Tipton, E., Yeager, D. S., Iachan, R. & Schneider, B. Designing probability samples to study treatment effect heterogeneity. in *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment.* (ed. Lavrakas, P. J.) (Wiley, in press).

53. Tipton, E. Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Eval. Rev.* **37,** 109–139 (2014).

54. Easton, J. Q., Johnson, E. & Sartain, L. *The predictive power of ninth-grade GPA*. (Chicago, IL: University of Chicago Consortium on School Research, 2017).

55. Allensworth, E. M. & Easton, J. Q. *The on-track indicator as a predictor of high school graduation*. (Consortium on Chicago School Research, University of Chicago, 2005).

56. Imai, K., Keele, L. & Tingley, D. A general approach to causal mediation analysis. *Psychol. Methods* **15,** 309–334 (2010).

57. Hill, C. J., Bloom, H. S., Black, A. R. & Lipsey, M. W. Empirical benchmarks for interpreting effect sizes in research. *Child Dev. Perspect.* **2,** 172–177 (2008).

58. Somers, M.-A. *et al. The enhanced reading opportunities study final report: The impact of supplemental literacy courses for struggling ninth-grade readers*. (Institute of Education Sciences, 2010).

59. Weiss, M. J. *et al.* How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *J. Res. Educ. Eff.* **10,** 843–876 (2017).

60. E.M. Fahle *et al. Stanford Education Data Archive: Technical documentation (Version 2.0).* (2017).

61. Hong, Y., Chiu, C., Dweck, C. S., Lin, D. M.-S. & Wan, W. Implicit theories, attributions, and coping: A meaning system approach. *J. Pers. Soc. Psychol.* **77,** 588–599 (1999).

62. Duckworth, A. L. & Yeager, D. S. Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educ. Res.* **44,** 237–251 (2015).

63. Paluck, E. L. Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *J. Pers. Soc. Psychol.* **96,** 574–587 (2009).

64. Murphy, M. C. & Dweck, C. S. A culture of genius: How an organization's lay theory shapes people's cognition, affect, and behavior. *Pers. Soc. Psychol. Bull.* **36,** 283–296 (2010).

65. Leslie, S.-J., Cimpian, A., Meyer, M. & Freeland, E. Expectations of brilliance underlie gender distributions across academic disciplines. *Science* **347,** 262–265 (2015).

66. Green, D. P. & Kern, H. L. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin. Q.* **76,** 491–511 (2012).

67. Thaler, R. H. & Sunstein, C. R. *Nudge: Improving decisions about health, wealth, and happiness*. (Yale University Press, 2008).

68. Allcott, H. Site selection bias in program evaluation. *Q. J. Econ.* **130,** 1117–1165 (2015).

Study Information

1. Title

1.1. Provide the working title of your study. It may be the same title that you submit for publication of your final manuscript, but it is not a requirement.

Average effects and cross-site variability of effects of a growth mindset intervention on 9th grade achievement in a national probability sample.

2. Authorship

Authors: David S. Yeager*, Paul Hanselman*, Carol Dweck, Chandra Muller, Robert Crosnoe, Barbara Schneider, Greg Walton, Dave Paunesku, Beth Tipton, Chris Hulleman, Angela Duckworth
* Primary authors

Input: Andy Gelman (Columbia), Jordan Axt (Virginia), Todd Rogers (Harvard), Mike Weiss (MDRC).

3. Primary Research Questions

3.1. Please list each research question included in this study.

RQ 1: What is the average treatment effect (ATE) of a Growth Mindset (GM) intervention on the GPA of 9th grade *students* in regular U.S. public high schools?

RQ 2: What is the conditional average treatment effect (CATE) of a GM intervention on the GPA of 9th grade *previously low-performing students* in regular U.S. public high schools?

RQ 3: How much does the CATE of a GM intervention (on the GPA of 9th grade *previously low-performing students*) vary across U.S. public high schools?

RQ 4: Do school-level factors explain the variability in the size of the CATE of the GM (on GPA for *previously low-performing students* in U.S. public high schools)?

GPA and *Previously low-performing students* are defined in the measured variables section of the analysis plan.

4. Hypotheses

4.1. For each of the research questions listed in the previous section, provide one or multiple specific and testable hypotheses. Please state if the hypotheses are directional or non-directional. If directional, state the direction. A predicted effect is also appropriate here.

H 1: **ATE for all students.** We hypothesize a *very small* (near zero) positive effect of a GM intervention for 9th grade *students* in regular U.S. public high schools (on GPA).

H 2: **CATE for previously low-performing students**: We hypothesize a *moderate* positive effect of a GM intervention for *previously low-performing 9th grade students* in regular U.S. public high schools (on GPA).[1]

[1] Note: We hypothesize a near zero effect for previously high-performing students because:
a) growth mindset theory predicts improvements for struggling students, not students who are unchallenged (e.g. Burnette et al. 2013);
b) there is range restriction for previously high-achievers;

H 3: Cross-school variation in CATE. We hypothesize that there will be significant cross-school variation in the school-average effect of the GM intervention for 9th grade *previously low-performing students* in regular U.S. public high schools (on GPA).

H 4: Explaining cross-school variation in CATE. Research question 4 involves confirmatory analyses of previously-untested hypotheses. In particular, we hypothesize that:

H 4a. Among previously low-performing students*,* the school-average effect of the GM intervention will vary based on *school achievement level.*[2]

Directionally, we hypothesize that the CATE will be:

i. Smallest (and possibly zero) in the lowest-achievement schools.[3]
ii. Significant and positive in medium-achievement schools, and larger than in the lowest-performing schools.[4]
iii. Significant and positive, but of unknown relative magnitude, in the highest-achievement schools.[5]

Supplemental analyses will test for the effects for different strata of schools defined in the initial sampling plan.

H 4b. Among previously low-performing students*,* the school-average effect of the GM intervention will vary based on *school mindset saturation level*.

There are two competing directional hypotheses:[6]

i. *Larger effects on GPA in higher mindset saturation schools*. The reason why is that the environment reinforces the message over time. Giving the intervention in a high mindset saturation school is like "planting a seed in tilled soil".
ii. *Larger effects on GPA in lower mindset saturation schools*. The reason why is that in high mindset saturation schools students are already receiving growth mindset from their teachers and peers (because the control group is getting "treated") – the intervention is a "drop in the bucket". Meanwhile, in lower mindset saturation schools, students are most in need of a growth mindset – the intervention is like "water on parched soil."

We define *school achievement level* and *school mindset saturation level* below.

c) prior research that we are replicating (e.g. Paunesku et al., 2015; Yeager et al., 2016) only finds benefits for low-achieving students and does not focus on main effects in the full sample.

Thus, the ATE for the average student (RQ 1), which includes previously low- and high-performing students, is expected to be very small and positive. The effect for previously low-performers is expected to be moderate positive, relatively larger than for the full sample, and statistically significant.

[2] The conceptual hypothesis is that the rigor and standards in the school will interact with the treatment effect, under the theory that mindset interventions allow students to take better advantage of the instruction in the school.

[3] The rationale is that even though students in low-performing schools may face low levels of motivation, motivation may be less consequential for grades in schools with inadequate instruction or unsafe learning environments;

[4] The rationale is that student motivation may suffer in such schools and therefore may be lifted by the intervention, because instruction and learning environments are adequate but motivation is sub-optimal;

[5] The reason why we do not have predictions for higher-performing schools is that, on the one hand, they could have optimal motivation already and show weaker growth mindset treatment effects. Hence high-achieving schools might show the same effects as low-achieving schools. On the other hand higher-performing schools could have high rigor but a fixed mindset culture that could benefit more from a growth mindset treatment. Hence high-achieving schools might show stronger effects than low-achieving schools.

[6] These were first tested in Paunesku's dissertation.

Sampling Plan

In this section we will ask you to describe how you plan to collect samples, as well as the number of samples you plan to collect and your rationale for this decision. Please keep in mind that the data described in this section should be the actual data used for analysis, so if you are using a subset of a larger dataset, please describe the subset that will actually be used in your study.

5. Existing data

5.1. Preregistration is designed to make clear the distinction between confirmatory tests, specified prior to seeing the data, and exploratory analyses conducted after observing the data. Therefore, creating a research plan in which existing data will be used presents unique challenges. Please select the description that best describes your situation. Please do not hesitate to contact us if you have questions about how to answer this question (prereg@cos.io).

6. Explanation of existing data

6.1. If you indicate that you will be using some data that already exist in this study, please describe the steps you have taken to assure that you are unaware of any patterns or summary statistics in the data. This may include an explanation of how access to the data has been limited, who has observed the data, or how you have avoided observing any analysis of the specific data you will use in your study. The purpose of this question is to assure that the line between confirmatory and exploratory analysis is clear.

All of the present research questions concern the effect of an intervention on students' **Error! Reference source not found.** assigned from the point of intervention through the end of 9th grade. Students' grades have been recorded by school districts but have not yet been delivered to the researchers for 62 of the 66 schools in the study. Most of the school districts have delivered their datasets to a third-party research firm, ICF international, which is cleaning and merging the data. ICF international has not yet shared the full grades dataset with the research team.

ICF shared an "early release" of 4 of the 66 schools to the research team so that the team could provide feedback on the data cleaning and merging process and make additional requests for formatting and information that could be applied to the full set of 66 schools. Furthermore, data from those 4 schools were cleaned and analyzed by the research team, to inform the pre-registered analysis plan.

In sum, 62 of the 66 schools' achievement data are not yet delivered to the research team by the third-party research firm. Therefore we are not yet able to test any of the four research questions above.

7. Data collection procedures.

7.1. Please describe the process by which you will collect your data. If you are using human subjects, this should include the population from which you obtain subjects, recruitment efforts, payment for participation, how subjects will be selected for eligibility from the initial pool (e.g. inclusion and exclusion rules), and your study timeline. For studies that don't include human subjects, include information about how you will collect samples, duration of data gathering efforts, source or location of samples, or batch numbers you will use.

A research firm selected a sample of schools and recruited them into the study. A school liaison, working with the research firm, helped students complete the materials in a school computer lab. The sampling plan is described in the methodological report for the study.

8. Sample size

8.1. Describe the sample size of your study. How many units will be analyzed in the study? This could be the number of people, birds, classrooms, plots, interactions, or countries included. If the units are not individuals, then describe the size requirements for each unit. If you are using a clustered or multilevel design, how many units are you collecting at each level of the analysis?

School level: We took a stratified random sample of approximately 150 high schools from the universe of *all* regular U.S. public high schools.[7] 76 schools agreed to participate and collected student survey data and 66 provided student record data. The primary analytic sample for the present study will be the 66 schools with student achievement records.

Student level: Students are nested within schools. Our target was to include all 9th grade students within each randomly selected school. Students are included in analyses of treatment effects provided that they (a) saw the first page of treatment or control content, and (b) have student records data (for calculating GPA).

There were approximately 16,000 students who began Session 1 in the 76 schools, but we do not yet know the sample size for the subset with student records because the student records have not yet been delivered.

9. Sample size rationale

9.1. This could include a power analysis or an arbitrary constraint such as time, money, or personnel.

We recruited as many schools as could be recruited in the period between April 2015 and February 2016 (when the final schools implemented the treatment). The plan was for all schools to complete the intervention by the second month of 9th grade, but we extended the window until February of 2016 to increase sample size.

10. Stopping rule

10.1. If your data collection procedures do not give you full control over your exact sample size, specify how you will decide when to terminate your data collection.

Our data collection procedures did not give us full control over exact sample size. Termination of data collection occurred when it was too late in the year to include more schools (February 2016).

Variables

11. Manipulated variables

11.1. Describe all variables you plan to manipulate and the levels or treatment arms of each variable. For observational studies and meta-analyses, simply state that this is not applicable.

[7] A list of all U.S. public high schools was obtained from the Common Core of Data (National Center for Education Statistics) and supplemented through private databases (see Tipton, Yeager et al., in press).

We manipulated the materials during individual computer activities students completed at school. Students were randomly assigned by the computer program to be presented with either a growth mindset treatment or a control activity.

12. Measured variables

12.1. Describe each variable that you will measure. This will include outcome measures, as well as any predictors or covariates that you will measure. You do not need to include any variables that you plan on collecting if they are not going to be included in the confirmatory analyses of this study.

Outcome Measure(s):

GPA: GPA serves as the single confirmatory outcome measure for all hypotheses discussed in this analysis plan. GPA refers to the end-of-the-school-year GPA based on grades in core courses only. Grades are defined as grades on a 0-4.33 point scale. Core courses refer to math, science, social studies, and English/Language Arts. Grades in these core courses will be averaged (unweighted) to calculate GPA. Plans for data processing of grades are provided in the *Indices* section of the analysis plan.

Analyses of other configurations of grades are possible, but the confirmatory GPA variable will drive the main "story" regarding the effectiveness of the GM intervention. We will also explore the GM intervention's effects:

- In specific subjects (e.g., Math).
- On "poor performance" at the end of 9th grade, such that 1 = D/F average in core courses, 0 = satisfactory performance (C- or above).

Attitudes: Students self-reported a number of attitudes at pre-test and post-test, and analyses of these were pre-registered prior to data delivery (https://osf.io/byc2e/). Students self-reported mindsets and we measured their behavior on the "make-a-worksheet" challenge-seeking task (as interim outcomes). We collected measures of treatment fidelity (described in the exploratory analyses).

Student-level subgroup(s):

To answer RQ 2-4, we must define who is considered a "previously low-performing student." We do that here.

Previously low-performing students are defined as students who were earning grades lower (or equal) to 50 percent of his or her 9th grade school peers, prior to random assignment.[8] At an operational level, this is a student whose pre-random assignment GPA is at or below the 50th percentile of his/her 9th grade peers.

School-level subgroup(s):

To answer research question 4, we must define school achievement-level and school mindset saturation level. We do that here.

School-level achievement is defined as a latent variable derived from school-level achievement data. When testing for non-linear differences, we break school-level achievement into three categories, which align with the sampling plan and the hypotheses described in section 4.1:

[8] At a theoretical level, this is a student whose grades are not already maximal and who might show higher grades if motivated.

i. *Lowest-achievement schools* are those schools in the bottom quartile of the school-level achievement index.
ii. *Medium-achievement schools* are those schools that fall above the 25th and below the 75th percentile on the school-level achievement index.
iii. *Highest-achievement schools* are those schools in the top quartile of the school-level achievement index.

The *School-level achievement index*
, or measure, is described in section 13.

School-level mindset saturation is defined as the prevalence of growth mindset thinking in the school environment. A continuous variable will test the competing hypotheses described in section 4.1. The *School-level mindset saturation* are described in section 13.

Student-level Covariates

- Student male/female identification
- Student race/ethnicity
  - Dummy variables for Asian/Asian-American, Hispanic/Latino/a, Black/African-American, or other, with the referent group white students
- Student special education status (when available), dummy variable
- Student maternal education, dichotomized (1=four-year degree or higher, 0=less than a four-year degree)
- Student self-reported expectations for success (unless multi-collinearity with prior achievement is too high)
- ***Missing data***. We will not use list-wise deletion of cases that are missing covariates. We will impute missing covariates using the missing value dummy method, unless an alternative method is recommended by our statistician advisors.
- ***Collinearity***. We will remove a covariate from the models if it is too highly correlated with others, if there is excessive missing data, if it increases standard errors due to multi-collinearity, or if it prevents the model from converging.

13. Indices

13.1. If any measurements are going to be combined into an index (or even a mean), what measures will you use and how will they be combined? Include either a formula or a precise description of your method. If you are using a more complicated statistical method to combine measures (e.g. a factor analysis), you can note that here but describe the exact method in the analysis plan section.

For Outcome Measure(s):

Processing of grades for calculating GPA: Here is how we will process both pre- and post-intervention grades:

- We will analyze grades at the **term** level (e.g. fall or spring semester, or, in block schedules, a quarter). When only independent marking period grades are provided (e.g., marking periods 1-3, but not fall semester) then we will aggregate them to the term level (except in the case of missing pre-treatment data, as noted below).
- We will analyze only **core course** grades. We define core courses as math, science, social studies, and English/language arts. Non-core courses are electives, such as art, PE, computers or music. Non-core courses also include “support” classes, such as a lab class that is co-enrolled with a science class.

    - Core course designation will be made through a combination of course catalogs from schools and coding of course names. Coding of core courses will be independent of knowledge of the effect on outcomes of the study, and all syntax will be retained to enable robustness checks.
- If a school has a non-standard schedule (e.g. a block schedule) we may need to create a school-specific rule. We will annotate the syntax in the grades processing file, along with the justification. These decisions will be made prior to merging data with the randomized condition variable.
- Grades will be provided as letter grades (e.g., A, B, C). Core course grades will be re-coded on a 0 to 4.33 point scale, with 0 referring to "F" and 4.33 referring to "A+." Some schools will only report up to an A and so 4.0 will be the max grade for them. We will test the impact of putting all schools on the same scale (from 0 to 4).

GPA at the end of 9th grade.

- **Goal**: The main outcome is GPA in core courses at the end of 9th grade, weighting each core course equally. The initial plan was to average Fall 2015 and Spring 2016 achievement. However, some schools delivered the intervention in Spring 2016.
- **Confirmatory Operationalization.**
    - In schools that delivered the intervention in Fall of 2015, the outcome will be the average of Fall 2015 and Spring 2016 core course grades.
    - In schools that delivered the intervention in Spring of 2016, the outcome will be Spring 2016 core course grades only.
    - An exception will be if a school uses block scheduling and an entire quarter's grade is self-contained. In that event, we will look at the timing of the delivery of the treatment and the beginning and end of the quarter, to determine whether grades were recorded post-intervention or pre-intervention.
- **Missing data**.
    - We will use list-wise deletion of cases that are missing the primary outcome variable.
    - We will examine the impact of differential attrition on our inferences (for instance, perhaps the treatment kept marginal students from dropping out) and develop adjustments if attrition is differential.

For Student-level Subgroup(s)

Previously low-performing students:

- **Goal**. Conceptually, we wish to know if the treatment benefits students who were not already earning very high grades prior to receiving the intervention. The design of the study called for a fall intervention, and so we expected to use 8th grade achievement to define this subgroup. However some schools gave the intervention in the spring term of 9th grade (or perhaps after a full quarter's grades were recorded, for block schedules). Therefore grades from the fall of 9th grade are the most recent term. We will use the most recent term to create the pre-intervention low-performing student subgroup.
- **Confirmatory Operationalization.**
    - In schools that delivered the intervention in the Fall of 2015 prior GPA will be the average grade in 8th grade core courses, or if these were not provided, it will be 8th grade spring in core courses.
    - In schools that delivered the intervention in the Spring of 2016, the pre-intervention GPA will be the Fall 2015 GPA in core courses (i.e. Fall of 9th grade).
    - The exception to these two rules would be schools using block scheduling on a quarter system and where the treatment was delivered in the Fall but after an entire quarter's grades were recorded. In such cases, the completed first quarter Fall 2015 grades will be prior achievement and the remaining three quarters will be the outcome.
    - Pre-intervention GPA will be z-scored *within* schools.

To be precise, let:

$PreTreatGPA_{ij}$= the 8th grade GPA of student *i* attending 9th grade school *j* (among students where the study was implemented in in the fall of 9th grade), or the fall 9th grade GPA of student *i* attending 9th grade school *j* (among students where the study was implemented in the spring of 9th grade or after a Fall 8th grade block was finished).

$MedianPreTreatGPA_j$= the median pretreatment GPA of students attending 9th grade school *j*.

So:

$$Low\_performing_{ij} = \begin{cases} 1 \; if \; PreTreatGPA_{ij} \leq Median \; PreTreatGPA_j \\ 0 \; otherwise \end{cases}$$

Missing data.

- When students do not have 8th grade achievement but at least some of their grades were reported on a progress report in 9th grade prior to the delivery of the treatment, such as first quarter grades, then these will constitute pre-treatment GPA.
- When students do not have any of these grades, we will impute prior achievement values using their 8th grade test scores and self-reports of expectations for success in the coming year (cf. Hulleman & Harackeiwicz, 2009).

<u>For School-level Subgroup(s)</u>

<u>School-level achievement index</u>

- **Goal**. The goal for the school achievement variable is to understand whether treatment effects are different at school with different levels of rigor and standards. As a proxy for this, we created a latent variable of school achievement level for the purposes of stratification when randomly sampling schools to participate in this project (see Tipton, Yeager et al.).[9] This same latent variable will be used for subgroup analyses.
- **Confirmatory Operationalizations**.
  - There will be two operationalizations. The first is a continuous school achievement level variable, z-scored in the full population of approximately 12,000 regular U.S. public high schools (see Tipton, Yeager et al.).
  - A second operationalization is three categories of school-level achievement level: low (bottom quartile), medium (25th -75th), and high (top quartile), which is how this variable was used in the stratified sampling plan.
- **Missing data**.
  - There is no missing data on the school achievement level variable.

<u>School-level mindset saturation index</u>

- **Goal**. The goal is to assess whether environments with a strong mindset climate have weaker or stronger effects. However there is no established measure of mindset saturation.
- **Confirmatory Operationalizations.** We will test and report two operationalizations:
  - *Self-report.* The average "fixed mindset rating" on a 6-point scale for students in the school, measured prior to random assignment (both treatment and control group). The advantage of this measure is that it is a direct assessment of the construct – the prevalence of fixed/growth mindset thinking. The disadvantage of this measure is the potential for "reference bias" in making between-school comparisons (see Duckworth & Yeager, 2015). Another disadvantage is that peers

[9] As described in Figure 1 in Tipton, Yeager et al., the prior achievement variable was a composite of PSAT scores, AP scores, % AP Calculus test-takers, rating from greatschools.org (which are mostly composed of state test scores), and a state-level constant from the NAEP.

may conform more to perceived actions than private beliefs. Then again, reference bias may be minimal for growth mindset (see West et al. 2017).
  - *Behavior.* The number of challenging math assignments that students chose on the make-a-worksheet task, in the control group. The advantage of this measure is that it may not be subject to reference bias. Another advantage is that the mindset saturation in a school might be communicated more by how students *act* than what students say they *believe* (e.g. Haimovitz & Dweck; also see Paluck, 2009). A disadvantage of this behavioral measure is challenge-seeking is only a proxy for growth mindset and is only modestly correlated with growth mindset.
- **Missing data**.
  - NA

Design Plan

14. Study type

This is a randomized controlled trial. The researcher randomly assigned treatments to study subjects.

15. Blinding

Personnel who interact directly with the study subjects (either human or non-human subjects) were not aware of the assigned treatments.

16. Study design

16.1. Describe your study design. Examples include two-group, factorial, randomized block, and repeated measures. Is it a between (unpaired), within-subject (paired), or mixed design? Describe any counterbalancing required. Typical study designs for observation studies include cohort, cross sectional, and case-control studies.

Two-group randomized block design. Within each school (block) students are randomized to treatment or control. In addition, stratified random sampling was used to select schools.

17. Randomization

17.1. If you are doing a randomized study, how will you randomize, and at what level?

Randomization occurs at the student level via the computer after students log in to the system. Students, teachers, facilitators and researchers are all blind to condition.

Analysis Plan

You may describe one or more confirmatory analysis in this preregistration. Please remember that all analyses specified below must be reported in the final article, and any additional analyses must be noted as exploratory or hypothesis generating.

A confirmatory analysis plan must state up front which variables are predictors (independent) and which are the outcomes (dependent), otherwise it is an exploratory analysis. You are allowed to describe any exploratory work here, but a clear confirmatory analysis is required.

18. Statistical models

RQ 1. ATE for all students.

To estimate the average treatment effects (ATE), we use the following fixed effects model:

$$Y_i = \alpha + \beta \cdot T_i + \gamma \cdot A_i + \sum_{k=1}^{K} \theta_k \cdot X_{ki} + \sum_{j=1}^{j} \rho_j \cdot S_{ji} + e_i \qquad (1)$$

where:

Y_i = the outcome for student *i* (*GPA*)
T_i = 1 if student *i* was randomized to treatment and zero otherwise,
A_i = the prior achievement for student *i*, z-scored within schools
X_{ki} = school-mean-centered baseline covariate *k* for student *i* (see section 12)
S_{ji} = indicator variable indicating student *i* attends school *j*

The model will use person-level survey weights (which include school-level adjustments) and will not include any school-level covariates. It will use cluster-robust standard errors, clustered at the school level, to account for the nesting of students within schools. The parameter of interest is β, the average effect of the GM intervention for all students.

RQ 2: ATE for previously low-performing students.

To answer research question two we will use equation (1) on the subsample of previously low-performing students (i.e. students below the median within their school). The parameter of interest is β, the average effect of the GM intervention for previously low-performing students.

We use the above model (rather than the random effects model in RQ3 and 4) because RQ1 and RQ2 seek to estimate the effect for the average student, not the average school.

Below are the conclusions we would draw from the analyses in RQ1 and RQ2.

		Full Sample (RQ 1)	
		$\hat{\boldsymbol{\beta}} > 0, p < .05$	$\hat{\boldsymbol{\beta}} \approx 0, p > .05$
Previously low-performing students (RQ 2)	$\hat{\boldsymbol{\beta}} > 0,$ $p < .05$	1. Replicated Yeager/Paunesku low-performer effect and surprisingly showed a main effect as well. *Program was effective, on average, for the full sample and for previously low-performers.*	2. Replicated Yeager/Paunesku low-performer effect and replicated Yeager et al. non-significant main effect. *Program was effective, on average, for previously low-performing students. Results were as expected.*
	$\hat{\boldsymbol{\beta}} \approx 0,$ $p > .05$	3. Failed to replicate Yeager/Paunesku low-performer effect, but surprisingly showed a main effect. *Program was effective, on average, for the full-sample, but not effective, on average, for the expected subgroup.*	4. Failed to replicate Yeager/Paunesku low-performer effect, replicated non-significant main effect. *Program was not effective, on average, for all students or for low-performers.*

RQ 3: Variability in effects across schools, among previously low-achieving students.

To estimate variability in the treatment effect across schools, we will estimate a mixed effects model in the subset of previously low-performing students, using the model described by Bloom et al. (2017):

Level one (students)

$$Y_{ij} = \alpha_j + \beta_j \cdot T_{ij} + \gamma \cdot A_{ij} + \sum_{k=1}^{K} \theta_k \cdot X_{kij} + e_{ij} \qquad e_{ij} \sim N(0, \sigma_T^2) \qquad (2)$$

Level two (schools)

$$\beta_j = \beta + r_j \qquad r_j \sim N(0, \tau^2) \qquad (3)$$

where:

Y_{ij} = the outcome for low-achieving student *i* from school *j* (*GPA*)
T_{ij} = 1 if student *i* from school *j* was randomized to treatment and zero otherwise,
A_{ij} = the prior achievement for student *i* from school *j*, z-scored within schools
X_{kij} = school-mean-centered baseline covariate *k* for student *i* from school *j*, (see section 12)

For each school (j) this model allows for a fixed school-specific intercept (α_j), to account for the possibility of differences across schools in the proportion of students who are randomly assigned to the treatment vs. control group. This model allows for the treatment effect among low-achieving students to vary randomly across schools, β_j, with variance τ^2. Note that the model allows the student-level residual variance to be different for treatment and control group members (represented by the subscript in the term σ_T^2). These analyses will use survey weights and not include any school-level covariates. The parameter of interest for RQ 3 is τ, the standard deviation of the school-level distribution of average treatment effects.

We will conclude the intervention effects vary across schools when either a permutation test or a Q-statistic from meta-analysis shows that τ is different from zero (see Bloom, Raudenbush, Weiss & Porter, 2016). We will interpret the practical significance of our estimate of τ by comparing it to published benchmarks in program evaluation research (Weiss et al., 2017).

Here are the conclusions we would draw from the analysis in RQ3:

- $\hat{\tau} > 0$, p < .05: The effectiveness of the GM intervention varies across schools.
- $\hat{\tau} \approx 0$, p > .05: There is no discernable evidence that the effectiveness of the GM intervention varies across schools.

If $\hat{\tau}$ > 0 and p < .05, we will also estimate and graphically present the school-level distribution of average GM effects, as described in Bloom, Raudenbush, Weiss & Porter, 2016.

RQ 4: Predicting variability in effects across schools, among previously low-achieving students. Regardless of the answer to RQs 1-3, we will test whether school factors predict variation in the GM intervention's effects among schools —i.e. the β_j's in equation (3). The moderators are school achievement level and mindset saturation. All models will control for percent minority (black, Latino/a, or Native American) because these could be confounded with school achievement level and mindset saturation. These are confirmatory analyses of exploratory hypotheses – thus the approach to the analyses is more flexible than the approach for RQ 1- 3. This will also require a more cautious interpretation.

To preview, we will conduct four parametric tests:

1. School achievement as a continuous variable
2. School achievement as a categorical variable
3. Mindset saturation assessed via self-reports
4. Mindset saturation assessed via behavior

Then we will estimate a flexible, non-parametric model that likely will use Bayesian inference.

As a first test, we will examine independent, linear predictors. Specifically, we estimate a two-level mixed effects model. The level 1 model is specified in equation (2). The level 2 model is in equation (4) below:

Level two (schools)

$$\beta_j = \beta + \delta \cdot A_j + \pi \cdot M_j + \lambda \cdot S_j + r_j \qquad r_j \sim N(0, \tau^2) \qquad (4)$$

Where:

A_j = The grand-mean centered school achievement level for school j, coded continuously
M_j = The grand-mean centered percent minority (black, Latino/a or Native American) in school j
S_j = The grand-mean centered saturation of fixed mindset for school j

The significance and direction of δ will answer RQ 4a. The significance and direction of λ will answer RQ 4b. We do not have a substantive hypothesis about the π parameter, for minority composition, but would attempt to interpret and understand it if it was significant.

We will test whether there is a significant reduction in τ^2 as a result of the inclusion of school-level covariates (i.e. comparison of questions (3) and (4)). This will answer the research question of whether these three school-level factors in general explain variability in the GM effect among schools.

Second, we will use the school-achievement level variable coded into the three categories that constituted the sampling strata (bottom 25%, middle 50%, and top 25%) and conduct planned contrasts of subgroup ATEs. In the second model mindset saturation will still be a continuous variable.

Third, with consultation from statisticians, we will evaluate potential non-parametric models to examine the independent and interactive impact of school-level moderators on between-school variability in the treatment impact β_j in equation 4 (e.g. likely a variation on Bayesian Additive Regression Trees). These models will test robustness of results to potential confounds in the school-level moderators (such as rural/urban or poverty concentration). To avoid over-interpretation of results, we will provide the statisticians with a dataset where the name of the variable and the meaning of the value labels are masked. The initial summary of the significant moderators will be generated by the statisticians, blind to the identities of the variables or of the treatment or control values.

19. Transformations

19.1. If you plan on transforming, centering, recoding the data, or will require a coding scheme for categorical variables, please describe that process.

See indices above.

20. Follow-up analyses

20.1. If not specified previously, will you be conducting any confirmatory analyses to follow up on effects in your statistical model, such as subgroup analyses, pairwise or complex contrasts, or follow-up tests from interactions. Remember that any analyses not specified in this research plan must be noted as exploratory.

See primary analyses above and planned analyses below.

We will also examine whether the data meet the assumptions of the linear models. If they do not, we will adjust the model and possibly the estimation methods for standard errors to fit the data as appropriate. We expect the linear models with robust standard errors to be appropriate, however.

For research question 4, we will test the planned sub-groups of "low" "medium" and "high" achieving schools, and we will allow the penalized non-parametric models to tell us where subgroup effects are appearing.

Analyses of the characteristics of schools that did and did not agree to participate will be used to assess whether the answers to the RQs generalize the population of *all* 9th grade regular U.S. public high schools or only those represented by the sample that agreed to participate. We expect that non-participation will be unrelated to observable characteristics following non-response adjustment (i.e. weighting).

21. Inference criteria

21.1. What criteria will you use to make inferences? Please describe the information you will use (e.g. p-values, Bayes factors, specific model fit indices), as well as cut-off criterion, where appropriate. Will you be using one or two tailed tests for each of your analyses? If you are comparing multiple conditions or testing multiple hypotheses, will you account for this?

p<.05, two-tailed, for RQ1-3. For RQ1-3 we limit the potential for a multiple testing problem by using one outcome (GPA), measured in a one way, to answer these confirmatory research questions, resulting in three null hypothesis tests.

For RQ4, we reduce multiple testing by using two operationalizations of school achievement (continuous vs. dichotomous) and two operationalizations of mindset saturation (self-report and behavioral). These four null hypothesis tests will use *p* <.05. We will then use Bayesian inference with penalties for over-fitting for the flexible, non-parametric models; because the models are Bayesian, they do not involve null hypothesis tests.

22. Data exclusion

22.1. How will you determine what data or samples, if any, to exclude from your analyses? How will outliers be handled?

Participants will be included as long as they saw the first page of the treatment or control exercises and had linked student transcript data.

23. Missing data

23.1. How will you deal with incomplete or missing data?

For academic outcomes we will default to listwise deletion for missing data. We will examine whether there was any attrition from the study (i.e. mid-semester dropouts) and whether that was differential by condition. If so, we will consult with statistical experts to develop a missing data plan (e.g. weights or propensity scores).

For prior achievement, we will impute missing grades using test scores or self-reported expectations for doing well that year (when students have data on those variables). Imputed values will be z-scored so that they are on the same metric as the grades. This approach comes in part from Yeager, Romero et al. (2016), who created a prior achievement composite with grades, test scores, and self-reported expectancy for grades. Expectancies were an intervention moderator in a related intervention (Hulleman & Harakeiwicz, 2009).

24. Planned additional analyses (optional)

These four sets of planned exploratory analyses supplement the primary research questions above and will be reported in the manuscript or supplement regardless of the outcomes:

1. **Poor performance rate**. We will report a secondary outcome of poor performance, defined as a D/F average in core courses at the end of 9th grade. We will attempt to replicate the results reported in Yeager, Romero et al. (2016) and Paunesku et al. (2015), which found reductions in poor performance rates for treated individuals (also see Yeager et al. 2014, *JEP:General*). Poor performance rates furthermore represent a conceptual replication of higher-education interventions, which found treatment effects on full-

time enrollment rates (Yeager et al., 2016). Although GPA is the primary outcome, the poor performance rate is also highly practically relevant because it is a strong predictor of eventual high school graduation (see Allensworth et al. 2005).

2. **Results for different courses**: We will report results separately by course (math, English, social studies, etc.) either in the paper or in an online supplement. There might be larger effects in math and science, under the assumption that lay beliefs about fixed ability are stronger in math and science and therefore might benefit more from correction via a growth mindset treatment.

3. **Intervention fidelity**: We will assess the implementation fidelity of our treatment and control conditions with the following measures: (1) the percentage of open-ended questions that students answered during their on-line sessions, (2) the percentage of screens that students opened (and presumably viewed) during their on-line sessions, (3) the student-level response rate, (4) the amount of distraction that students reported experiencing during their on-line sessions, and (5) the amount of distraction that students reported other students experienced during their on-line sessions. We will create a composite of all or a subset of these (using factor analysis or analogous data-reduction methods) and aggregate to the student and school level. We will explore whether intervention fidelity explains differences in treatment impact, and whether it is a mechanism for potential moderation by achievement level.

4. **Strength of manipulation check**: We will explore whether different schools show different treatment effects because they were more or less successful at delivering the treatment in a way that caused students to change their attitudes and interim behaviors, as measured by the size of the treatment effect on manipulation checks (self-reported mindsets and challenge-seeking behavior, after receiving the treatment) across schools.

The below additional planned exploratory analyses test alternative research questions. They could be presented as secondary analyses for the primary paper, or they could constitute papers of their own. Although these questions are not fully developed, we pre-register them here so that they are listed prior to seeing or analyzing any results, so as to constrain researcher degrees of freedom.

5. **Student-level moderators**: (1) academically negatively-stereotyped minority students (e.g., black, Latino, native American) vs. academically non-negatively stereotyped students (white or Asian-American students), (2) females vs. males (especially in quantitative classes), (3) and students who are socioeconomically disadvantaged (defined either by parental education/occupation or by free/reduced price lunch status) vs. advantaged students, (4) school track (i.e. advanced math vs. regular math); (5) attitudinal measures obtained at the beginning of students' first on-line session, such as initial growth mindsets (to replicate the marginally-significant moderation in Blackwell et al., 2007), expectations for academic success (conceptually replicating Hulleman & Harackeiwicz, 2009), or math anxiety. In general, we will test the conceptual hypothesis that students who face more disadvantage or have greater vulnerability might also show stronger treatment effects.

6. **Interaction of achievement level and mindset saturation**: When investigating research question 4, achievement level and mindset saturation could potentially interact. The greatest treatment effects might occur in places where there is the highest school-achievement level, but the weakest school level mindset saturation.

7. **Adding convenience sample schools to increase cross-site statistical power**. A planned supplemental analysis for Research Questions 3 and 4 will combine the treatment effect estimates obtained in the pilot study (Yeager et al. 2016) and in a replication in a convenience sample of urban district schools

(Hanselman et al. in prep) with the national study estimates, to increase the number of schools by 18. This will increase our power to detect cross-site variation in treatment effects. After merging these schools' data with the national sample, we will re-conduct the analyses for Research Questions 3 and 4. We will do this and report it (in the paper or the supplement) no matter what the results of the primary analysis are.

8. **Timing**: We will explore whether timing during the year (e.g. August vs. January), and timing during the day moderated treatment impacts. We have a working hypothesis that receiving an intervention during a busy time (e.g., right before thanksgiving or a holiday, on a Monday or Friday, the last period of the day) will show weaker effects. We will explore the hypothesis that timing within the year is a predictor of school likelihood of compliance (under the assumption that schools that participated earlier were more willing partners), and so experiments conducted earlier in the year might show larger treatment effects.

9. **School minority composition and stereotype threat**: we will test the exploratory hypothesis that negatively-stereotyped minority students in low-minority high schools might benefit most from the treatment. This would be a conceptual replication of a study of affirmation (a different psychological intervention) by Hanselman et al., 2014 and a potential test of stereotype threat explanations for growth mindset intervention effects.

Script (Optional)

The purpose of a fully commented analysis script is to unambiguously provide the responses to all of the questions raised in the analysis section. This step is not common, but we encourage you to try to create an analysis script, refine it using a modeled dataset, and use it in place of your written analysis plan.

25. Analysis scripts (Optional)

25.1. (Optional) Upload an analysis script with clear comments. This optional step is helpful in order to create a process that is completely transparent and increase the likelihood that your analysis can be replicated. We recommend that you run the code on a simulated dataset in order to check that it will run without errors.

NA

Other

26. Other

26.1. If there is any additional information that you feel needs to be included in your preregistration, please enter it here.

Other analyses are possible with this dataset. For instance, we plan to study whether variance in the treatment impact across math classes varies due to characteristics of teachers and classrooms. We also plan to conduct correlational analyses of the math classroom data. We will strip the present dataset of the teacher identifiers so that we can pre-register those analyses prior to conducting them.

An earlier version of this pre-registration was uploaded but not approved by researchers. It was "frozen" by the OSF robots after researchers did not approve it within 48 hours, but the plan was not complete because it had not yet been reviewed by MDRC. The present version has been reviewed by MDRC and is final and complete. We have "withdrawn" the previous frozen version.

NSLM Main Impacts and Heterogeneity of Impacts Supplementary Materials

February, 2018

Contents

2 Overview

This document reproduces statistics reported in the main text as well as supplemental analyses of the impacts of the growth mindset intervention.

First, we present preliminary analyses, such as balance tests of the effectiveness of random assignment.

Second, we present the results to answer each of our four primary research questions. For the core academic grades models, we report a series of sensitivity analyses.

Third, we present descriptive information about program implementation.

Fourth, we present data on the treatment effects on growth mindset self-reports, and evidence that the treatment effects did not vary across schools for students overall or for lower-achieving students in particular.

Unless otherwise noted, analyses employ weights to represent the population of regular public high schools in the U.S. (as pre-specified). Weights were provided by the survey research firm.

3 Preliminary Analyses

3.1 Defining Key Variables

- `Fixed Mindset Scale` = Post-intervention mindset 3-item scale; values range from 1 (most growth mindset) to 6 (most fixed mindset).
- `Growth Mindset` = Post-intervention growth mindset indicator (less than 3.0 on 3-item scale 1-6 fixed mindset scale)
- `Hard Problems` = Willingness to seek out challenges in math (number of hard problems selected in make-a-math-worksheet task); also referred to as "challenge-seeking" and the basis for a school mindset saturation measure
- `Self-report Growth Mindset Norm` = School average of growth mindset self-reports noted above, estimated from all students prior to random assignment. In the pre-registration, we called this "mindset saturation."
  - `Behavioral Growth Mindset Norm` = School average of hard problems students chose on the "make-a-math-worksheet" task, estimated from all students in the control group who completed the Time 2 survey. In the pre-registration, we called this "mindset saturation."
- `GPA` = Post-intervention grades in core academic courses (mathematics, English/ELA, science, social studies; omitting support courses) in 9th grade.
- `D/F Avg` = Core GPA in D/F Range (less than 2.0)

3.2 Descriptive statistics for full sample (unweighted)

	mean	sd	min	max	n
Female	0.490	0.500	0	1.0	13367
Maternal College	0.289	0.453	0	1.0	13421
Asian	0.038	0.191	0	1.0	13348
Black	0.112	0.315	0	1.0	13348
Hispanic	0.244	0.429	0	1.0	13348
White	0.430	0.495	0	1.0	13348
Other/Multiple Race/Ethnicity	0.176	0.381	0	1.0	13348
Growth Mindset (Dichotomous)	0.584	0.493	0	1.0	11979
Hard Problems	2.968	2.490	0	8.0	11628
GPA	2.587	1.041	0	4.3	12546
D/F Avg	0.264	0.441	0	1.0	12546

3.3 Descriptive statistics for previously lower-achieving students (unweighted)

We hypothesize larger intervention effects on GPA for lower-achieving students, defined as those with achievement below the school median prior to random assignment.

	mean	sd	min	max	n
Female	0.411	0.492	0	1.000	6807
Maternal College	0.216	0.412	0	1.000	6840
Asian	0.023	0.149	0	1.000	6796
Black	0.127	0.333	0	1.000	6796
Hispanic	0.284	0.451	0	1.000	6796
White	0.363	0.481	0	1.000	6796
Other/Multiple Race/Ethnicity	0.204	0.403	0	1.000	6796
Growth Mindset (Dichotomous)	0.513	0.500	0	1.000	5941
Hard Problems	2.663	2.389	0	8.000	5752
GPA	2.012	0.923	0	4.225	6219
D/F Avg	0.450	0.498	0	1.000	6219

3.4 Descriptive statistics for previously higher-achieving students (unweighted)

Higher-achieving students are defined as those above the school median prior to random assignment.

	mean	sd	min	max	n
Female	0.571	0.495	0	1.0	6553
Maternal College	0.365	0.481	0	1.0	6569
Asian	0.053	0.225	0	1.0	6545
Black	0.097	0.296	0	1.0	6545
Hispanic	0.202	0.402	0	1.0	6545
White	0.500	0.500	0	1.0	6545
Other/Multiple Race/Ethnicity	0.147	0.355	0	1.0	6545
Growth Mindset (Dichotomous)	0.655	0.475	0	1.0	6034
Hard Problems	3.267	2.551	0	8.0	5873
GPA	3.153	0.815	0	4.3	6323
D/F Avg	0.080	0.272	0	1.0	6323

4 Experimental Balance and Attrition

4.1 Balance on pre-treatment characteristics

Random assignment was effective at producing balance between groups in terms of characteristics measured prior to random assignment.

Variable	Trt Mean	Trt SD	Trt N	Ctl Mean	Ctl SD	Ctl N	Diff	p	Pooled S
Female	0.492	NA	6673	0.488	NA	6694	0.004	0.674	0.50
Maternal College	0.285	NA	6702	0.293	NA	6719	-0.008	0.294	0.45
Black	0.114	NA	6651	0.110	NA	6697	0.005	0.424	0.31
Asian	0.037	NA	6651	0.039	NA	6697	-0.002	0.548	0.19
Hispanic	0.246	NA	6651	0.241	NA	6697	0.005	0.529	0.42
White	0.432	NA	6651	0.429	NA	6697	0.003	0.769	0.49
Pre-treatment Fixed Mindset Scale	3.052	1.150	6692	3.066	1.164	6708	-0.013	0.505	1.15
Pre-treatment GPA	2.777	0.966	5649	2.801	0.964	5688	-0.025	0.176	0.96

4.2 Attrition among randomized students

The treatment and control groups did not differ in terms of missing data for the outcome variables measured in session 2.

Outcome	Trt Attrition	Trt N	Ctl Attrition	Ctl N	p
Fixed Mindset Scale	0.110	6702	0.105	6719	0.360
Growth Mindset Indicator	0.110	6702	0.105	6719	0.360
Number of Hard Problems Selected	0.133	6702	0.134	6719	0.884
Hypothetical Challenge-seeking	0.119	6702	0.112	6719	0.215

4.3 Calculations for CONSORT report

4.3.1 Enrollment

Participants were identified for participation by the third party research firm (ICF International) in consultation with school officials by virtue of grade membership and enrollment in targetted classes.

Considered	Parental Refusal	Intention to Treat (Randomized)
13617	73	13544

4.3.2 Allocation

Students were randomized at the start of the computerized activity. Students received the allocated intervention for session 1. Some students were absent and received no session 2 materials. Other students incorrectly inputted their names at session 2, and they were always given the control group materials.

All analyses are Intention-to-Treat, regardless of whether students saw the session 2 materials.

treatment	Total	As Allocated, Both Sessions	Absent	Non-allocated Session 2 Materials
0	6779	6219	553	7
1	6765	6131	573	61

4.3.3 Follow-up

There are two primary reasons why participants were lost to follow-up for the primary analyses of GPA outcomes. First, one school did not provide administrative records. Second, some students' GPAs could not be matched with the administrative data, either because their names could not be matched or because they left the school. We cannot discern which of the two sources of non-response at follow-up were at play. However, if the treatment kept some marginal students from dropping out, then that would bias effect sizes in the direction of smaller, more conservative estimates.

treatment	Intention to Treat	Non-reporting School	Grades not Available	Analytic Sample
0	6779	60	479	6296
1	6765	63	513	6250

5 Pre-Registered Treatment Impacts on Academic GPA

Here we report results from analyses exactly as specified in the pre-registered analysis plan.

We pre-specified four core questions about the impacts of the growth mindset treatment on core academic GPA:

1. What is the average treatment effect (ATE) of a Growth Mindset (GM) intervention on the GPA of 9th grade students in regular U.S. public high schools?
2. What is the conditional average treatment effect (CATE) of a GM intervention on the GPA of 9th grade previously lower-performing students in regular U.S. public high schools?
3. How much does the CATE of a GM intervention (on the GPA of 9th grade previously lower-performing students) vary across U.S. public high schools?
4. Do school-level factors explain the variability in the size of the CATE of the GM (on GPA for previously lower-performing students in U.S. public high schools)?

Because the models to answer RQs 3 and 4 depend on the model specification decisions in RQs 1 and 2, we report a series robustness checks and sensitivity analyses for RQs 1 and 2.

5.1 Main Academic Impacts Overall (Pre-registered RQ1)

Below we present the results for the primary specification, with pre-specified model options:

Estimated impact on core academic GPA for all students:

estimate	stderr	t	dof	p	n	n_schools	ctl_mn	ctl_sd
0.034	0.011	3.092	46	0.003	12542	65	2.584	1.044

5.2 Main Academic Impacts for Previously Low-achieving Students (Pre-registered RQ 2)

Results for the primary specification, with pre-specified model options:

Estimated impact on core academic GPA for low-achieving students:

estimate	stderr	t	dof	p	n	n_schools	ctl_mn	ctl_sd
0.083	0.026	3.144	46	0.003	6219	65	1.993	0.928

5.3 Additional Impacts Overall and by Prior Achievement

In addition to impacts on GPA, we hypothesized reductions in the likelihood of a D/F average among low-achieving students. The pattern of results is similar for this outcome as for Grade Point Average. A significant overall impact is driven by a larger effect among previously low-achieving students.

We did not expect impacts for high-achieving students on either outcome, and we find no evidence of effects for this group.

Mindset treatment impact estimates on 9th grade grades in core academic courses (65 schools):

Outcome	Sample	estimate	stderr	t	dof	p	n	n_schools	ctl_mn	ctl_sd
Core GPA	Full Sample	0.034	0.011	3.092	46	0.003	12542	65	2.584	1.044
Core GPA	Lower-achievers	0.083	0.026	3.144	46	0.003	6219	65	1.993	0.928
Core GPA	Higher-achievers	-0.006	0.018	-0.327	46	0.745	6323	65	3.164	0.796
D/F Average GPA	Full Sample	-0.028	0.009	-3.056	46	0.004	12542	65	0.268	0.443
D/F Average GPA	Lower-achievers	-0.059	0.017	-3.439	46	0.001	6219	65	0.464	0.499
D/F Average GPA	Higher-achievers	0.003	0.009	0.307	46	0.760	6323	65	0.077	0.266

5.4 Robustness and Sensitivity Analyses for RQ1 and RQ2

In addition to the pre-specified analyses, we considered the sensitivity of results to several alternative specifications, listed below. Note that pre-specified options are highlighted in bold.

Survey weights:

1. **Grade 9 weights** [pre-specified] = records weighted by weighted based on sampling design and non-response (including missing grade 9 GPA outcomes); weights calculated by survey firm
2. Grade 9 weights trimmed = records weighted by a modified version of the Grade 9 weights, with weights top coded to the 3rd highest value within school achievement groups
3. Design weights = records weighted by the inverse of treatment selection for the school, given the sampling design
4. No weights = all individual records assigned a constant weight, maintaining clustering corrections for strata and primary sampling unit

Alternate GPA Outcomes:

1. **Grade 9 post core** [pre-specified] = GPA in core academic courses (Mathematics, English Lanugage Arts, Science, Social Studies) from the intervention term to the end of the year; support courses not included
2. Grade 9 post academic = GPA in all academic courses (including support courses and Foreign Lanugage, etc.) from the intervention term to the end of the year
3. Grade 9 average core = GPA in core academic courses for all of 9th grade; expected to yeild *conservative* estimates because includes some pre-intervention information

Alternative Covariates:

1. None
2. School fixed effects = indicators for each school
3. Prior achievement = school fixed effects and prior GPA
4. **Full** [pre-specified] = school fixed effects, prior GPA, and demographic/academic characteristics (*)

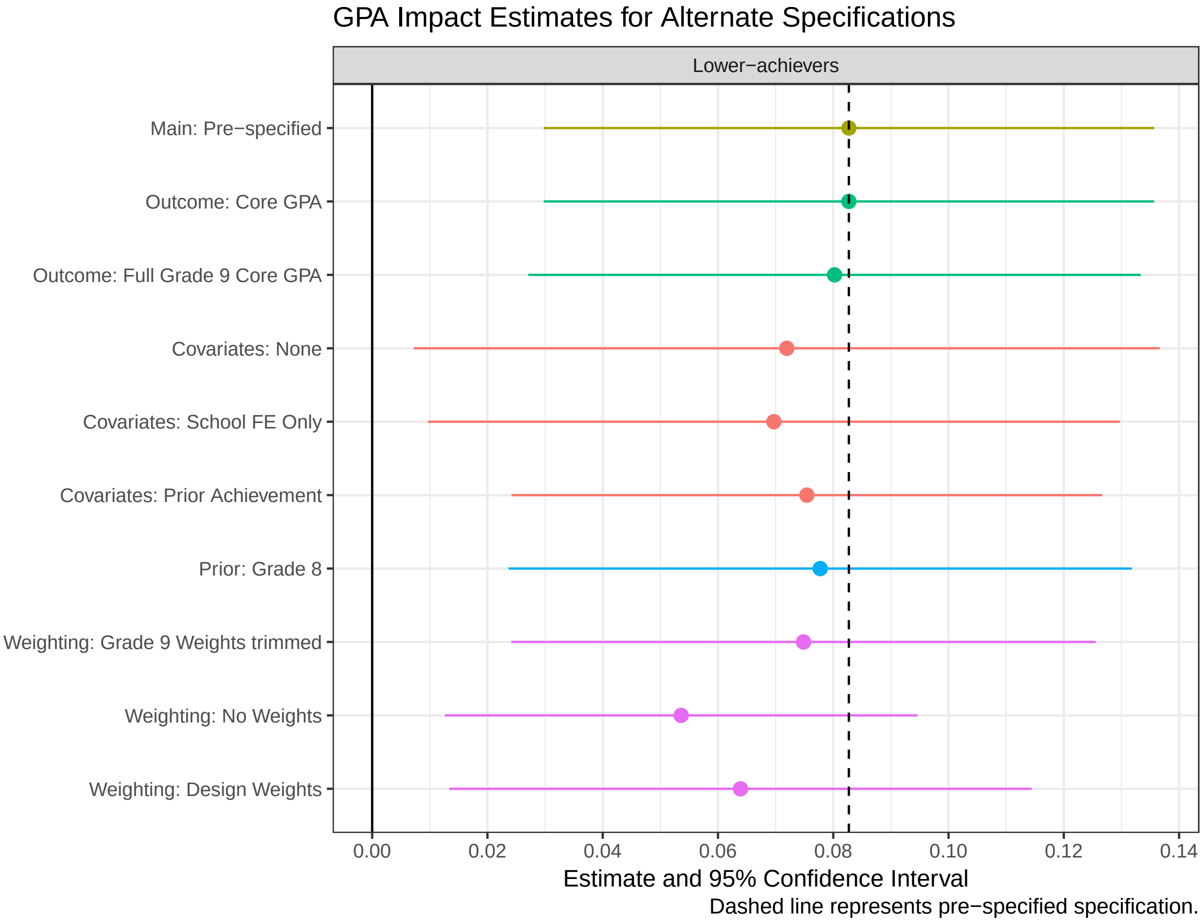
(*) Covariates:

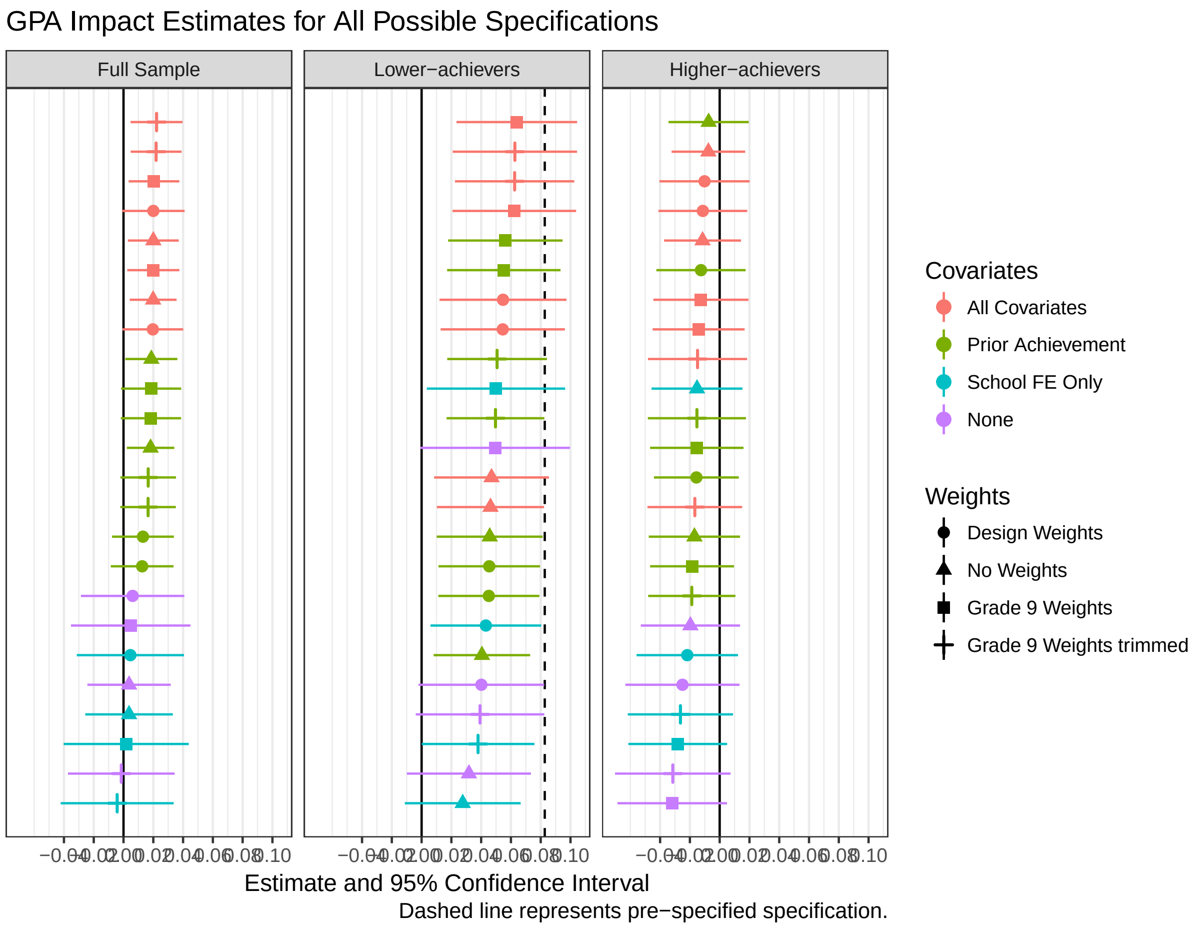
- standardized achievement, imputed zero if missing
- indicator for missing standardized achievement
- pre-treatment expectancy for school success, imputed zero if missing
- indicator for missing expectancy
- prior gpa self-report, imputed zero if missing
- indicator for missing prior gpa self-report
- gender
- race/ethnicity indicators (Asian, Black, Hispanic, Native American, Multi-racial, White)
- parental education categorical indicators (1-8)
- English Language Learner classification
- Special Education classification
- First year freshman indicator
- Free/reduced lunch indicator

Alternative prior achievement specification:

1. **Most recent pre-intervention GPA** [pre-specified] = Most recent pre-treatment GPA: grade 8 if intervention conducted in semester 1; grade 9 semester 1 if intervention conducted in semester 2
2. Grade 8 = Grade 8 GPA for all students

5.5 Summary of alternate specifications




GPA Impact Estimates for All Possible Specifications
Full Sample
Lower−achievers
Higher−achievers
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
Estimate and 95% Confidence Interval
Dashed line represents pre−specified specification.
Covariates
All Covariates
Prior Achievement
School FE Only
None
Weights
Design Weights
No Weights
Grade 9 Weights
Grade 9 Weights trimmed

5.6 Alternate Specification Estimates for Low-achieving Students

Alternate Weights Specifications

Weights	estimate	stderr	t	dof	p	n	n_schools	ctl_mn	ctl_sd
Grade 9 Weights	0.083	0.026	3.144	46	0.003	6219	65	1.993	0.928
Grade 9 Weights trimmed	0.075	0.025	2.972	46	0.005	6219	65	1.993	0.928
No Weights	0.054	0.020	2.631	46	0.012	6219	65	1.993	0.928
Design Weights	0.064	0.025	2.546	46	0.014	6219	65	1.993	0.928

Alternate Outcome Specifications

Outcome	estimate	stderr	t	dof	p	n	n_schools	ctl_mn	ctl_sd
All Academic GPA	0.064	0.020	3.178	46	0.003	6226	65	2.132	0.876
Core GPA	0.083	0.026	3.144	46	0.003	6219	65	1.993	0.928
Full Grade 9 Core GPA	0.080	0.026	3.039	46	0.004	6232	65	1.996	0.902

Alternate Covariate Specifications

Covariates	estimate	stderr	t	dof	p	n	n_schools	ctl_mn	ctl_sd
None	0.072	0.032	2.238	46	0.030	6219	65	1.993	0.928
School FE Only	0.070	0.030	2.338	46	0.024	6219	65	1.993	0.928
Prior Achievement	0.075	0.025	2.963	46	0.005	6219	65	1.993	0.928
All Covariates	0.083	0.026	3.144	46	0.003	6219	65	1.993	0.928

Alternate Prior Grades Specifications

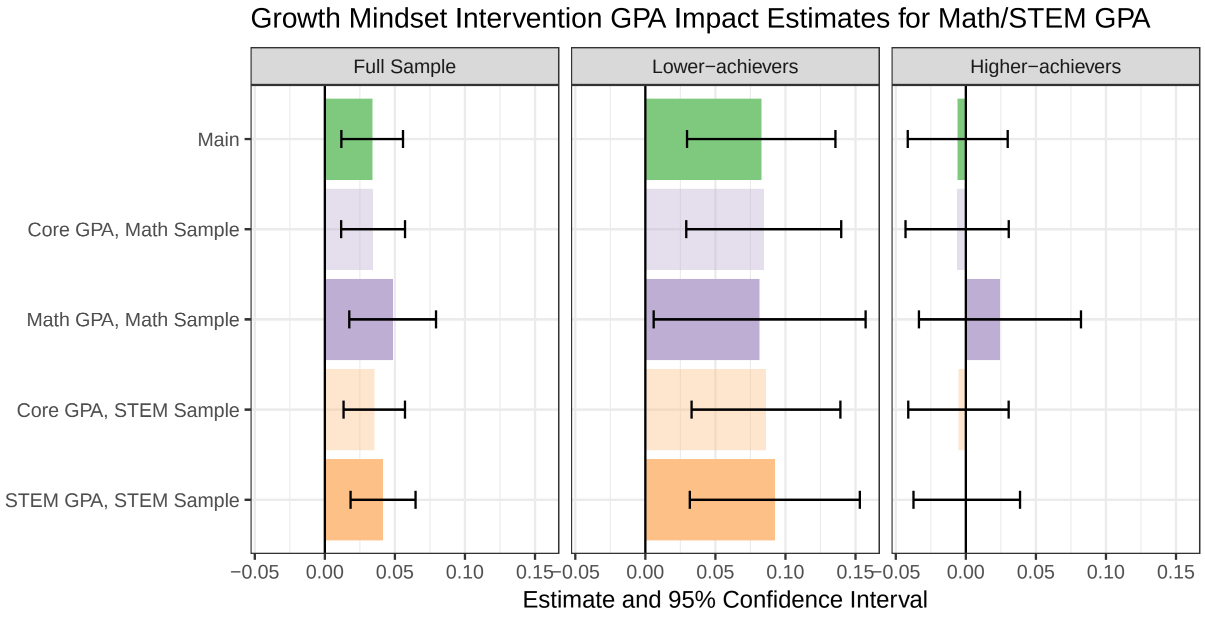
Prior GPA	estimate	stderr	t	dof	p	n	n_schools	ctl_mn	ctl_sd
Pre-intervention	0.083	0.026	3.144	46	0.003	6219	65	1.993	0.928
Grade 8	0.078	0.027	2.894	46	0.006	6407	65	2.055	0.962

5.7 Mathematics and STEM GPA

Given study focus on mathematics, we consider impacts on mathematics-specific and STEM (mathematics and science) GPA. Because some students are not observed in a mathematics or STEM course, we also re-estimate primary impacts (on core academic GPA) for the subset of students with non-missing Mathematics or GPA grades.

1. Main Impact = Core academic GPA (main specification above)
2. Core GPA, Math Subsample = Core academic GPA among students with non-missing mathematics-specific GPA
3. Mathematics GPA = Mathematics-specific GPA
4. Core GPA, STEM Subsample = Core academic GPA among students with non-missing STEM-specific GPA
5. STEM GPA = GPA in mathematics and science courses

Sample	Subject Specification	estimate	stderr	t	p	min95	max95	n
Full Sample	Main	0.034	0.011	3.092	0.003	0.012	0.056	12542
Full Sample	Core GPA, Math Sample	0.034	0.011	3.040	0.004	0.012	0.057	11893
Full Sample	Math GPA, Math Sample	0.048	0.015	3.145	0.003	0.017	0.079	11893
Full Sample	Core GPA, STEM Sample	0.035	0.011	3.238	0.002	0.013	0.057	12251
Full Sample	STEM GPA, STEM Sample	0.042	0.012	3.606	0.001	0.018	0.065	12251
Lower-achievers	Main	0.083	0.026	3.144	0.003	0.030	0.136	6219
Lower-achievers	Core GPA, Math Sample	0.084	0.027	3.076	0.004	0.029	0.140	5880
Lower-achievers	Math GPA, Math Sample	0.082	0.038	2.171	0.035	0.006	0.157	5880
Lower-achievers	Core GPA, STEM Sample	0.086	0.026	3.263	0.002	0.033	0.139	6051
Lower-achievers	STEM GPA, STEM Sample	0.092	0.030	3.066	0.004	0.032	0.153	6051
Higher-achievers	Main	-0.006	0.018	-0.327	0.745	-0.041	0.030	6323
Higher-achievers	Core GPA, Math Sample	-0.006	0.018	-0.340	0.736	-0.043	0.031	6013
Higher-achievers	Math GPA, Math Sample	0.024	0.029	0.847	0.401	-0.033	0.082	6013
Higher-achievers	Core GPA, STEM Sample	-0.005	0.018	-0.297	0.768	-0.041	0.031	6200
Higher-achievers	STEM GPA, STEM Sample	0.001	0.019	0.034	0.973	-0.037	0.039	6200



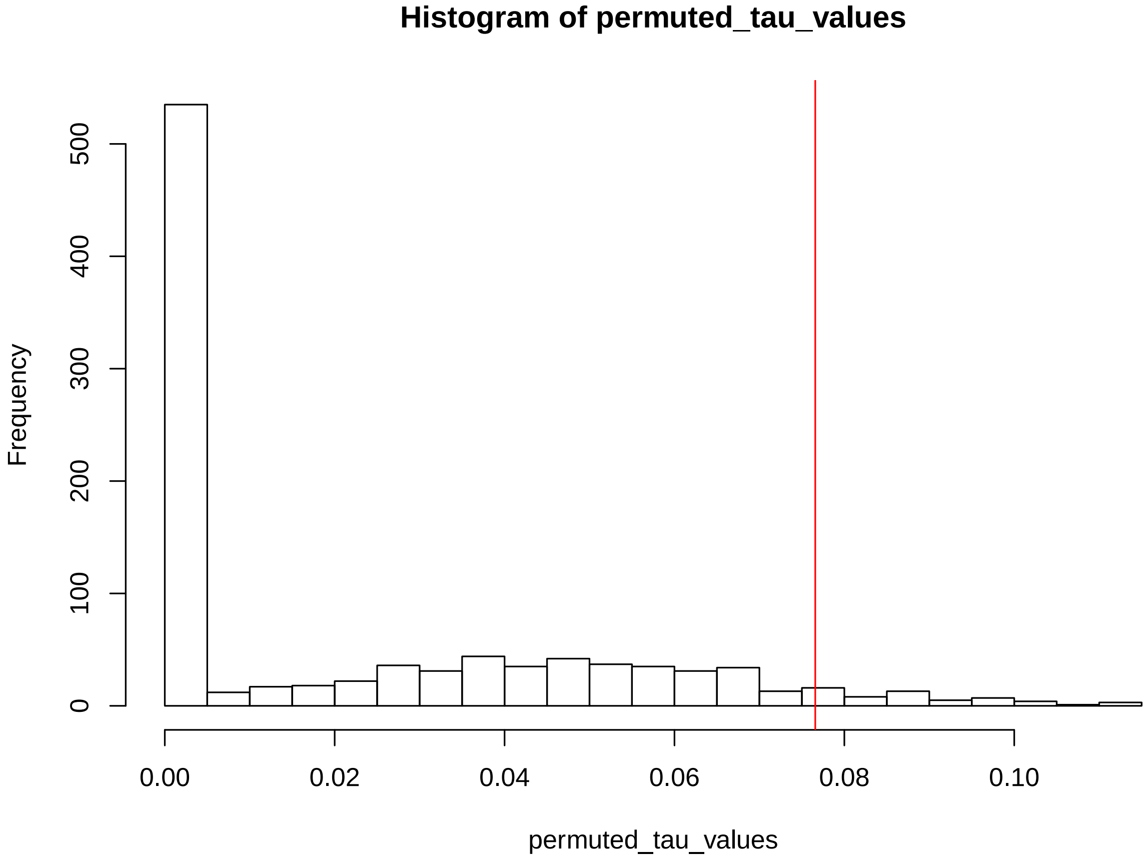
5.8 School Heterogeneity (Pre-registered RQ3)

The parameter of interest is tau, the standard deviation of intervention impacts across schools. Multi-level model heterogeneity analyses are estimated with restricted maximum likelihood and do not account for sampling weights. In the unweighted sample, the estimated treatment effect for lower-achieving students on post-intervention GPA in an average schools is 0.054. The estimated standard deviation of school impacts is 0.077.

Summary of School Heterogeneity Estimates for Low-achieving Students

Mean.Impact	Impact.SE	Impact.t	tau	tau.p.value.Q.method	tau.p.value.permutations
0.054	0.02	2.644	0.077	0.024	0.049

We present two statistical tests of the hypothesis that tau is greater than zero. The first is the Q-statistic proposed by Bloom et al. (2017). The second is a permutation test in which we randomly re-assign cases to schools and recalculate tau. The distribution of the estimate of tau across 999 permutations, with the observed value



marked, is:

5.9 School moderation (Pre-registered RQ4)

We tested two school moderators:

1. School achievment = composite of standardized achievement (PSAT) and other indicators.
2. School mindset norm = school mean number of "hard" problem selections among control students on the worksheet task (school mindset saturation)

School moderation analyses do not employ survey weights because it is not (in our view) a settled statistical issue to implement survey weights in mixed effects models. The net result of not using weights is that the average effect size is smaller and the random slopes are shrinking to a the unweighted mean, not the weighted mean. However our primary interest was in estimating the presence and functional form of the moderators, not the size of the treatment effects within sub-groups.

5.9.1 School Achievement Level

Following the pre-registered analysis plan, we divide schools into 3 categories base on a school achievement composite constructed from PSAT scores, state standardized tests, and AP participation and scores (See Tipton, Yeager, et al., in press, for details).

- The low group is schools below the 25th percentile.
- The medium group is schools between the 25th and 75th percentile.
- The high group is schools above the 75th percentile.

5.9.2 School Mindset Norm

The school mindset norm is defined as the prevalence of growth mindset thinking and behavior in the school environment (in the pre-registration, we called this "mindset saturation").

We test the hypothesis that there will be larger effects on GPA in higher mindset norm schools. The reason why might be that the environment reinforces the message over time. Giving the intervention in a high mindset norm schools might be like "planting a seed in tilled soil."

At the same time, it was possible that students might benefit most when they attend schools with unsupportive norms. Giving the treatment in low mindset norm schools might be like "water on parched soil."

Based on our pre-registered plan, we divided schools into "high" and "low" mindset norm groups based on mean challenge-seeking behavior in the control group. Challenge-seeking, as noted, is measured by the number of hard problems selected on the behavioral make-a-math-worksheet task. We label schools with above (below) average mean number of hard problem selects as high (low) behavioral growth mindset norm.

Estimates are from models comparable those considered for RQ3: school-level treatment slopes are random, intercepts are fixed, and analyses are unweighted.

Selected coefficients from school moderation of effects on core academic GPA among lower-achieving students (N=6,219)

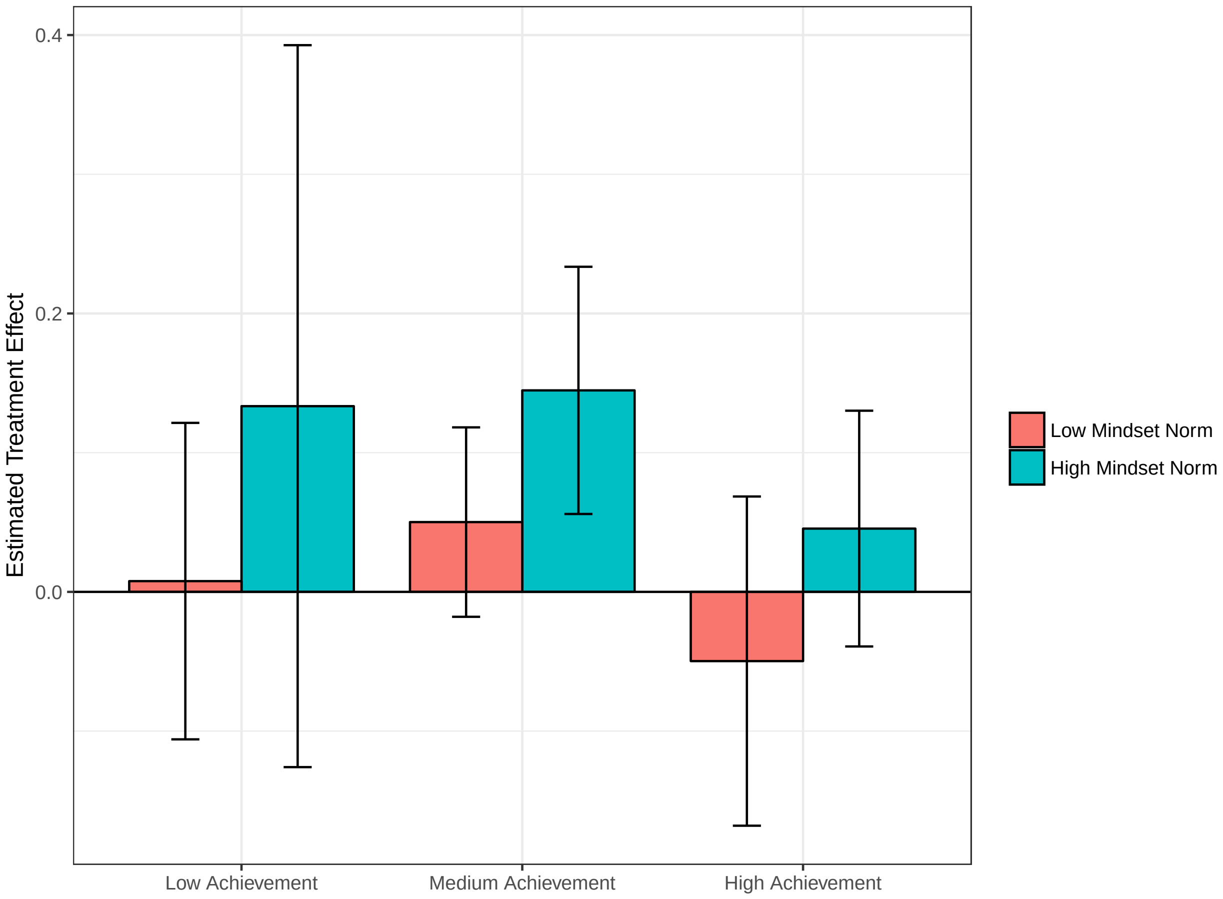
Parameter	Estimate	SE	z	p
Treatment	0.054	0.020	2.73	0.006
School Achievement	1.724	0.299	5.76	0.000
School Mindset Norm	-0.229	0.167	-1.37	0.170
School Proportion Stereotyped Minority	2.183	0.588	3.71	0.000
Treatment x Achievement	-0.066	0.029	-2.25	0.024
Treatment x Mindset Norm	0.127	0.054	2.33	0.020
Treatment x Minority	-0.107	0.097	-1.10	0.270

Post-hoc comparisons of school achievement sub-groups in terms of treatment effects on core academic GPA among lower-achieving students (N=6,219)

Parameter	Estimate	SE	z	p
Treatment	0.091	0.027	3.32	0.001
School Achievement Low (vs. Middle)	0.590	0.182	3.25	0.001
School Achievement High (vs. Middle)	0.603	0.103	5.83	0.000
School Mindset Norm	0.933	0.143	6.52	0.000
School Proportion Stereotyped Minority	-0.556	0.223	-2.49	0.013
Treatment x Low Achievement	-0.026	0.074	-0.36	0.721
Treatment x High Achievement	-0.096	0.047	-2.04	0.041
Treatment x Mindset Norm	0.108	0.056	1.94	0.053
Treatment x Minority	-0.025	0.106	-0.24	0.812

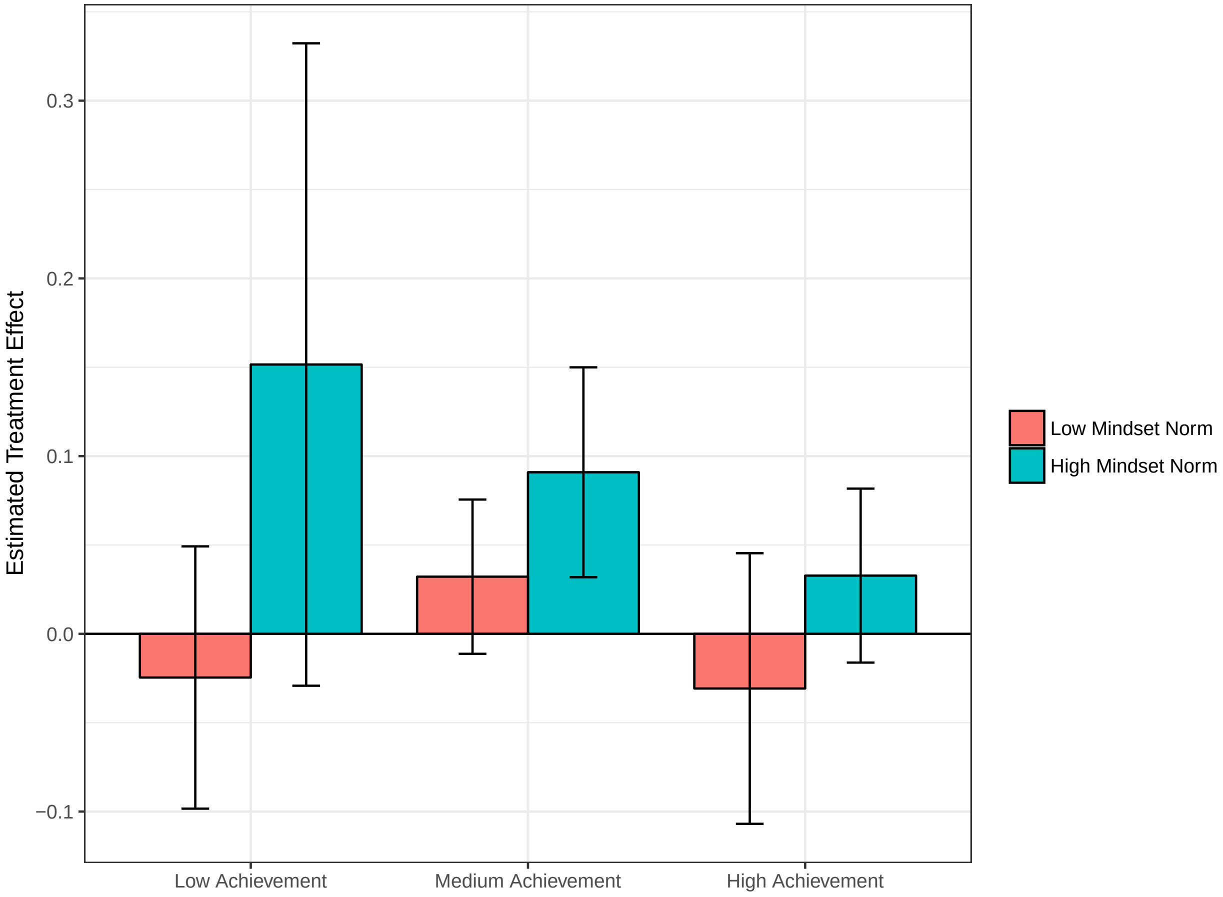
6 Figures depicting variation in treatment effects (RQ4)

6.1 Bar chart depicting conditional average treatment effects by school achievement and mindset norms groups, estimated in the pre-registered mixed effects model



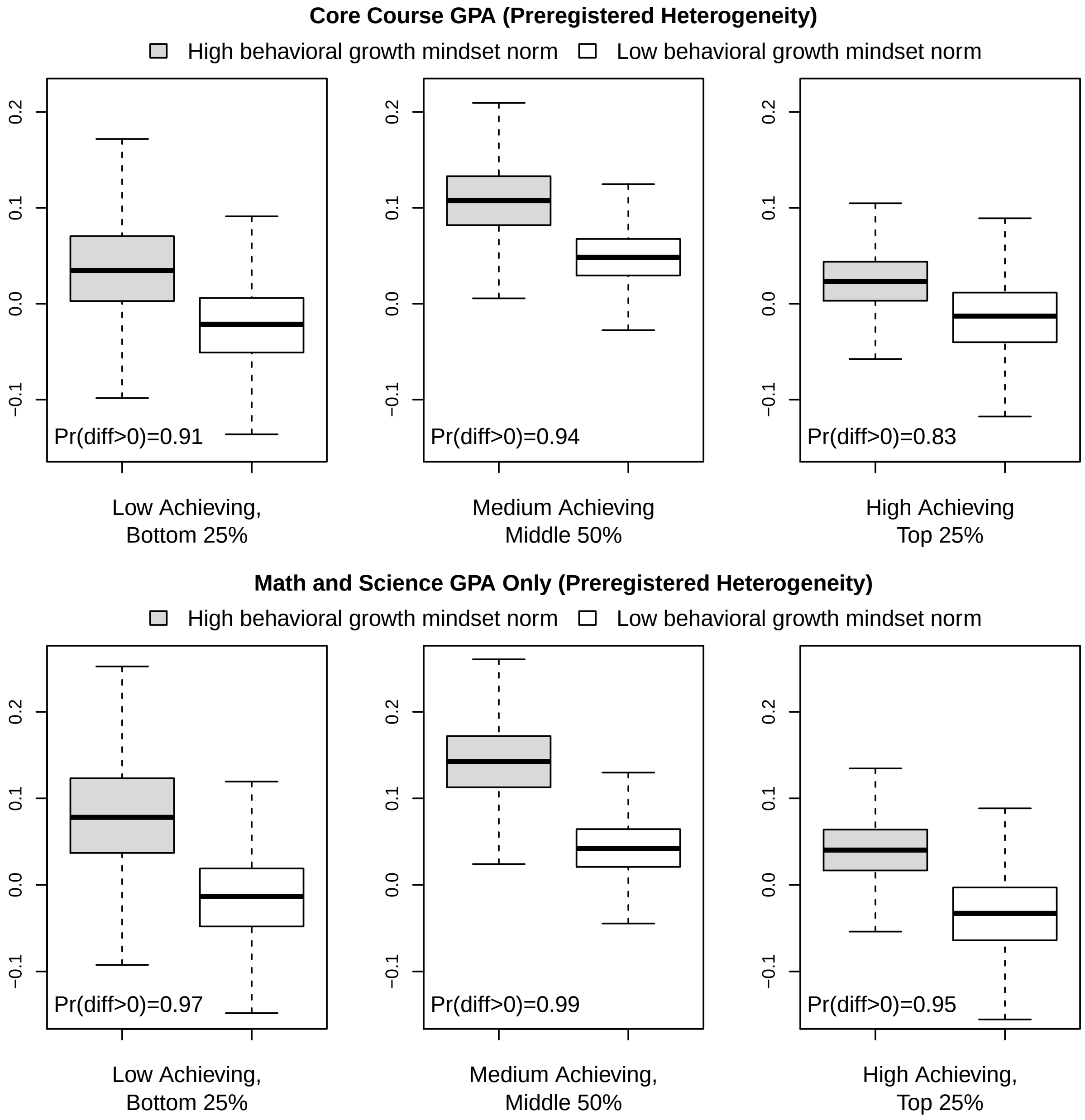
Achievement Level	Growth Mindset Norm	Estimate	SE	z	p	N	Schools
Low Achievement	Low Mindset Norm	0.008	0.058	0.13	0.894	698	7
Low Achievement	High Mindset Norm	0.133	0.132	1.01	0.313	115	3
Medium Achievement	Low Mindset Norm	0.050	0.035	1.44	0.149	2009	21
Medium Achievement	High Mindset Norm	0.145	0.045	3.19	0.001	1197	11
High Achievement	Low Mindset Norm	-0.050	0.060	-0.82	0.410	660	7
High Achievement	High Mindset Norm	0.045	0.043	1.05	0.293	2852	8

6.2 Bar chart depicting conditional average treatment effects in STEM GPA by school achievement and mindset norms groups, estimated in the preregistered mixed effects model

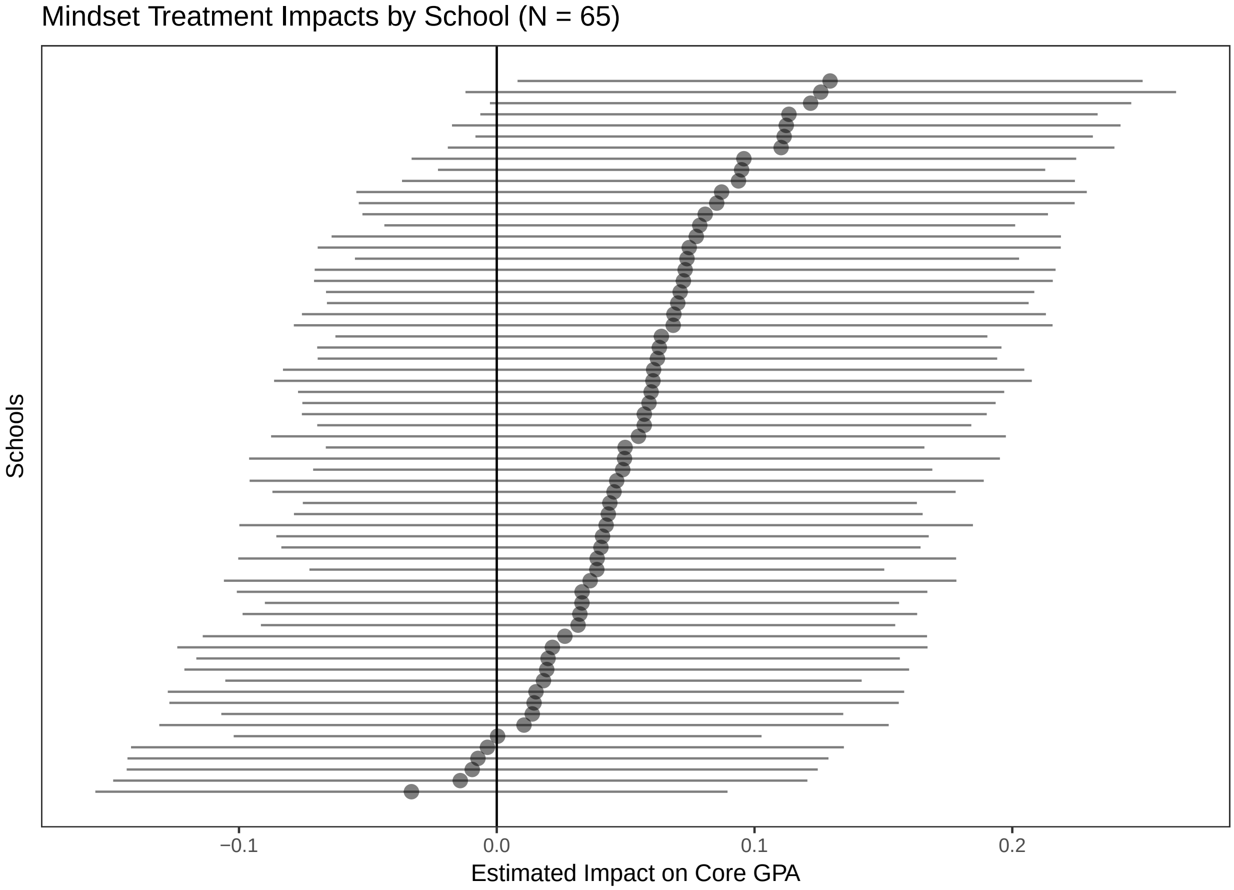


Achievement Level	Growth Mindset Norm	Estimate	SE	z	p	N	Schools
Low Achievement	Low Mindset Norm	-0.025	0.038	-0.65	0.514	1390	9
Low Achievement	High Mindset Norm	0.152	0.092	1.64	0.100	232	3
Medium Achievement	Low Mindset Norm	0.032	0.022	1.45	0.146	4005	24
Medium Achievement	High Mindset Norm	0.091	0.030	3.02	0.003	2167	12
High Achievement	Low Mindset Norm	-0.031	0.039	-0.79	0.428	1301	8
High Achievement	High Mindset Norm	0.033	0.025	1.31	0.190	3156	9

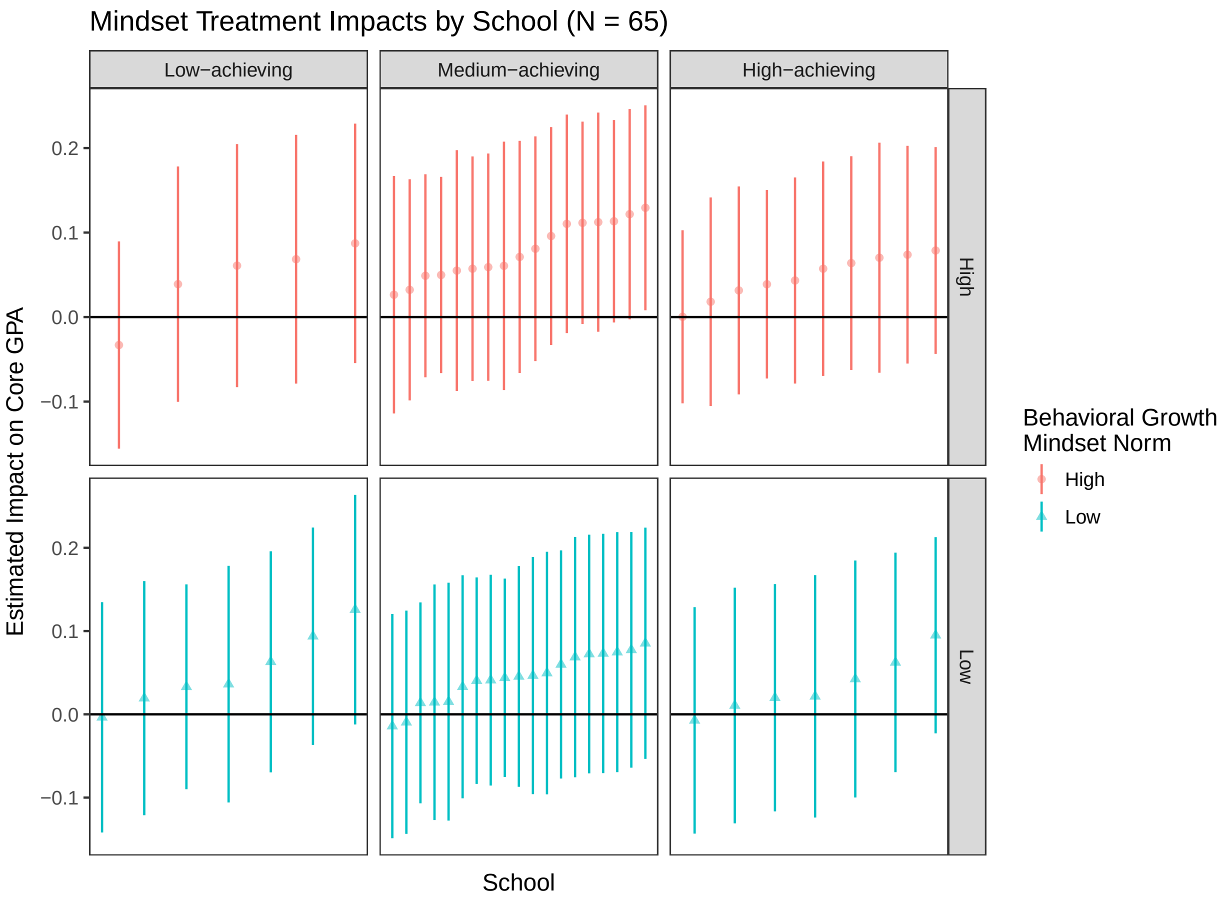
6.3 Box plots depicting posterior distributions for conditional average treatment effects estimated in the hierarchical Bayesian Additive Regression Trees (BART) model



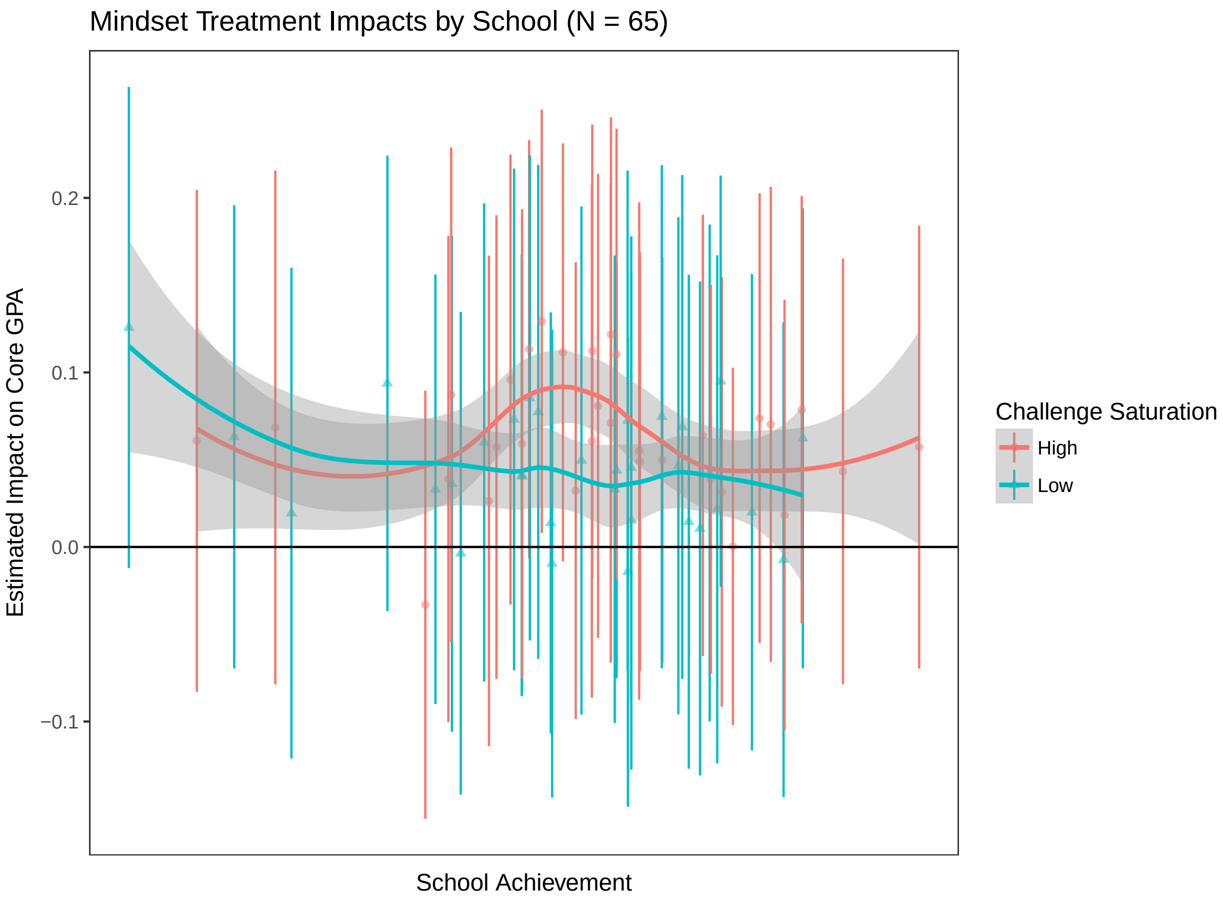
6.4 Variation in growth mindset intervention impacts on core GPA across 65 schools



Points and 95% confidence intervals report Empirical Bayes estimates.

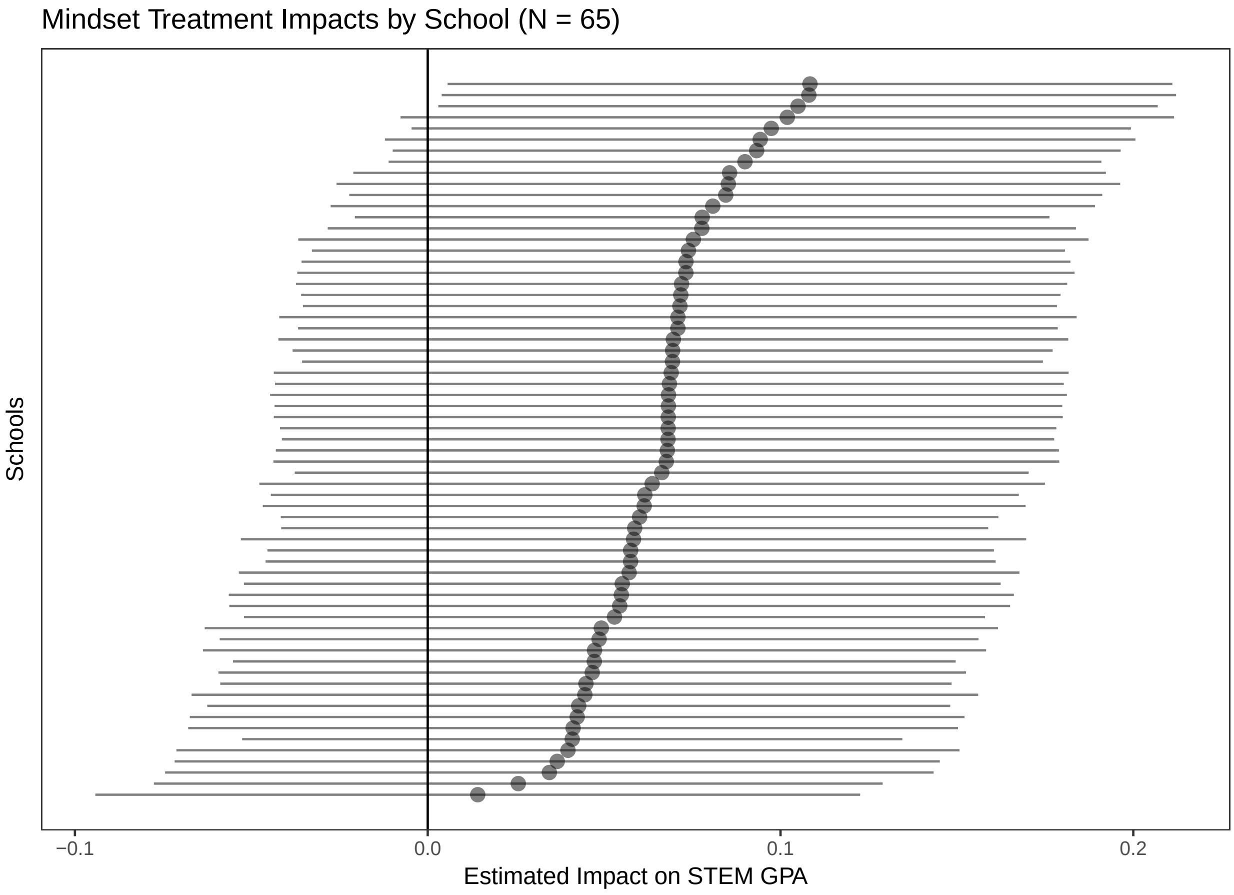


Points and 95% confidence intervals report Empirical Bayes estimates.

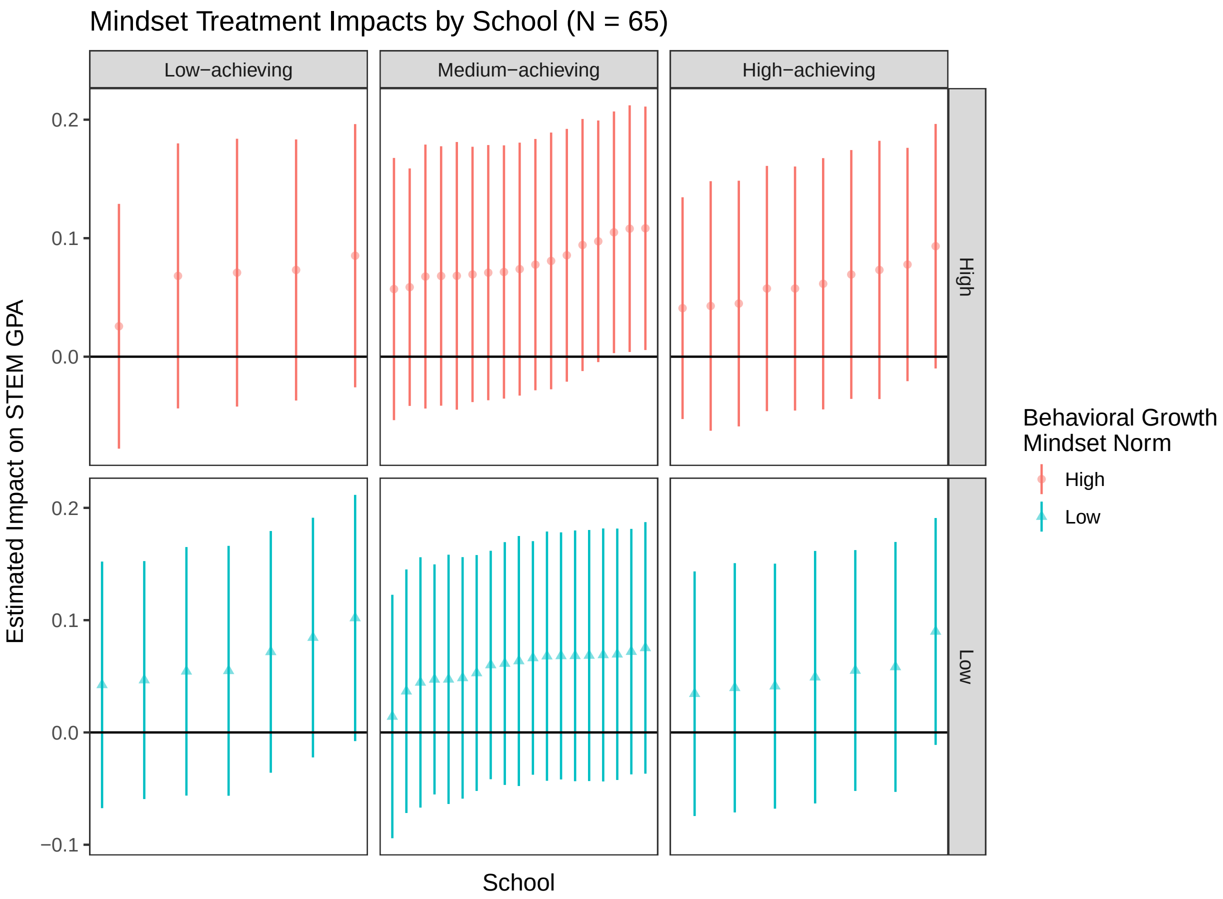


Points and 95% confidence intervals report Empirical Bayes estimates.

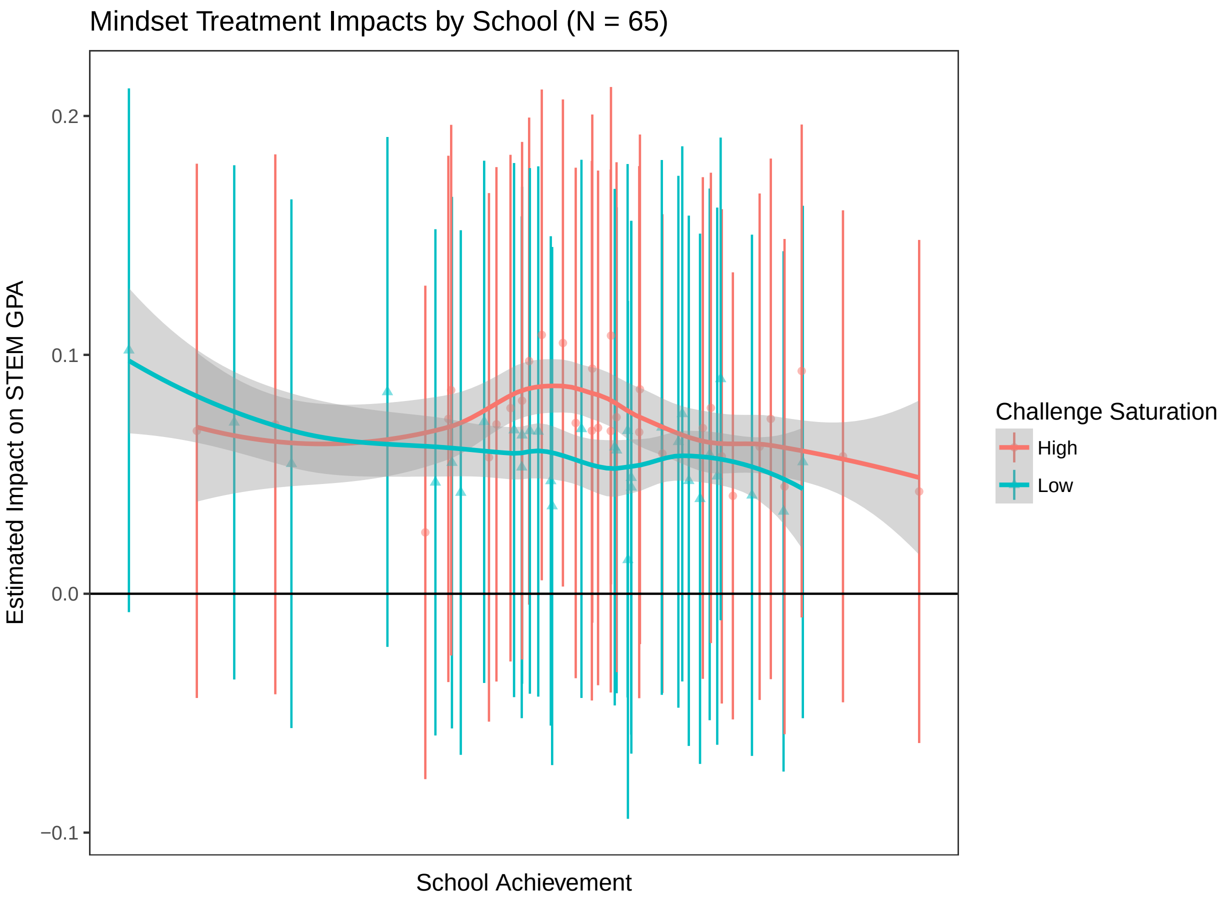
6.5 Variation in growth mindset intervention impacts on STEM GPA across 65 schools



Points and 95% confidence intervals report Empirical Bayes estimates.



Points and 95% confidence intervals report Empirical Bayes estimates.

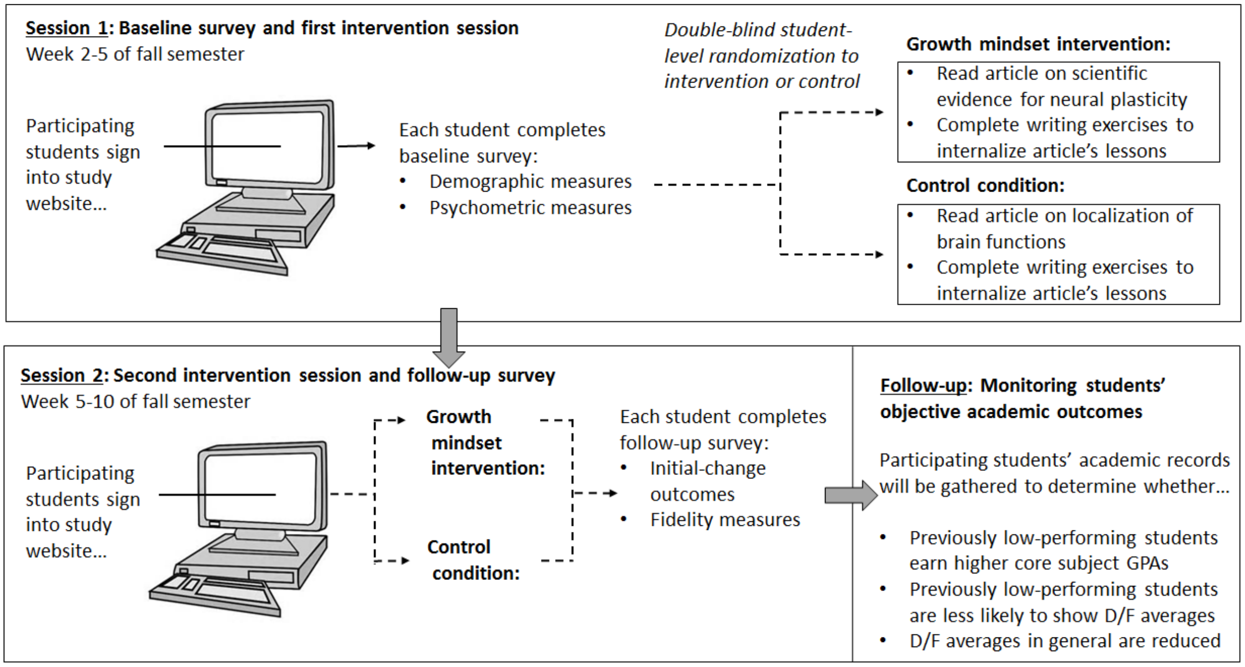


Points and 95% confidence intervals report Empirical Bayes estimates.

7 Data on Implementation of the Intervention in the National Sample

7.1 Overview of the student participation process

The participation process is depicted below in the Figure below. As it shows, the design of the study called for schools to deliver both sessions of the intervention to all students in the school in the fall of 2015, and for the two sessions to be roughly three weeks apart.



Session 1: Baseline survey and first intervention session
Week 2-5 of fall semester
Participating students sign into study website...
Each student completes baseline survey:
• Demographic measures
• Psychometric measures
Double-blind student-level randomization to intervention or control
Growth mindset intervention:
• Read article on scientific evidence for neural plasticity
• Complete writing exercises to internalize article's lessons
Control condition:
• Read article on localization of brain functions
• Complete writing exercises to internalize article's lessons
Session 2: Second intervention session and follow-up survey
Week 5-10 of fall semester
Participating students sign into study website...
Growth mindset intervention:
Control condition:
Each student completes follow-up survey:
• Initial-change outcomes
• Fidelity measures
Follow-up: Monitoring students' objective academic outcomes
Participating students' academic records will be gathered to determine whether...
• Previously low-performing students earn higher core subject GPAs
• Previously low-performing students are less likely to show D/F averages
• D/F averages in general are reduced

Figure 1: Study Overview.

Below, we present data on the actual timing of the intervention sessions in the schools.

The figures show that, overall, schools were quite compliant with the timing requests. Some schools, however, implemented the intervention in the spring. Moreover, students varied in how long they had in between sessions.

These descriptive statistics have three implications for our study. First, although our planned analysis was to use 8th grade GPA as the prior achievement variable and 9th grade fall and spring as the outcome variable, in some schools it was necessary to use fall 9th grade as the prior achievement variable. In sensitivity analyses, we examine whether or not the choice of a prior achievement term affects our conclusions.

Second, note that some students received the treatment very late in the year, and it was not uncommon for spring-implementation schools to deliver the second half of the treatment well into March–just two months before the school year was over. This necessarily limits the potential for the treatment to affect their grades. This problem is especially acute when considering that some schools only provided a year-end grade, not broken out by semester. Therefore some students' outcomes were already mostly determined before random assignment. Thus, the treatment effect sizes in the study are conservative relative to what might be gained with ideal timing of implementation.

Third, the relatively high rate of compliance overall testifies to the scalability of the treatment and to the effectiveness of the study procedures designed by the reserach team and by the independent data collection firm ICF International.

7.2 Student survey response rates were quite high

The table below depicts descriptive statistics for the proportion of eligible students in the school who were in the intent-to-treat sample. The few cases with very low response rates came from schools that required signed consent from parents.

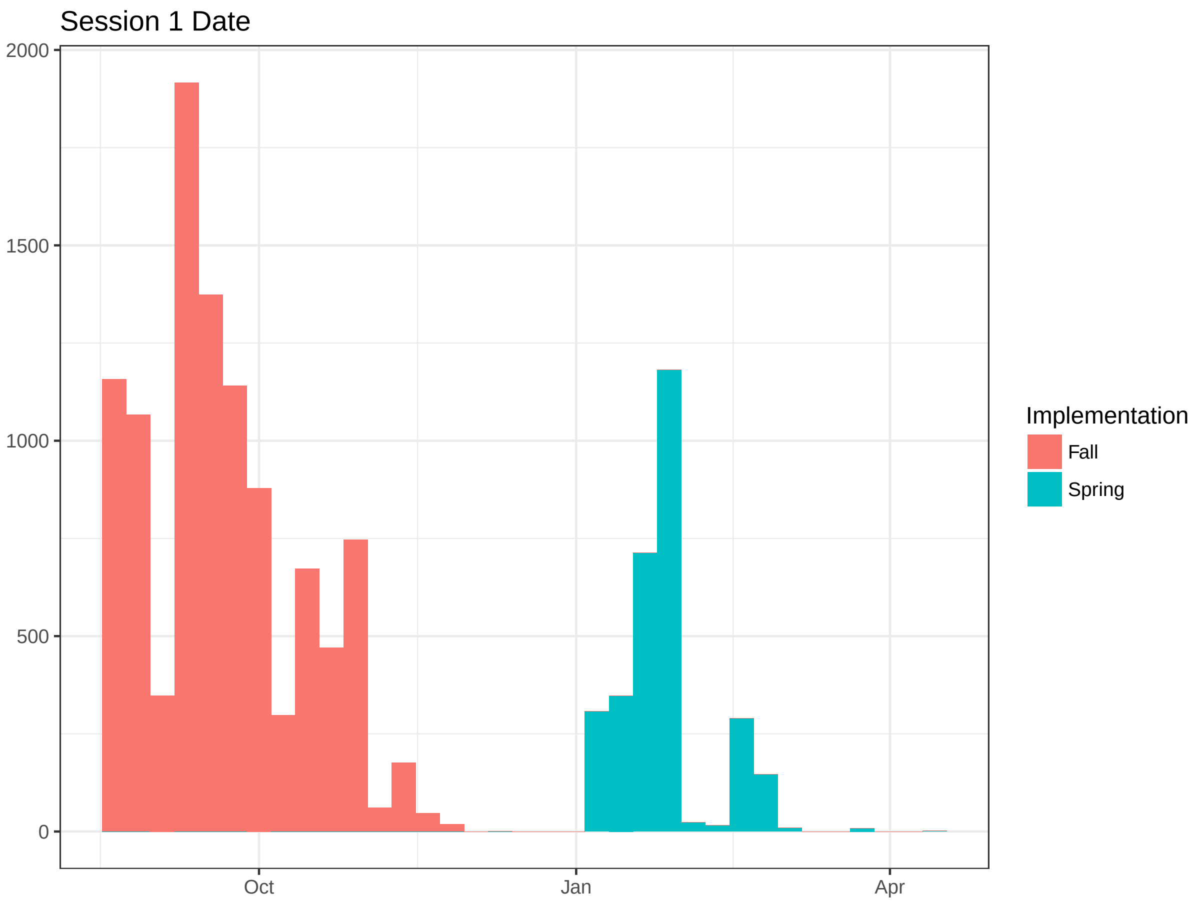
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.9400  0.9800  0.9346  1.0000  1.0000
```

7.3 Timing of intervention sessions within the school year

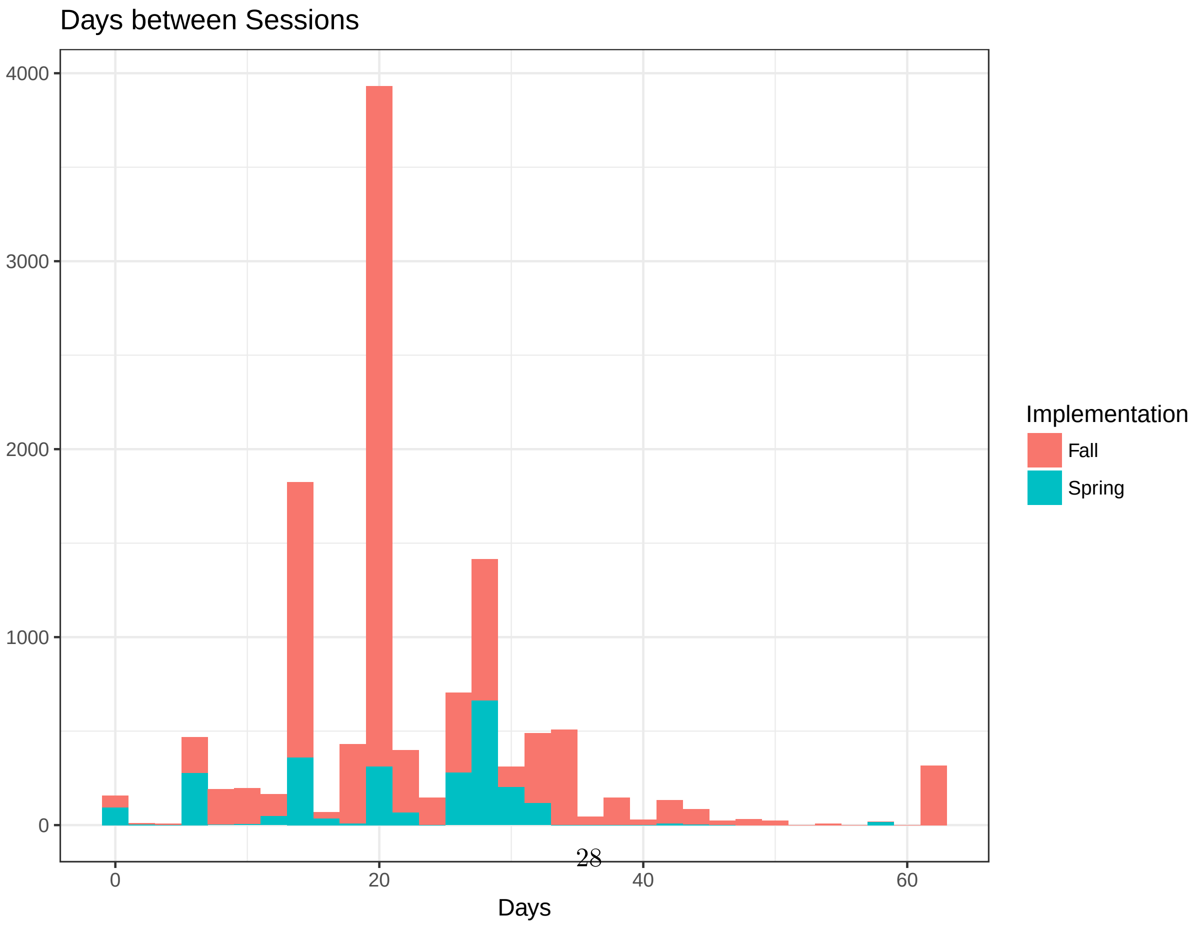
Three quarters of schools implemented the intervention in the fall, as planned. Most schools were able to follow the request to space the two treatment sessions 3 to 4 weeks apart.

Implementation	Schools	School Prop	Students	Student Prop
Fall	54	0.818	10372	0.773
Spring	12	0.182	3049	0.227

Implementation	Proportion	S1 Median	S1 SD	S1 N	S2 Median	S2 SD	S2 N	S1-S2 Med (days)	S1-S2 SD
Fall	0.773	2015-09-17	22.431	10372	2015-10-12	22.951	9775	21	10.771
Spring	0.227	2016-01-26	13.114	3049	2016-02-18	15.901	2520	27	9.374



Session 1 Date

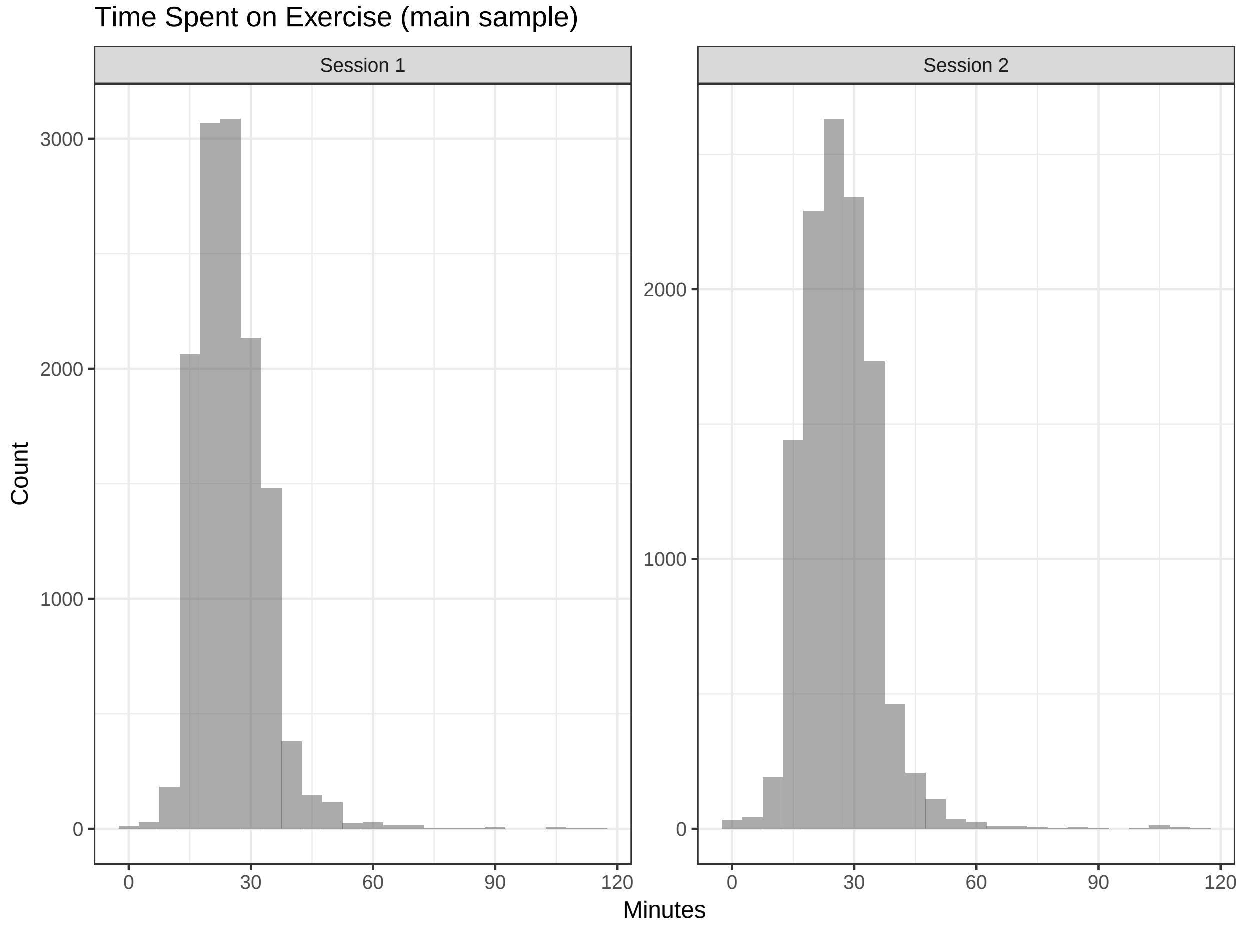


Days between Sessions

7.4 How long did students spend on the intervention exercises?

Each treatment session took students about 25 minutes on average, for a total of 50 minutes overall. This is notable because the primary analyses are looking for effects of this 50-minute experience on grades across all core classes at the end of the school year, sometimes many months later.

The control group was shorter in session 1, corresponding to someone less content than in the treatment condition. The control group was a longer in session 2 because students answered extra questions about the classroom and school climate.



```
##    s1_minutes       s2_minutes
##  Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 19.67   1st Qu.: 21.83
##  Median : 24.03   Median : 26.22
##  Mean   : 25.36   Mean   : 26.70
##  3rd Qu.: 28.40   3rd Qu.: 30.58
##  Max.   :113.60   Max.   :113.60
##  NA's   :623      NA's   :1826
```

Table 24: Session time by experimental group

Variable	Trt Mean	Trt SD	Trt N	Ctl Mean	Ctl SD	Ctl N	Diff	p	Pooled SD	d
s1_minutes	26.413	8.681	6334	24.325	7.898	6464	2.088	0	8.299	0.252
s2_minutes	25.860	9.109	5812	27.546	9.302	5783	-1.686	0	9.206	-0.183

Variable	Trt Mean	Trt SD	Trt N	Ctl Mean	Ctl SD	Ctl N	Diff	p	Pooled SD	d
## Session Co	mpletion									

	.
Started Session 1	1.000
Finished Session 1 (of all)	0.869
Started Session 2 (of all)	0.916
Finished Session 2 (of S2 starters)	0.792

Status	Trt Mean	Trt SD	Trt N	Ctl Mean	Ctl SD	Ctl N	Diff	p	Pooled SD	d
Started Session 1	1.000	NA	6702	1.000	NA	6719	0.000	0.890	0.000	NaN
Finished Session 1	0.847	NA	6702	0.891	NA	6719	-0.044	0.000	0.337	-0.131
Started Session 2	0.915	NA	6702	0.918	NA	6719	-0.003	0.525	0.277	-0.012
Finished Session 2	0.794	NA	6129	0.790	NA	6166	0.004	0.574	0.406	0.011

8 Effects of the Intervention on Self-Reported Mindset

In supplementary analyses, we asked: beyond grades, did the treatment increase motivation and course-taking?

There were effects of the intervention on reporting a growth mindset (>4 on the 1-6 growth mindset scale) for all students:

Mindset treatment impact on fixed mindset scale

Sample	estimate	stderr	t	dof	p	n	n_schools	ctl_mn	ctl_sd
Full Sample	-0.418	0.022	-19.010	46	0	11350	65	2.897	1.168
Lower-achievers	-0.381	0.045	-8.489	46	0	5507	65	3.076	1.194
Higher-achievers	-0.437	0.045	-9.707	46	0	5843	65	2.721	1.115

Mindset treatment impact on growth mindset (indicator)

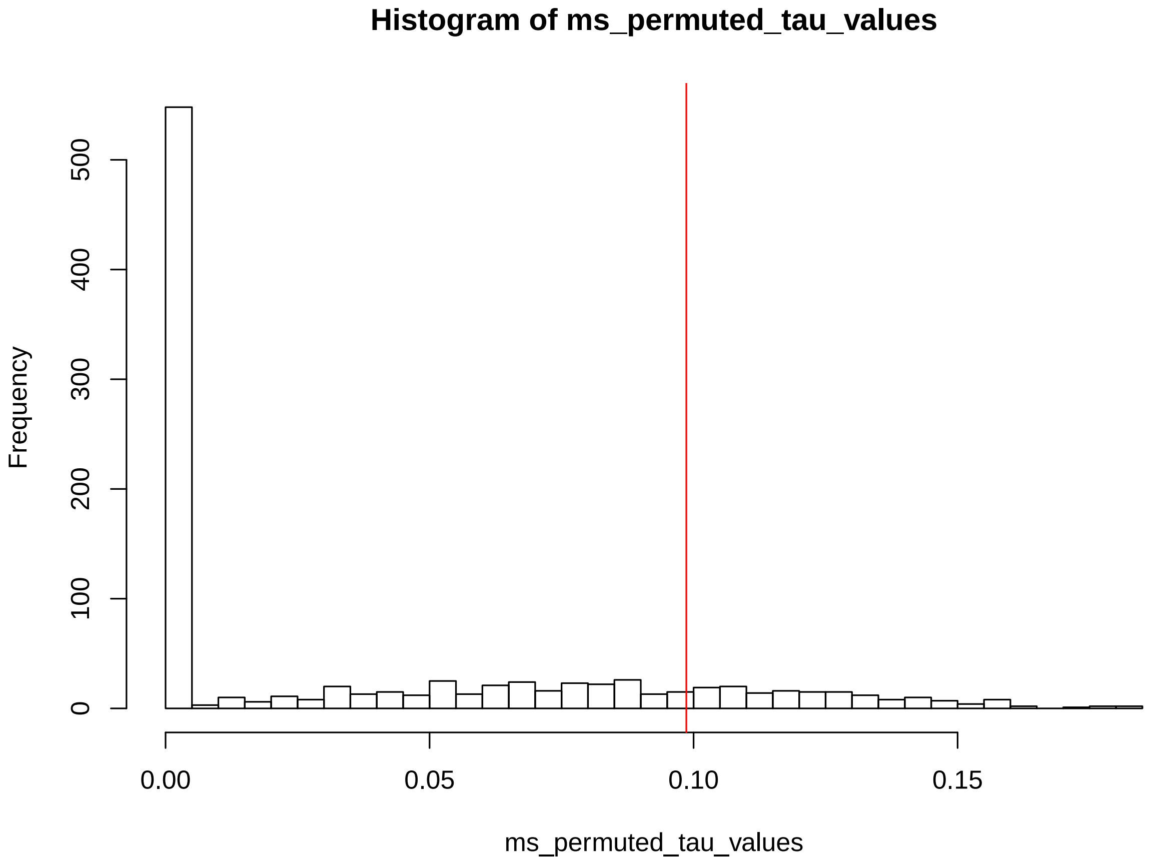
Sample	estimate	stderr	t	dof	p	n	n_schools	ctl_mn	ctl_sd
Full Sample	0.158	0.009	17.179	46	0	11350	65	0.511	0.500
Lower-achievers	0.147	0.016	8.962	46	0	5507	65	0.444	0.497
Higher-achievers	0.163	0.015	11.178	46	0	5843	65	0.577	0.494

8.1 School Variation in Impacts on Fixed Mindset Scale

Summary of School Heterogeneity Estimates for Growth Mindset Outcome among Low-achieving Students

Mean.Impact	Impact.SE	Impact.t	tau	tau.p.value.Q.method	tau.p.value.permutations
-0.377	0.033	-11.269	0.099	0.443	0.247

The distribution of the estimate of tau across 999 permutations, with the observed value marked, is:



8.2 Observational association between pre-treatment fixed mindset and advanced mathematics in 10th grade among control students

Two observational analyses: bivariate correlation, and regression model predicting Algebra II or higher mathematics with fixed mindset and control variables.

```
## 
##  Pearson's product-moment correlation
## 
## data:  analysis_df[analysis_df$treatment == 0, ][["s1_fixedmindset3"]] and analysis_df[analysis_df$t
## t = -10.94, df = 3653, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
```