A Scalable, Theory-Based Intervention to Influence Teachers' Student Engagement Practices Improves Academic Performance in a State-Wide Sample

David Yeager

dyeager@utexas.edu

University of Texas at Austin https://orcid.org/0000-0002-8522-9503

Cameron Hecht

Department of Psychology, University of Rochester https://orcid.org/0000-0003-4842-6003

Christopher Bryan

University of Texas at Austin https://orcid.org/0000-0002-1582-4411

Matthew Giani

Department of Sociology, University of Texas at Austin

Robert Crosnoe

The University of Texas at Austin

Chandra Muller

University of Texas at Austin

Elizabeth Tipton

Northwestern University https://orcid.org/0000-0001-5608-1282

Jared Murray

McCombs School of Business, University of Texas at Austin

Ryan Smith

Texas Behavioral Science and Policy Institute at the Population Research Center, University of Texas at Austin

Shuqi Zhang

Texas Behavioral Science and Policy Institute at the Population Research Center, University of Texas at Austin https://orcid.org/0000-0003-0459-0688

Meghann Johnson

Alpha Motivation Lab

Megan Smith

Alpha Motivation Lab

Jenny Buontempo

Texas Behavioral Science and Policy Institute at the Population Research Center, University of Texas at Austin

Rebecca Boylan

Texas Behavioral Science and Policy Institute at the Population Research Center, University of Texas at Austin

Shannon Green

Texas Behavioral Science and Policy Institute at the Population Research Center, University of Texas at Austin

Mary Murphy

Apple University and Equity Accelerator

Carol Dweck

Department of Psychology, University of Texas at Austin

Article

Keywords:

Posted Date: December 17th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-8302846/v1

Additional Declarations: Yes there is potential Competing Interest. None of the authors owns a financial stake in any growth mindset product or service for students or teachers. The FUSE program is now available (after evaluation) for purchase at cost to school districts via the University of Texas at Austin, a non-profit academic institution. None of the authors receives pay, owns equity, or has any arrangement for future profit from the FUSE program, and all funds raised from the FUSE program go to support the research and training mission at UT Austin.

A Scalable, Theory-Based Intervention to Influence Teachers' Student Engagement Practices Improves Academic Performance in a State-Wide Sample

Authors

David S. Yeager*[1,2,3], Cameron Hecht[4], Christopher Bryan*[2,5], Matthew Giani*[2,6], Robert Crosnoe*[2,6], Chandra Muller*[2,6], Elizabeth Tipton*[7], Jared Murray*[5], Ryan Smith[2], Shuqi Zhang[2], Meghann Johnson[3], Megan Smith[3], Jenny Buontempo[2], Rebecca Boylan[2], Shannon Green[2], Mary C. Murphy*[8,9], & Carol S. Dweck*[10].

[1] Department of Psychology, University of Texas at Austin; [2] Texas Behavioral Science and Policy Institute at the Population Research Center, University of Texas at Austin; [3] Alpha Motivation Lab; [4] Department of Psychology, University of Rochester; [5] McCombs School of Business, University of Texas at Austin; [6] Department of Sociology, University of Texas at Austin; [7] Statistics for Evidence-Based Policy and Practice Center, Northwestern University; [8] Department of Psychological and Brain Sciences, Indiana University; [9] Apple University and Equity Accelerator; [10] Department of Psychology, Stanford University. *Co-Investigators for the FUSE evaluation project. Contact: David Yeager (yeagerds@austin.utexas.edu).

Abstract

What can be done about the crisis of student disengagement? A theoretical analysis led to a novel behavioral intervention that motivated and coached teachers to cultivate a classroom "culture of learning." This was intended to contradict the typical, disengaging "culture of judgment and evaluation" in secondary schools. The program, called the Fellowship Using the Science of Engagement (FUSE), influenced teachers' light-touch practices and involved changes to classroom language and communication around student mistakes, confusion, and grades. In doing so, the FUSE program honored teachers as collaborators (rather than passive recipients of "expert wisdom") and connected them virtually with other teachers in the program and with coaches who were also practicing educators. FUSE was evaluated in 6th to 9th grade math classes in a diverse, state-wide sample of 80 Texas public schools (*N* = 152 teachers; *N*=12,432 students). A control group (randomly assigned) received a version of the FUSE program that taught principles of cognitive science and how to apply them to teachers' math instruction. The pre-registered, conservative, Bayesian analysis showed that the FUSE treatment program changed teachers' beliefs and behaviors (average treatment effects from .14 to .51 *s.d.*), led to an estimated effect on student math performance equivalent to an additional 4 months of student learning, and reduced the proportion of teachers who reported feeling "burnt out" by half, while improving teacher well-being by .25 *s.d.* This is the only known teacher intervention to influence teacher behavior, student performance, and teacher well-being longitudinally in a pre-registered randomized trial that was conducted in a scaled-up manner. Given the relatively low cost of the program (~$25 per student per year) this study highlights the ability of behaviorally-informed interventions to influence teachers' subtle, culture-building practices and points to their role as an important route to educational improvement.

INTRODUCTION

Academic disengagement threatens to make global education systems less effective, damaging the occupational and developmental trajectories of youth around the world [1–3]. Disengagement includes such things as avoiding intellectual challenges or failing to learn from academic mistakes. Large surveys find that 72% of secondary students give up when academic work gets hard[2] and fully 63% prefer easier assignments that they can do well on without really thinking (vs. harder assignments that could teach them something new)[4]. Add to this the fact that 19% of students globally have missed school for three or more months because they simply did not like school[5]. Furthermore, student engagement is the leading predictor of burnout among teachers—a major global problem.[6]

Conventional approaches to increasing student engagement, such as the adoption of new curricula or educational technologies, are lengthy and require major investments of money and effort—and are generally ineffective[1,7]. For example, the adoption of new curricula tends to require costly teacher professional learning programs, but these programs tend to yield null, negligible, or highly variable impacts on student learning[8,9]. Meanwhile advances in educational technology have often been touted as solutions to the engagement crisis, but these apps tend to have unproven effectiveness and more of them in schools is unlikely to solve the problem[10]. Indeed, the average school district in the U.S. already uses over 1,500 different educational applications—a number that continues to grow[10]—and the problem of disengagement has only worsened over the last decade[11].

Here we evaluated a novel teacher professional learning program that was relatively short and inexpensive—and had demonstrable effects at scale. This program did not require changes to the curriculum. Instead it focused on changes to *classroom culture*, which is shaped by behaviors that were almost entirely under teachers' control, such as how teachers talk to their students about challenges, struggles, mistakes, or grades[12–17]. This teacher professional learning program could, in principle, be delivered entirely virtually, and, as shown below, it impacted student engagement and performance in a state-wide sample of 37 school districts. This makes the program more readily scalable than conventional reforms.

The present research advanced the literature by applying theoretical insights from behavioral science (see Extended Data Table 2) to yield meaningful impacts on both students and teachers in a rigorous, pre-registered evaluation study[18,19]. The study also has implications for organizational cultures in general. Many workplaces face problems with an organizational culture that undermines performance, such as high-tech corporations, service and retail organizations, competitive athletics, and more [12,17,20–22]. In each context, a culture of learning could promote innovation and increase employee satisfaction and well-being[12,21]. The current research could therefore serve as a model for how professional learning programs could create such a culture.

The Culture of Learning Program

The Fellowship Using the Science of Engagement (FUSE) evaluated here is a year-long program that sought to shift secondary math classroom cultures by means of targeted professional learning experiences that honored teaches as "fellows" and provided them with a peer support network to guide implementation. (See Figure 2, Extended Data Table 2, and Extended Data Figure 5).

Background. The FUSE program focused on creating *culture of learning* classrooms, which are classroom cultures that convey the growth mindset belief that all students, even struggling students, can learn, grow, and succeed[13]. This focus came from our theoretical analysis[12–17] of the causes of student disengagement, and of teacher practices that could promote engagement (Figure 1). In a culture of learning, students are led to embrace challenges, correct their mistakes, and improve their knowledge and expertise over time [12–17]. Such a culture works because it contradicts the *culture of judgment and evaluation* that students typically encounter (compare Panel A in Figure 1 to Panel B) [12–17]. Many classroom cultures foster disengagement by sending evaluative messages to students—messages about measuring, ranking, and sorting students based on their performance on classroom work, homework, and tests. In such classrooms, students may be fearful of exposing ignorance (or feeling ignorant). They may therefore be reluctant to engage in their learning with the spirit of openness and vulnerability that is necessary for confronting confusion and learning from mistakes. This reluctance can start a cycle of poorer performance that evokes even more negative evaluation from teachers, accelerating students' disengagement from that judgmental culture (see Figure 1).

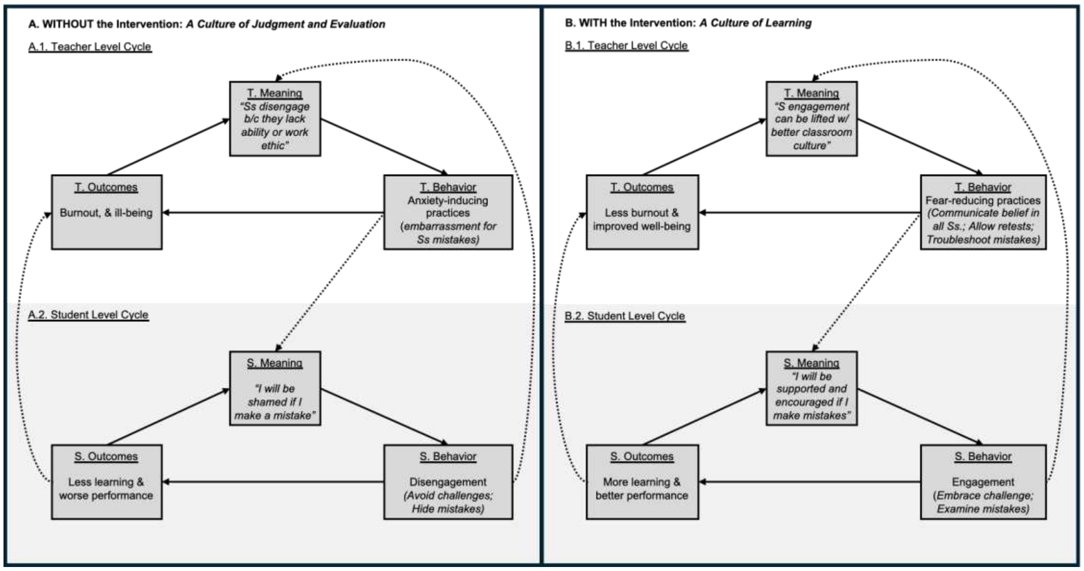

A. WITHOUT the Intervention: A Culture of Judgment and Evaluation
A.1. Teacher Level Cycle
T. Meaning
"Ss disengage b/c they lack ability or work ethic"
T. Outcomes
Burnout, & ill-being
T. Behavior
Anxiety-inducing practices (embarrassment for Ss mistakes)
A.2. Student Level Cycle
S. Meaning
"I will be shamed if I make a mistake"
S. Outcomes
Less learning & worse performance
S. Behavior
Disengagement (Avoid challenges; Hide mistakes)
B. WITH the Intervention: A Culture of Learning
B.1. Teacher Level Cycle
T. Meaning
"S engagement can be lifted w/ better classroom culture"
T. Outcomes
Less burnout & improved well-being
T. Behavior
Fear-reducing practices (Communicate belief in all Ss.; Allow retests; Troubleshoot mistakes)
B.2. Student Level Cycle
S. Meaning
"I will be supported and encouraged if I make mistakes"
S. Outcomes
More learning & better performance
S. Behavior
Engagement (Embrace challenge; Examine mistakes)

Figure 1. A simplified process model of how teacher and student meaning-making and behaviors mutually influence classroom culture and outcomes (A) without the intervention (Culture of Judgment and Evaluation), and (B) with the intervention (Culture of Learning). Note. The theoretical model was generated from insights integrated from previous research [12–17] and initially tested in a nationally-representative survey study we conducted (Extended Data Table 1). Classroom culture is the result of teacher meaning-making tendencies (e.g. beliefs and situational appraisals), leading to visible teacher behaviors, which shape student meaning-making tendencies, behaviors, and outcomes [23]. These student-level processes in turn influence teacher outcomes and reinforce teacher meaning-making tendencies. This theory suggests that an intervention program's targeted shifts in teachers' meaning-making and behaviors—at critical times, such as early in the school year—might start a new, self-reinforcing process that results in better outcomes for students and teachers over time (Panel B).

Practices targeted by the program. In the FUSE program, we provided teachers with professional learning that helped them to communicate a culture of learning (and overturn the culture of judgment and evaluation) through relatively short, targeted messages to students (see Table 1 and Extended Data Figure 5). These messages conveyed that all students, regardless of their past performance, can learn, improve, and do well if they work hard and seek appropriate help from teachers and peers. Examples of these messages include the ideas (a) that students' mistakes provide useful information to teachers about what students need and about how the teachers can be more effective in their teaching, (b) that test scores represent students' performance at one moment in time and, if not optimal, need not determine how well they can perform in the future, and (c) that the teacher has rigorous but flexible classroom grading policies if students correct their mistakes and show evidence of mastering the material (via retesting or revision). (see Supplemental Online Materials).

Treatment Group (Culture-of-Learning)	**Control Group (Cognitive-Science-of-Learning)**
Honorific framing as a "fellowship"	
Keynote lecture on the science of adolescent development (e.g. brain, hormones, motivation)	Keynote lecture on the science of adolescent attention and memory (e.g., encoding, storage, retrieval)
Practice 1: "Mining" mistakes as learning opportunities	Practice 1: Frequent testing and knowledge maps
Practice 2: Retesting and revision policies	Practice 2: Spaced and interleaved practice
Practice 3: Culture-of-learning speeches	Practice 3: Engaging brains with prediction
Practice 4: Collaborative troubleshooting routines	Practice 4: Spaced practice: A closer look
Practice 5: Exam speeches	Practice 5: Optimizing memory encoding
Student modules: Growth mindset and stress-can-be-enhancing mindset	Student modules: Cognitive science of learning and memory
Reports on student survey data	
Coaching from experienced teachers and former fellows	
Peer networks with other current fellows	
Final "showcase" presentation	

Table 1. Comparison of the treatment (culture-of-learning) and control (cognitive-science-of-learning) interventions. Cells shaded darker grey were unique to the treatment (culture-of-learning) group. Cells without shading were unique to the control group. Cells shaded lighter grey were in consistent across the two conditions.

Teachers were taught to deliver these messages about their beliefs and policies at the beginning of the year, when the classroom culture was first established, and also at certain culture-defining moments, such as before and after a major assessment. This is because these are moments when students are looking for clues about how to interpret the meaning of current and/or future difficulty[24]. Importantly, although teachers in the culture-of-learning program were encouraged to be more supportive of student learning, they were not encouraged to make their courses any less rigorous or demanding, or to give students high grades without students earning them. Instead, they were taught how to maintain high standards while increasing the rate at which all students met those standards[17].

How the program changed behavior. To change teachers' behavior and address the "intention-to-action gap," that undermines so many behavioral programs [25], the first step was to reorient teachers' beliefs (or *mindsets*) to motivate teachers to adopt novel practices (see Figure 2). This goal came from pilot research that we conducted in a nationally-representative sample of teachers (*N*=980; see Extended Data Table 1), which showed that teachers' *fixed mindsets* (their beliefs that students' abilities were fixed and cannot change) were associated with (a) more judgmental interpretations of student mistakes, (b) a reduced likelihood of using culture of learning practices, and (c) more teacher language that communicated a culture of judgment and evaluation. Those pilot data led us to hypothesize that a first step in our program would be to promote more of a *growth mindset*—the belief that all students can learn and grow. Without ever mentioning fixed or growth mindsets, the FUSE program drew on behavioral science, and in particular on the social psychology of belief and behavior change [26–31], to shift teachers' mindsets (see Extended Data Table 2).

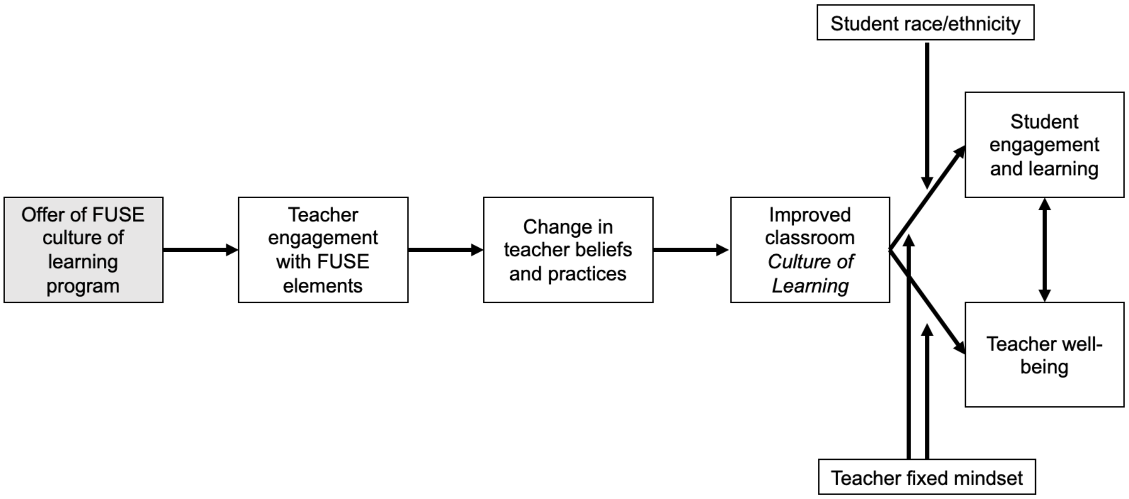

Student race/ethnicity
Offer of FUSE culture of learning program
Teacher engagement with FUSE elements
Change in teacher beliefs and practices
Improved classroom *Culture of Learning*
Student engagement and learning
Teacher well-being
Teacher fixed mindset

Figure 2. A simplified theory of change for the impact of the Fellowship Using the Science of Engagement (FUSE) on student and teacher outcomes, moderated by student and teacher factors.

The current study. In the present research, FUSE was evaluated in a randomized experiment conducted in a diverse, state-wide sample of schools with a pre-registered analysis plan. The sample of schools was first recruited to represent the population of schools in Texas (see Methods). All regular math teachers in grades six to nine in the participating schools were invited to join the FUSE program. Among teachers who elected to join the program, individuals were randomly assigned to a treatment (culture-of-learning) group or an active control (cognitive-science-of-learning) group (see Table 1). Because FUSE is a teacher-level treatment whose goal is to shift *classroom* culture and improve classroom outcomes, which is defined at the aggregate level, then all of the students nested under treatment teachers received the same condition (and treatment modules) and all students under control teachers received the control modules (see Figure 3) [32].

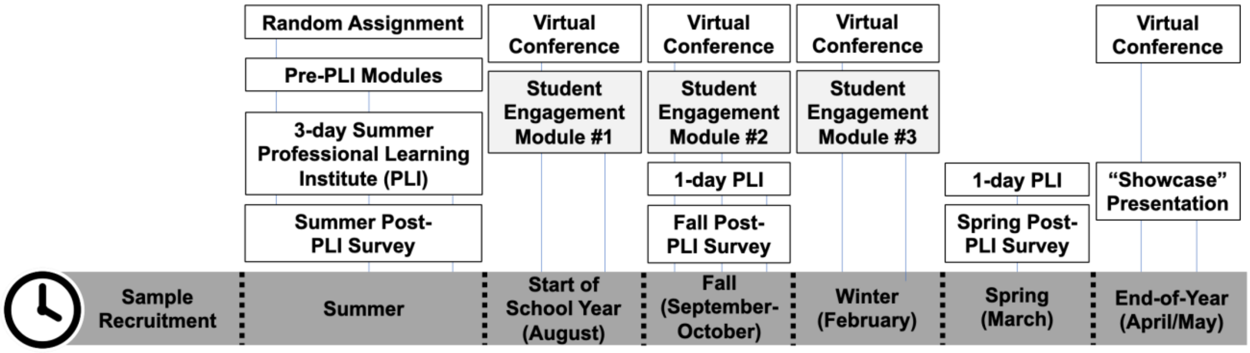


Figure 3. Timeline of teacher and student activities in the Fellowship Using the Science of Engagement (FUSE). Note: PLI = Professional Learning Institute. Light-gray shaded boxes represent student learning modules and data collection activities; the remaining boxes represent teacher-facing activities. All "modules" were online reading and writing exercises completed asynchronously. During the 3-day summer institute teachers were introduced to the first three practice modules; the fourth and firth practice modules were introduced during the Fall and Spring PLIs, respectively. Virtual conferences are 1 hour virtual (online) small-group coaching meetings involving FUSE fellows and near-peer coaches.

A focus on mathematics. We conducted the present research in secondary math classrooms (grades six to nine in the U.S.)[33] because students in this age range often report that math is hard and that they fear the judgment that comes from making mistakes in math [34,35]. When asked about math, students often agree, for example, that *"I get anxious when asking questions in class because I don't* *want to look stupid…"* [35]. In western culture, math performance is often portrayed as a reflection of true intellectual ability [36,37], which may cause students even greater fear of making mistakes that could reveal a lack of ability. These fears can be enhanced during the adolescent years due to the onset of pubertal maturation, which is known to sensitize the brain to social emotions such as embarrassment or shame[7,17,38]. Thus, a culture of learning could be a welcome antidote to the culture of judgment and evaluation in math that adolescents typically encounter. Further, the math curriculum in the U.S. is highly structured and sequenced, such that progress in one year can powerfully determine the available opportunities in subsequent years[39,40]. Thus a math classroom culture of learning that breaks a negative cycle could potentially have long-term beneficial effects[33].

Control condition. The control group learned instructional practices that aligned their pedagogy with insights from the cognitive science of learning and memory, such as interleaving and spaced repetition[41]. (See Table 1, Extended Data Table 2 and the Supplemental Online Materials) This control was chosen because cognitive science has demonstrated impacts on learning and performance, mostly in laboratory experiments in which researchers controlled the instructional materials and assessments. Such evidence led cognitive science tools to become popular in educational reform[42]. Although studies have not yet found that training teachers on cognitive science techniques improves student performance outcomes, this control condition remains a relatively high bar for the treatment group to clear. The program for both the treatment and control groups included key elements of effective professional learning programs, such as: (1) presenting information in a way that helps teachers (who are novices in these new techniques) develop expertise in them [41,43]; (2) live coaching (via virtual conferences) from experienced math teachers [44–46]; and (3) peer networks of other FUSE fellows [47] (see Table 1 and Figure 3).

Teacher well-being. Our theoretical analysis (see Figure 1) proposes that student disengagement would jeopardize teacher well-being. National surveys have found that a disengaged classroom is teachers' top complaint—and a leading source of psychological distress.[6] In particular, we hypothesized that teachers with more of a fixed mindset—whose students reported feeling more disrespected (see Extended Data Figure 3), and whose students were less-engaged—would typically show worse well-being (e.g. greater burnout, lower life satisfaction). The treatment, however, should be especially impactful at lifting teacher well-being for previously fixed-mindset teachers. Thus, we pre-registered measures of well-being as well as prior fixed mindset as a moderator (see Figure 2).

Prior research. Promising initial evidence for the efficacy of the FUSE materials for belief change came from two previous RCTs (one published[48] and one reported in Extended Data Figure 1). However, because the impacts of the fellowship, as a whole, have not yet been extended to test performance or to potential impacts on teacher well-being, a more definitive, pre-registered experimental evaluation was needed. Thus, we conducted the present research.

Bayesian, Machine-Learning Statistical Modeling

Pre-registered analyses were conducted using the *stochtree* package [49], which implements the Bayesian Causal Forest (BCF) model [50]—a popular method for evaluation studies of these kinds of interventions[51–53]. As a fully Bayesian method, BCF does not output conventional p-values used in frequentist methods. It instead outputs vectors of draws from the posterior distribution. Therefore we report posterior probabilities continuously[54–56] and follow our pre-registered criteria for interpreting the strength of the evidence (e.g. only interpreting effects with greater than .75 posterior probability of a difference), while placing greater emphasis on effect size than on null-hypothesis-rejection, per statistical guidelines [54]. Table 2, Figures 3 to 9, and Extended Data Figures 2 and 3 report the average treatment effects (ATEs) and conditional average treatment effects (CATES).

Effect Size Calculations

Standardized mean differences (*s.m.d.'s*; i.e. effect sizes) are reported in terms of between-teacher standard deviations (*s.d.'*s) because (a) randomization was conducted at the cluster (teacher) level, (b) research questions focused on teacher-level impacts, and (c) the measures were designed to provide reliable estimates at the teacher level, not at the student level (see Methods). The achievement test outcomes, however, are presented in terms of student-level standard deviations, to facilitate comparisons to achievement effects in the literature[57], even though in the present study these are known to be conservative (due to planned, random measurement error in our study design). As a secondary way of presenting effect sizes for student achievement outcomes, we standardized effects in terms of expected months of academic learning, using published grade-level benchmarks.[58]

RESULTS

Equivalence of Conditions

Teachers in the treatment and control conditions were similar in terms of demographic characteristics and classroom composition (see the Supplemental Online Material). Next, the treatment and control groups reported equivalent (and very high) ratings of the likelihood that they would recommend the program to a friend or colleague (on a scale of 0 to 10): Treatment mean (*m*) = 8.68, Control *m* = 8.62, *s.m.d.* = 0.04 *s.d.*, posterior probability that the ATE is greater than 0, pr(ATE>0)<.75. The conditions were also equivalent with respect to the number of PLI events attended synchronously, $\chi^2(3)=3.87$, $p=.27$, and in terms of the number of virtual conferences attended, $\chi^2(4)=4.58$, $p=.33$. Accordingly, both conditions influenced teachers' intentions to implement the associated practices (see Figure 4 and Extended Data Table 3). Thus, the study design conservatively tested the impact of the culture of learning practices compared to practices based on findings from cognitive science.

Manipulation Checks and Preliminary Analyses

We first checked whether the FUSE program changed teachers' mindsets and intended behaviors, which are the first steps in the theory of change (see Figure 2). This was important to demonstrate as a preliminary matter because if teachers' beliefs and intentions had *not* been influenced, it would have suggested serious problems with the persuasive approach of the FUSE treatment.

Consistent with our theory of change (Figure 2), at post-test the FUSE treatment program reduced teachers' reports of their *fixed mindsets* (the belief that student math ability is fixed and cannot change), *s.m.d.* = -0.33 *s.d.,* pr(ATE>0)<.995. (Also see impacts on related mindsets in Figure 4). In addition, treated teachers reported high levels of intention to implement the culture of learning practices (*M*=4.58 out of 5) and these levels were higher than the control group's already-high intention to implement those same practices, (*M*=4.33 out of 5), *s.m.d.* = 0.47 *s.d.*, pr(ATE>0)>.999. This suggested that the program successfully motivated teachers at the end of the initial three-day learning institute. Notably, because many interventions in the literature fail to change people's personally-held behaviors and their behavioral intentions, these results suggest that FUSE was able to overcome the barriers that block the theory of change for many professional learning programs.

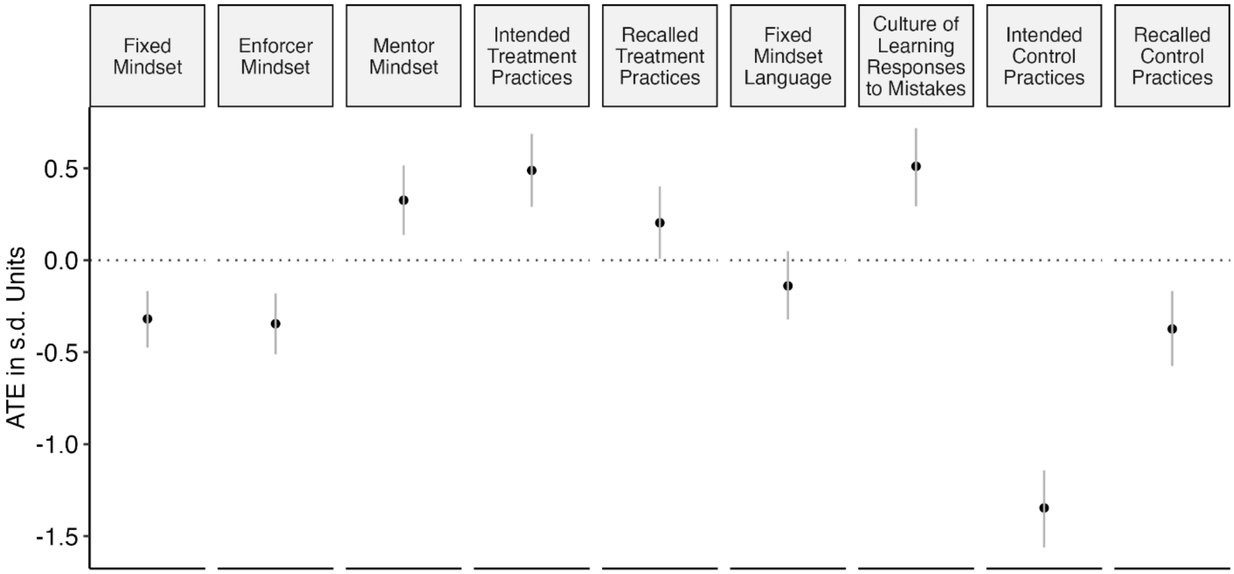


Figure 4. Teacher beliefs, intended behaviors, and behavior average treatment effects (ATEs) for the FUSE culture-of-learning program compared to the control program. Figure depicts the average treatment effects (ATEs) in pooled between-teacher *s.d.* units and the intervals from the 10th to the 90th percentiles. See Extended Data Table 3 for point estimates. The results are from BCF models, which include inverse probability weights (IPWs) to adjust for missing data/attrition. Sample items: *Fixed mindset*, "People have a certain amount of intelligence, and they really can't do much to change it;" *Enforcer mindset*, the belief that adolescents need excessively high standards and little support, "Most adolescents are incapable of behaving correctly unless there are severe consequences for disobedience;" *Mentor mindset*, the belief that adolescents need both high standards and high support, "Adolescents can display remarkable self-discipline if the classroom environment properly motivates them;" *Treatment (culture-of-learning) practices*, "Spend class time encouraging students to reveal and discuss their mistakes with the teacher or classmates;" *Fixed mindset language,* "How effective if this statement? "I will tell you the right facts and techniques for solving each kind of problem. It will be up to you to remember them;" *Culture of learning responses to mistakes:* See Extended Data Table 4; *Control (cognitive-science-of-learning*) practices, "Increase the frequency of activities in which students have to retrieve math knowledge from memory (e.g., more low-stakes quizzes or filling out a visual organizer without consulting course materials) so that students have to practice remembering the course content."

Next, analyses of data from two measures suggested that the FUSE treatment influenced teachers' actual behaviors—that is, their language and communication to convey a culture of learning—as expected by the theory of change (Figure 2). On the first measure, administered in the Fall several months post-randomization, treated teachers provided a more negative rating of a fixed mindset "speech" a teacher might give in class (see Figure 4), *s.m.d.* = -0.14 *s.d.*, pr(ATE<0)=.817. On the second measure, teachers gave free-response descriptions of the language they would use in response to three common student mistakes on Algebra or pre-Algebra concepts (for examples of language, see Extended Data Table 4). Responses were coded reliably (by coders blind to condition) into several categories that were combined into a single metric (see Methods). Treated teachers were more likely to use culture-of-learning-supportive language in response to student mistakes compared to control teachers, who were more likely to use language that created a culture of judgment and evaluation, *s.m.d.* = 0.51 *s.d.*, pr(ATE>0)>.999. For example, treated teachers were far more likely to ask students open-ended questions to try to understand the origins of their mistakes. Control teachers were more likely to use language intended to embarrass or rush a student who made a mistake. Third and finally, eight months after the start of the intervention, treated teachers were more likely to retrospectively report using culture-of-learning practices, *s.m.d.* = 0.20 *s.d.*, pr(ATE>0)=.912.

Overall, these results suggested that the FUSE program changed teachers' behaviors, thus resulting in an informative test of our primary hypotheses about the impact of culture of learning practices.

Impacts on an Engaging Classroom Culture

Did changes in teachers' practices improve students' actual experiences in the classroom? The first pre-registered outcome was students' overall, *gestalt* perceptions of the classroom culture of *respect* for students. Much theory has suggested and prior data have shown that adolescents are, by virtue of their stage of puberal maturation, especially attuned to the respect afforded them by adults, such that they tend to engage more deeply when they feel respected and they tend to disengage when they feel disrespected.[7,17,38] If treated teachers had truly created a culture of learning, we theorized, then students would feel more respected, and therefore more engaged, because in a culture of learning teachers value students even when they make mistakes. By contrast, in a culture of judgment and evaluation, which should be more prevalent in the absence of the FUSE treatment, teachers were theorized to treat student mistakes or confusion more harshly (e.g. through shame or blame), which should lead students to feel more disrespected. Thus our pre-registered test of whether teachers' culture-of-learning practices were likely to engage students concerned perceptions of respect.

As shown in Figure 4, the treatment improved the respectful classroom culture by *s.m.d.* = 0.16 *s.d.,* pr(ATE>0) = .875. An exploratory analysis found, surprisingly, that the impact on respect was the most striking early in the year, at the first measurement occasion, *s.m.d.* = 0.26 *s.d.*, pr(CATE)>0=.95, and weaker by the final measurement occasion, *s.m.d.* = 0.07 *s.d.*, pr(CATE>0)<.75, pr($\text{Diff}_{\text{CATES}}$>0)=.993. This may be because the teacher practices were a stark contrast to what students expected at the beginning of the year but became expected by the end.

A secondary pre-registered analysis found persistent, year-long treatment impacts on the overall culture of learning. This culture was measured by a composite of nine items that students answered twice, in the Fall (October) and in late Winter (February). Items included: "*My math teacher cares more about whether I learn and improve in class than whether I get the highest grade on my first try*" or "*In my math class, a lot of students hope that they will not be called on, because they are afraid they might say something wrong*" (reverse-scored). The effect across both times points was *s.m.d.* = .16 *s.d.*, pr(ATE>0)=.938. (see Figure 5 and Extended Data Table 3). Unlike the respect results, impacts were equal across the two time points. Similar findings were especially striking for the subset of three of these items that focused specifically on teachers' treatment of *mistakes* (see Figure 5 and Extended Data Table 3C, row 3). Thus, students in treated classrooms observed more of a culture of learning throughout the year—especially regarding the chance to learn from mistakes.

To explore whether students' experiences of this more respectful culture translated into greater behavioral engagement, on the Fall student survey we administered the "make-a-math-worksheet" behavioral task (see validation in [4]). Students were asked to choose either challenging math problems that could teach them something new or easy math problems that they could do well on but that would not teach them anything new. Following the analysis plan for this measure (from previous research[51]), we found that treated classrooms showed more engagement relative to control group classrooms, between-teacher *s.m.d.* = 0.45 *s.d.* pr(ATE<0)=.997 (see Figure 5 and ED Table 3).

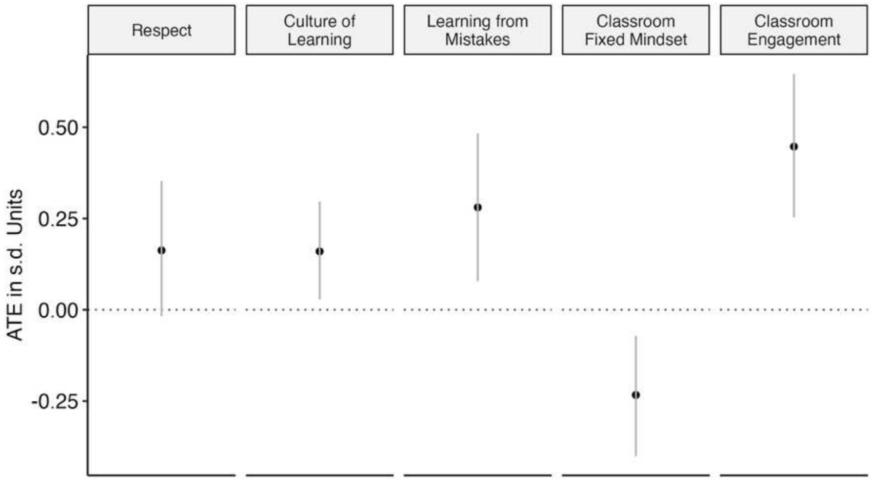


Figure 5. Classroom culture average treatment effects (ATEs) for the FUSE culture-of-learning program compared to the control program. Note: Figure depicts the average treatment effects (ATEs) in pooled between-teacher *s.d.* units and the intervals from the 10th to the 90th percentiles. Results from multilevel BCF models, which were fit by nesting students within teachers and estimating a random intercept for each teacher; during posterior summarization, results were aggregated to the teacher level to yield teacher-level treatment impacts, which were then averaged to produce the ATE estimates shown here. *Respect* and *Classroom Fixed Mindset* were rated at all three measurement occasions, leading to higher overall numbers of teachers. *Culture of Learning* and *Learning from Mistakes* were rated on the second and third student surveys, and *Classroom Engagement* was rated on the second survey only, leading to lower sample sizes for those outcomes. See Extended Data Table 3 for point estimates.

Impacts on Math Test Performance

Main effects. Did this more respectful and engaging classroom culture of learning translate into an impact on math test performance? In the spring semester, approximately six months into the school year, students completed a validated math assessment developed by leaders in assessing math performance internationally (see methods and Extended Data Fig. 7). The scores in treated teachers' classrooms were .122 *s.d.* higher than the scores in control teachers' classrooms on the same set of test items, pr(ATE>0)=.989 (see Extended Data Table 4 for teacher-level *s.m.d.'s*). In terms of empirical benchmarks for interpreting effect sizes for students of this age[58], this treatment effect size corresponded to an additional four months of learning (out of nine in a year) compared to controls.

Heterogeneity of impacts.

Heterogeneity of impact across student racial and ethnic groups. It was important to examine heterogeneity of effects in addition to average impacts [59] because such analyses could reveal the kinds of students who tend to show greater gains. As a preliminary matter, we found strong group disparities in math achievement in the control condition, as shown in Figure 6 and Extended Data Table 4. For instance, compared to White, non-Hispanic students in the control group, Black students in the control group had lower average math test scores (see Figure 6).

The intervention program's impact among Black students was slightly larger than this achievement gap, and therefore Black students' mean performance in the treatment condition was the same as the mean among White, non-Hispanic students in the control group, pr($\text{Diff}_{\text{Black v. White}}$ >0)<.75 (see Figure 6 and Extended Data Table 4). Turning to disparities *within* the treatment group, the treatment impact among Black students (.20 *s.d.*) was slightly larger than the impact among White, non-Hispanic students (.14 *s.d.*; both pr(CATE>0) > .980). Thus we saw a 35% reduction in the Black-White achievement gap in the treatment group (see Figure 6; effects for other groups appear in Extended Data Table 4). (For parallel results showing a larger reduction of racial group disparities in students' reports of classroom respect*,* see Extended Data Figure 2).

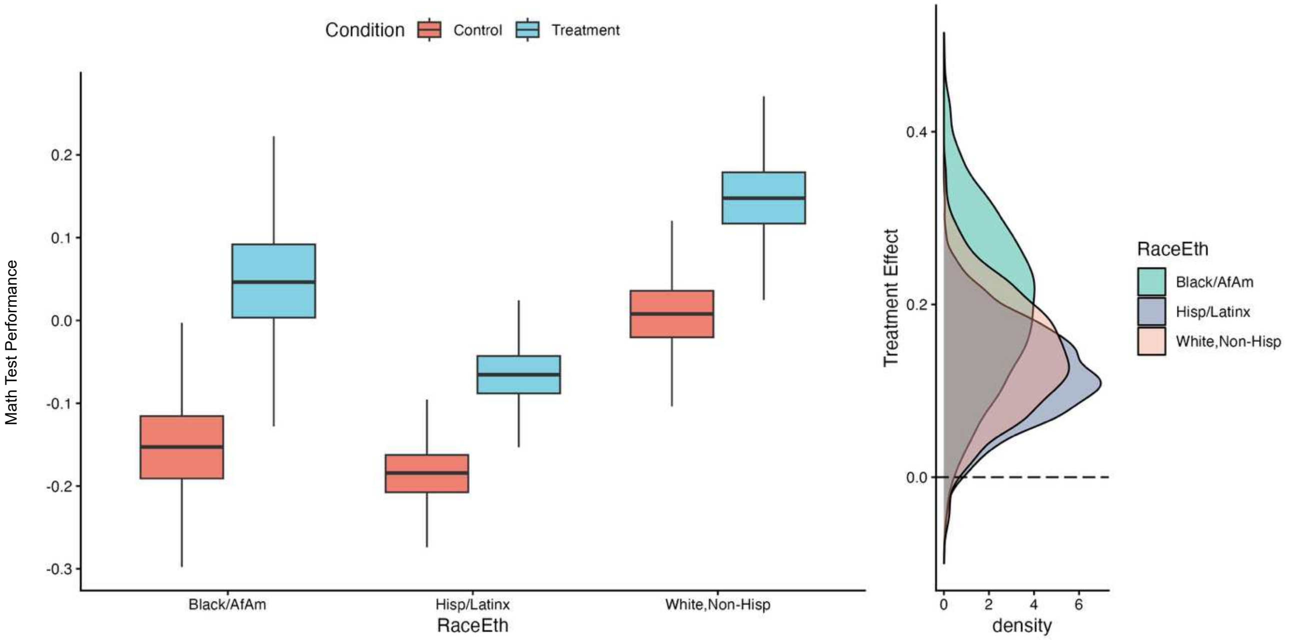


Figure 6. Heterogeneity in the impacts of the FUSE culture-of-learning program on classroom math test performance across student racial and ethnic groups. Note: Results are from multilevel BCF models, which were fit by nesting students within teachers and estimating a random intercept for each teacher. During summarization and analysis of race/ethnicity CATES, model fits were aggregated to "local identity groups" (i.e. race/ethnicity groups within teachers) and then averaged to yield teacher-level treatment impacts. See Extended Data Table 4 for point estimates. Posterior probabilities for interaction effects: pr($\text{Diff}_{\text{CATEs Black vs. White}}$>0) = .755; pr($\text{Diff}_{\text{CATEs Hispanic vs. White}}$>0) <.75; pr($\text{Diff}_{\text{CATEs Asian vs. White}}$>0) <.75; pr($\text{Diff}_{\text{CATEs Other vs. White}}$<0) = .855. The panel on the left is in the student-level-standardized metric (mean of 0 and *s.d.* of 1); effects presented in student-level *s.d.* units. See Extended Data Table 4 for achievement effects in teacher-level *s.d.* units.

Heterogeneity of impact across teachers. There was substantial variation in treatment impacts across teachers (see Figure 7A), suggesting that the program was more effective for some teachers than for others. Thus, we examined potential moderators of the treatment impact.

Moderation by teacher fixed mindset. The primary, pre-registered teacher-level moderator of impacts was teachers' pre-random-assignment mindsets. Previous research found that teachers reporting more of a *fixed mindset* had classrooms that were rated by students as having more of a culture of judgment and evaluation, and less of a culture of learning [12,14,60]. Indeed, in the present research, control condition teachers who reported more of a fixed mindset prior to random assignment were rated by students as being less respectful compared to teachers reporting more of a growth mindset, r = -.25 (see Extended Data Figure 2).

Next, moderation analyses found that teacher fixed mindset was associated with greater treatment impacts on both classroom perceptions of respect at r = .31 (see Extended Data Figure 2) and classroom math achievement at r = .25. (see Figure 7B and Extended Data Table 4). Breaking down the achievement moderation, there were twice-as-large improvements in math test performance among prior-fixed-mindset teachers, $\text{CATE}_{\text{Fixed mindset}} = 0.203$ *s.d.* (6.8 additional months of learning), compared to prior-growth-mindset teachers, $\text{CATE}_{\text{Growth mindset}} = 0.104$ *s.d.* (3.5 additional months of learning), $\text{pr}(\text{Diff}_{\text{CATEs}}>0)=.787$.

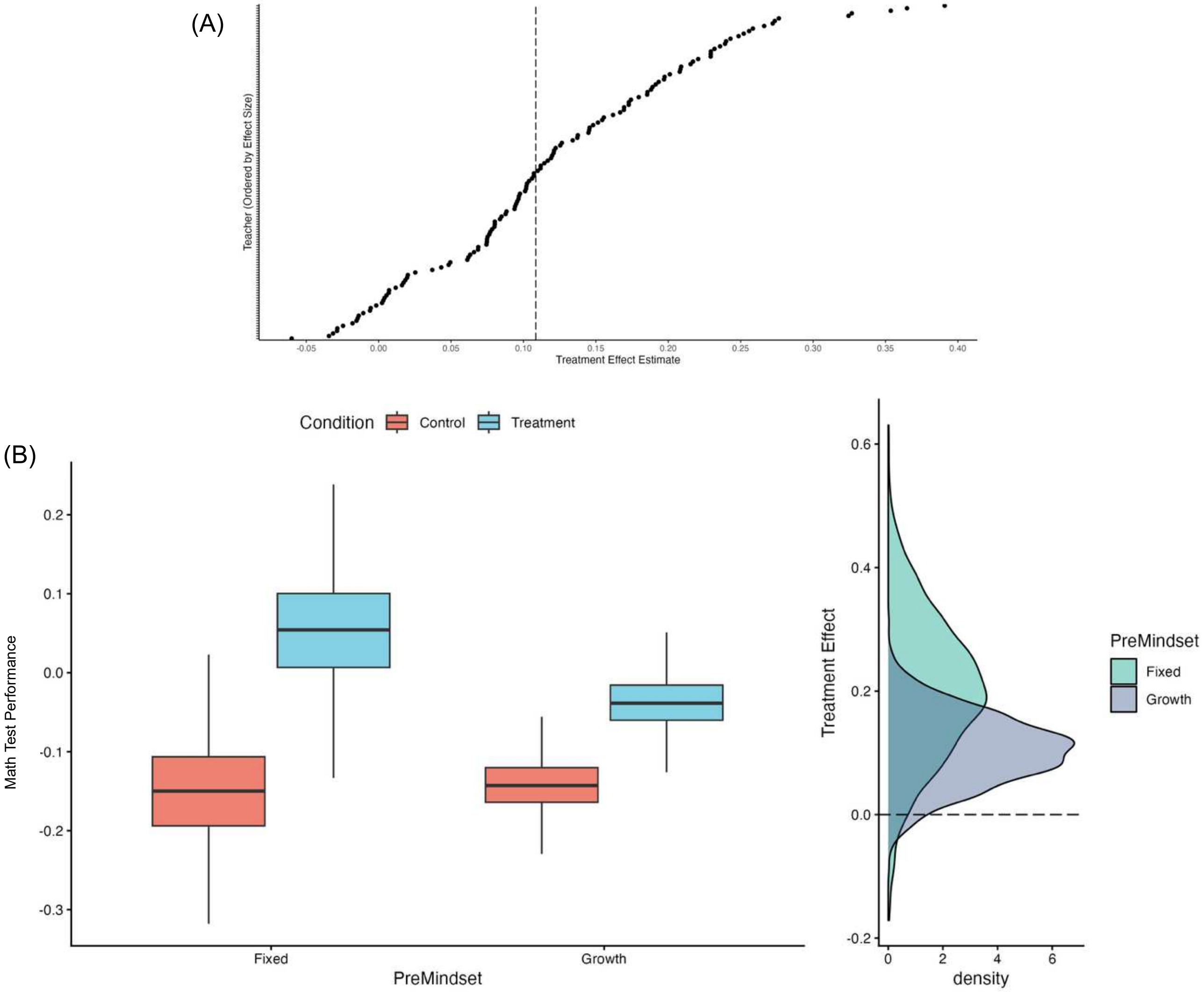


Figure 7. Between-teacher heterogeneity in the impacts of the FUSE culture-of-learning program on classroom math test performance (A) and moderation of effects by (B) teacher mindset. In (A), each dot is a teacher's estimated treatment impact and the dashed line is the median of the distribution of treatment impacts (i.e. the median treatment effect). Effects in (B) in the student-level-standardized metric (mean of 0 and *s.d.* of 1). Posterior probabilities for interaction effects for (B), across teacher mindset subgroups, $\text{pr}(\text{Diff}_{\text{CATES}}) = .80$. See Extended Data Table 4 for achievement effects scaled by teacher-level *s.d.* units.

This was interesting because, *a priori*, fixed mindset teachers might have been the most difficult to influence with a culture-of-learning treatment. Indeed, the nationally-representative survey to prepare for the current research (described previously; see Extended Data Table 1) found that fixed mindset teachers were the most skeptical of culture-of-learning/growth mindset practices. Nevertheless, we found larger impacts for the primary pre-registered classroom outcomes among previously-fixed-mindset teachers. This suggests that the rhetorical approach used here (called *values-alignment*[31,61]*,* as described in Extended Data Table 1 and the methods) was effective at persuading reluctant audiences, as intended.

Moderation by teacher fidelity of implementation. Next, an exploratory analysis examined whether greater exposure to culture-of-learning content was associated with greater treatment impact (compared to teachers with matching exposure to cognitive-science-of-learning). Treated teachers who attended more of the online coaching sessions from peer teachers (i.e. virtual conferences; see Figure 3) tended to show larger treatment impacts compared to teachers who attended fewer of the virtual conferences, $r = .56$. Treated teachers who never attended the virtual conferences (17% of teachers) showed no differences in their classroom test scores from controls who never attended virtual conferences, pr(CATE>0)<.75, but teachers who attended 3 or 4 virtual conferences (40% of teachers) showed test scores that were 0.26 *s.d.* higher compared to peer teachers in the control condition who attended the same number of virtual conferences (8 additional months of learning for their students), pr(CATE>0)=.995. This result supports the validity of the theory of change (Figures 2 and 3), while also pointing to the need for future research on increasing adherence with virtual coaching.

Moderation by School Context

A final exploratory heterogeneity analysis examined moderation of the test score impacts by school achievement level (which was a moderator in previous mindset studies [51]). This tested the possibility that schools that provide more resources to support teacher success and student learning might also show larger impacts for the FUSE program. Supporting our interpretation of school achievement level as an overall measure of school efficacy, cross-tabulations found that in the bottom tercile of school achievement ratings, 23% of teachers never attended a synchronous virtual conference, which were shown (above) to strongly predict program efficacy. In the top tercile of schools, only 10% of teachers never attended a virtual conference session, and fully 51% attended 3 or more. Thus, a school's rating may be interpreted as an overall proxy for the school's ability to support teachers' bandwidth or appetite for professional development.

Heterogeneity analyses found a strong positive association between the school's official effectiveness rating (i.e., state rating from 0 to 100) and the magnitude of the treatment impact on student test score performance, $r = .46$. Schools that were in the top two thirds of ratings (i.e. top and middle terciles) showed the expected positive impact (5.5 months of learning), pr(CATE>0)=.993. Schools in the bottom tercile showed no impact, pr(CATE>0)<.75, pr($\text{Diff}_{\text{CATEs}}$>0)=.984. This suggests that the teacher-behavior-change program does not operate in a vacuum; it may depend in part on the broader institution's ability to support teacher's behavior-change at a structural level.

Impacts on Teacher Well-Being

Our final pre-registered hypothesis was that if FUSE could address fixed-mindset-teachers' student disengagement, it could also improve their well-being (recall Figure 1). Burnout was measured with items such as "*The stress and disappointments involved in teaching aren't really worth it*", on the Fall and Spring surveys; Satisfaction with Life was measured with items such as "*If I could live my life over, I would change almost nothing*"), only on the Spring survey.

As a preliminary matter, teachers' pre-randomization fixed mindset was strongly correlated with burnout in the control condition, as hypothesized, $r_{\text{Fall}} = .55$ $r_{\text{Spring}} = .56$ (see Extended Data Figure 4). Translating this into percentages, in the control condition 34% of baseline-fixed-mindset teachers reported elevated burnout symptoms in the Spring, compared to 11% with baseline-growth-mindset, pr(Diff>0)=97.

Critically, teachers with prior fixed mindsets in the treatment condition tended to show the greatest reductions in burnout. There were strong correlations between prior fixed mindset and the magnitude of the reduction in burnout caused by the treatment, r_{Fall} = -.66, r_{Spring} = -.48 (see Figure 8 and Extended Data Figure 4). Interpreting this result, in the treatment condition just 16% of baseline-fixed-mindset teachers reported feeling burnt out in the Spring, a reduction in burnout of more than half relative to 34% in the control condition, pr(CATE<0)=.88. Among baseline-growth-mindset teachers, just 11% of teachers in the treatment and control conditions reported feeling burnt out (also see Figure 8). The results for the Fall survey were even more striking than the Spring (see Figure 8).

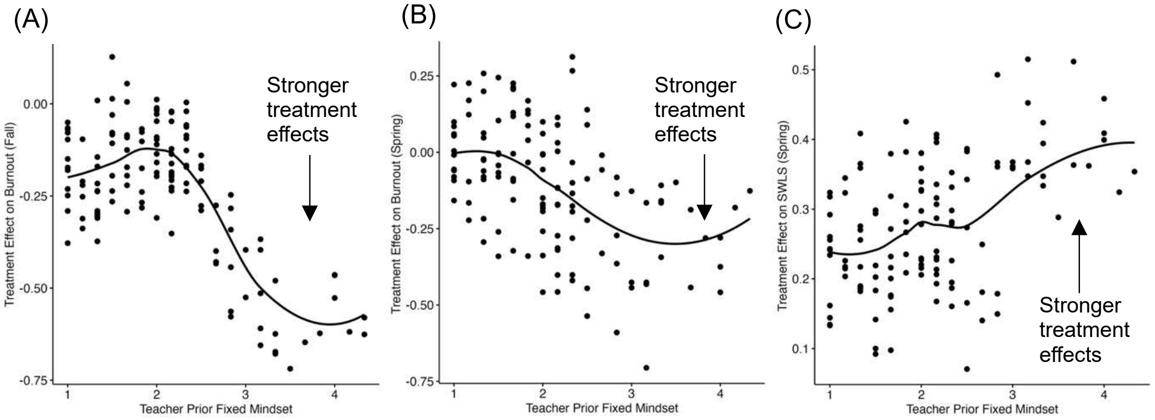


Figure 8. Moderation of the impacts of the FUSE culture-of-learning program as a function of teacher fixed mindset, for (A) reduced teacher burnout in the Fall (r_{Fall} = -.66), (B) reduced teacher burnout in the Spring (r_{Spring} = -.48), and (C) greater satisfaction with life in the Spring (r_{Spring} = .47). Note: Dots represent the mean of the posterior distribution for each teacher's estimated treatment impact in the BCF model.

Analyses also revealed similar, long-lasting impacts in the Spring on the Satisfaction With Life Scale (SWLS) [62], which is a globally-popular and well-validated measure of well-being. In the control condition 34% of baseline-fixed-mindset teachers reported being satisfied with their lives overall, compared to 44% among baseline-growth-mindset teachers, pr(Diff<0)=.77. There was a strong positive correlation between prior fixed mindsets and greater improvements in satisfaction with life; r_{Spring} = .47 (see Figure 8 and Extended Data Figure 4). Translating this into percentages, the treatment increased the percentage of baseline-fixed-mindset teachers who were satisfied with their lives to 60%, pr(CATE>0)=.92, which is a 76% improvement in well-being. Among baseline-growth-mindset teachers, again 60% reported being satisfied with their lives, which was also a meaningful improvement pr(CATE>0)=.88. Overall, the culture-of-learning program tended to improve teacher well-being overall and most strikingly among teachers with more of a prior fixed mindset.

Interestingly, we also found main effects of the treatment on well-being measures: $\text{Burnout}_{\text{Fall}}$ *s.m.d.* = -0.25 *s.d.,* pr(ATE<0) =.956; $\text{Burnout}_{\text{Spring}}$ *s.m.d.* = -0.11 *s.d.,* pr(ATE<0) = .78; $\text{SWLS}_{\text{Spring}}$ *s.m.d.* = 0.27 *s.d.,* pr(ATE>0)=.959. (See Figure 9). Translating this into percentages, in the Fall there was a reduction in burnout rates from 28% of teachers in the control condition reporting elevated levels of burnout symptoms to 14% in the treatment group, corresponding to a 50% overall reduction.

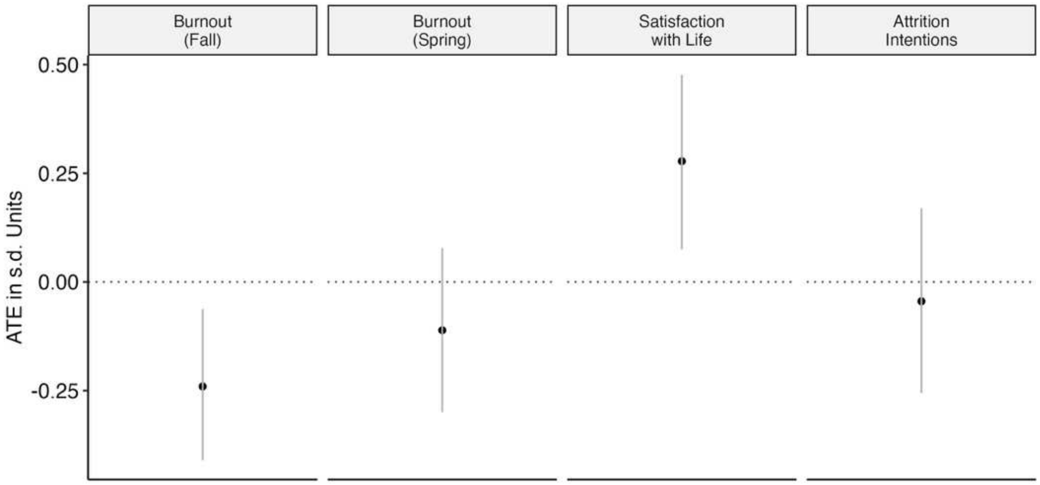


Figure 9. Teacher well-being average treatment effects (ATEs) for the FUSE culture-of-learning program compared to the control program. Note: Figure depicts the average treatment effects (ATEs) in pooled between-teacher *s.d.* units and the intervals from the 10th to the 90th percentiles. Results are from BCF models, which include inverse probability weights (IPWs) to adjust for missing data/attrition. Sample items: *Burnout*: "The stress and disappointments involved in teaching aren't really worth it;" *Satisfaction with life:* "If I could live my life over, I would change almost nothing." See Extended Data Table 3 for point estimates.

Attrition Intentions

Counter to our pre-registered hypotheses, the FUSE program did not impact teachers' self-reported intentions to leave their current teaching job, when those intentions were measured in March, the end of the observation period, *s.m.d.* = -0.04, pr(ATE<0)<.75 (See Figure 9). Speculating, perhaps control group teachers had not yet been discouraged enough to leave their position, in part because the state test was administered in April (and so they did not yet know if the year had gone poorly). Thus, although the treatment led to a greater sense of well-being, it might not have prevented treated teachers from intending to quit at our time of measurement because it did not yet present them with clear evidence of their greater effectiveness.

DISCUSSION

This experiment found that a relatively light-touch intervention for teachers—one that was delivered at scale, in 80 schools using mostly virtual training sessions —could have a meaningful impact on teachers' mindsets and behaviors, on students' math test performance, and on teacher well-being. These findings support several key conclusions.

First, we have now discovered a treatment for student disengagement that shows real promise for impact not only on the student experience but also student challenge-seeking behavior and performance. This is valuable because, as we noted at the outset, there is currently a major crisis of student engagement[1], which poses considerable risk for the future of education and perhaps society.

Furthermore, although FUSE was helpful for all groups, the primary, pre-registered heterogeneity findings found that FUSE showed the largest impacts among teachers (e.g. those with a fixed mindset) and students (e.g. minoritized students with lower performance) who might have been most in need of such a treatment. Without the FUSE program, teachers with a fixed mindset tended to have less-engaged students. And Black students tended to report being less respected, especially by fixed mindset teachers, compared to White students. Those students also, unsurprisingly, reported less engagement and showed poorer performance. FUSE's impacts on both respect and test performance were especially large for teachers with a prior fixed mindset, reducing the negative impacts of teachers' prior culture of judgment and evaluation, and narrowing group disparities by 35% (for achievement) or completely (for perceptions of respect).

Potentially qualifying this finding, we found that the school's overall level of functioning mattered. Schools in the bottom tercile of the state's rating system, which ranks schools for their ability to improve and perform well for all groups of students, did not show meaningful test performance impacts, perhaps because such schools lacked resources to support teachers' implementation of the program. Thus, these heterogeneity analyses were consistent with the Mindset × Context framework[15,40], which predicts that a social-psychological intervention would be most effective for the most psychologically-vulnerable individuals (e.g,. those with a fixed mindset) in the most structurally-supported contexts (e.g. those with sufficient structural/educational resources).

Second, we have identified an effective means for generating teacher behavior change. This was important in part due to the large number of past experiments that have found little to no impact of even heavy-handed and time-intensive teacher professional learning programs[63]. Indeed, the problem of professional behavior change, in general, is a large one, with many touted and popular programs failing to influence workplace behavior in rigorous experiments[64]. We suspect that the FUSE approach was effective where other programs were not in part because it used the *values-alignment*[31,61] method of persuasion. Values-alignment holds that it is generally more effective to reframe a behavior as being consistent with a group's core values, rather than trying to change what they value. Here we used this approach to address something teachers already cared about deeply: obtaining the willing engagement and learning of students in their classrooms, without coercion. By framing FUSE as a means to attain what, to teachers, was a core criterion for professional status and respect[48], then the program was able to instill greater and more lasting changes in teachers' beliefs and attitudes and behaviors.

The impacts on teacher well-being were impressive in part because the SWLS has been extensively validated around the world, most recently in a global study of nearly 57,000 people from 65 countries [65], and scores on the SWLS have been correlated with physical health, employment, and mortality[66,67]. However, few pre-registered, longitudinal studies have found intervention impacts on the SWLS over time[68]. Indeed, one large and comprehensive research synthesis concluded that "a scientific foundation is lacking" for the most common interventions to promote people's happiness or satisfaction with life [68,69]. Often, promising early studies were later found to have yielded inflated effect sizes, due to a combination of small samples and flexibility in statistical analysis. When larger, pre-registered, replications were conducted, they yielded much smaller or null effects on well-being [68,69]. And yet in the present study we found enduring impacts on SWLS of approximately a quarter of a standard deviation. Notably, the FUSE intervention did not explicitly target global well-being. Instead, it relied on a presumed link between workplace satisfaction and overall satisfaction with life. Thus, when FUSE helped teachers to create the kinds of engaging, equitable classrooms that many teachers aspired to, then it helped them to be more satisfied with life in general.

Although the present study involved over 12,000 students and a statewide, diverse sample of 37 school districts, a limitation of the study is that it lacked the sample size and representativeness to conduct truly conclusive and comprehensive heterogeneity analyses. Preliminary analyses suggested that the findings were strongest among teachers who were the most vulnerable (e.g. those with a fixed mindset) in the most supportive contexts (e.g. higher-achieving schools), but such findings require further interrogation and replication in larger and more representative samples.

It is also important to clarify that the present study's control group—focusing on cognitive science principles of learning and memory—does not in any way challenge the validity of that subfield's basic scientific conclusions. Ours was not a laboratory study testing whether cognitive science principles can affect memory. Instead, we tested *whether instructing teachers about these cognitive science findings* (and providing a network in which to discuss them) could influence their instructional practices in a way that improved student learning as much as the culture of learning practices. Notably, we did find that the control version of the FUSE program was effective at communicating the science, and that teachers and districts found real value in it. That knowledge, however, did not translate into better performance, presumably because it left the culture of judgment and evaluation intact. Future research may well find that training teachers on the science of learning could impact their students' outcomes—for example by combining the two versions of the fellowship together, so they build on one another—but we did not test this in the present study.

Finally, the present study represents a success story for cumulative, reproducible, and transparent research. The FUSE program had its origins in pre-registered RCTs showing that *student* growth mindset interventions were more effective among teachers with more of a growth mindset, and less effective among teachers with more of a fixed mindset[53]. Next, pre-registered studies manipulated teachers' fixed vs. growth mindset messages to students, using randomized experiments, confirming the causal results of the measured moderator in previous studies [16]. That program of research yielded a new theoretical model, the Mindset × Context framework[15,40], which in turn informed a long process of qualitative and observational research to identify the FUSE teacher practices that would best support students' mindsets[17]. Ultimately, this line of work yielded the present *culture of learning* principles and practices at the heart of the FUSE program. Finally, here we showed that motivating a new group of teachers to use those culture-of-learning practices and supporting them in that implementation with a virtual network and fellowship could improve student engagement and performance, most strikingly among the fixed mindset teachers who were the targets of the initial research.

The present research therefore serves as an empirical example of the value of systematically studying heterogeneity in field experimental research—first correlationally, later using laboratory methods, and finally using a scaled-up field experiment that manipulates a previous moderator[16]. Heterogeneity is not, as some have claimed, solely a secondary analysis that researchers conduct to salvage an otherwise-ineffective study. Instead, we argue that the analysis of heterogeneity should be at the core of how behavioral scientists routinely conduct their programs of research, so that more reliable and impactful solutions can be discovered[59]. The fact that in the present research this process yielded an intervention that addressed problems that so far have eluded behavioral scientists—while also suggesting novel methods for changing workplace culture—supports our claim[59] that a rigorous approach to studying heterogeneous effects of behavioral-science-based interventions, aided by modern statistical methods that reduce the chance of false-positive results, can play a useful role in the field's search for large and lasting impacts on policy-relevant outcomes.

METHODS

Ethics Approval and Consent

This study was approved by the IRB at the University of Texas at Austin (STUDY0003823). All teachers provided active consent to participate and all students provided active assent. Each school in the study signed a contract authorizing the FUSE program delivery and data sharing.

Pre-registration

The comprehensive pre-registration, covering the primary and secondary outcomes was posted and frozen prior to the accessing or analysis of any data on the secure server (https://osf.io/6nwuq). A preliminary pre-registration for the immediate post-test (summer PLI) outcomes was posted and frozen prior to the collection of any outcome data (https://osf.io/vwsry), and this was later updated by the comprehensive pre-registration. In the body of the text we designate primary, secondary, and exploratory analyses. For a complete reporting of the comprehensive pre-registration and the findings for each research question, see the Supplemental Online Materials.

Participants

Participants were a maximum of *N* = 152 teachers and their *N*=12,432 6th to 9th grade math students in 80 schools from 37 public school districts in Texas. In terms of demographics. 72.8% were women (the rest men); 47.2% of teachers were Hispanic/Latinx (the rest were non-Hispanic); 86% were White, 4% were Black, 2% were Asian and the rest were from other racial groups. The schools in which teachers were employed were diverse and resembled state-wide benchmarks: 21.6% were from urban areas; 13.5% from suburban areas; 27% from towns; and 37.8% from rural areas, according to U.S. Census classifications. More detailed descriptions of the characteristics of the teachers and students appear in the supplemental online materials (Supplemental Table 1).

Procedure

Sampling. For this first evaluation of FUSE, the scientific objective was to have a large and diverse-enough sample to estimate a reliable average treatment impact and begin to understand moderation of impacts as well as generalizability. Therefore, a diverse sample of school districts across Texas was drawn and then recruited, using data from the Education Research Center (ERC), Texas's statewide longitudinal data system. All high schools in Texas (schools that serve at least 9th-12th grade) were initially placed into one of five strata based on geography (South Texas, the Gulf Coast, Central Texas, the DFW metroplex, and mostly rural North and West Texas) defined by Educational Service Center (ESC) areas. We created one additional stratum of schools across geographic regions with larger populations of Black students. Because Texas public schools are majority-Hispanic, but the Black population is much smaller (~12%), we created this high-proportion-Black stratum to ensure our sample contained a sufficient number of schools in which we could examine whether FUSE was more or less effective for Black students in different kinds of schools. The random sample contained 208 districts and 251 high schools sampled proportionally from each of six strata (the high-proportion-Black stratum and the five based on geography). Additionally, we recruited the largest feeder middle school for each high school in the sample with at least one feeder middle (some "high schools" in our sample are K-12 or 6-12 and therefore do not have a feeder middle). Within each recruited middle and high school, we invited all teachers of 6th-9th grade regular instruction (e.g. not special education) math classes, up to Algebra I but excluding Geometry and higher (because Geometry is not a tested subject in Texas). Although recruitment started with a random sample, we employed "snowball" sampling methods to meet sample size requirements within the year of sample recruitment. Once the randomly selected schools were contacted and interested, then they could recommend neighboring or similar districts to the recruitment team. Thus, the final sample is, by design, not a strictly representative sample, but rather a diverse sample that closely matches the measured characteristics of schools and students in the state of Texas. Notably, the present study is a substantial improvement over the status quo in educational evaluation research in the U.S., which typically includes either one of the top five largest districts (e.g. NYC; Houston; Chicago) or the school district(s) closest to the university(ies) leading the study [70].

The Supplemental Online Material (Supplemental Table 1) shows a comparison of characteristics of study schools relative to the sampling frame, consistent with an interpretation of comparatively high representativeness. There were small differences that were generally in the direction of the achieved sample including schools that were lower-performing and had a higher proportion of racial/ethnic minority students, thus yielding variation in potential moderating characteristics.

Recruitment. Once the sample was drawn, school recruitment was led by the Texas Behavioral Science and Policy Institute (TxBSPI) in partnership with ICF International, a global research firm specializing in the recruitment of scientific samples. TXBSPI collaborated with ICF on the development of all procedures and recruitment materials. Then contacts and recruitment were led by TXBSPI's team of former Texas superintendents, who had expertise in forming research partnerships with school districts and obtaining signed contracts. The recruitment process involved (a) sending printed invitation packets to all districts; (b) follow-up with emails and phone calls for all superintendents; (c) invitations for all superintendents to attend symposia hosted in Austin and in the different regions of Texas; (d) presentations at annual conferences for administrators and math coordinators; and (e) in-person visits to any district that did not respond to any of the above contact methods. All 208 districts were contacted no fewer than 15 times, until they have signed the contract, and were considered non-respondents after all conversion attempts are exhausted. Once leaders were contacted, then the FUSE facilitators (former secondary STEM teachers in Texas) presented to the relevant math departments to recruit teachers to consent to participate in the study. The two most common reasons for teachers in partner districts declining included change in roles (i.e. no longer planning to be a math teacher at the school) or a conflict with summer schedule (e.g. teaching summer school or family vacation).

Randomization. Once a list of consenting teachers was recruited then the list of teachers' names was sent to Dr. Elizabeth Tipton's research team at Northwestern University's STEPP (Statistics for Evidence-Based Policy and Practice) Center. Random assignment was done using a true random number generator (Random.org), blind to teachers' or schools' information or outcomes. In schools with 3 or more teachers, these teachers were randomized within schools to conditions. In schools with only 1 teacher, then teachers were randomized individually. In schools with only 2 teachers agreeing to join the program, both teachers were randomized to the same condition, because three teachers were needed to eventually estimate school-level impacts. (When analyses were re-conducting treating as a single unit all teachers/students who were randomized together, then the results were statistically identical). Several teachers were invited to the FUSE program by their school districts without first notifying the research team; anticipating this, the research team had a list of randomly sorted condition numbers that was used to assign teachers on-the-spot on the first day of the FUSE training. A comparison of the student and teacher characteristics between conditions appears in the Supplemental Online Materials (Supplemental Table 1); all variables in that comparison were included as potential covariates in the Bayesian, machine-learning model, to reduce any potential imbalances on observables that occurred due to chance.

Program delivery. FUSE teachers were invited to an in-person, three-day event in Austin, TX, called the summer Professional Learning Institute (PLI). Both treatment and control teachers attended simultaneously. Plenary sessions delivered content that was relevant to both conditions (e.g. about the logistics of the fellowship), including talks from the principal investigators that talked broadly about student engagement but did not mention condition-specific content. Randomized content was delivered to teachers in the two experimental groups by having teachers attend sessions in different meeting rooms, and by assigning teachers to different content on the online course management platform (Canvas). After teachers completed their condition-specific online modules (see Figure 3), teachers participated in synchronous, facilitated sessions with instructional coaches, on the topic of how to successfully implement their respective practices in the coming year. All PLI sessions were video recorded and teachers who could not attend the live sessions accessed the recorded sessions virtually. As shown in Figure 2, all subsequent content after the summer PLI was delivered virtually, including monthly virtual conferences and a one-day PLI in the Fall and Spring semesters. When teachers missed a session, they completed the PLIs virtually and asynchronously.

Data collection. ***School administrative data.*** Schools provided classroom rosters for each teachers' 6th to 9th grade math students, to compile master lists and merge data successfully. Publicly available data about each school and district was collected from the Texas ERC, including the school's 0 to 100 quality rating. ***Teacher reports.*** In the week before summer PLI, and during the final hour of each of the three PLI meetings, teachers were directed to a survey to report on their beliefs, attitudes, and behaviors (see Figure 2). ***Student reports.*** All student data was collected via three online modules delivered over the course of the year (see Figure 2), each lasting approximately 30 minutes. Students completed the modules via a private link on their personal or school devices quietly during their regular math classroom meetings. These modules had several purposes: to measure student self-reports (of their own attitudes/beliefs and their perceptions of teachers); to deliver student modules that were designed to educate students about their brains and how learning happens (these differed depending on the condition); and to administer the standardized math assessment.

Data processing and blinding. All data were processed by two professional research database analysts, each with experience creating and maintaining large, publicly-available longitudinal datasets. The database administrators processed all data independently of the principal investigators and did not test any primary research hypotheses. All decisions about data merging and harmonization (e.g. matching records; handling duplicate records) and inclusion or exclusion in the sample were made following standard operating procedures for the Texas Behavioral Science and Policy institute and were made blind to the impact of any decisions on the results of the study. The experimental condition variables were processed in a separate file from the datasets containing outcome variables. Only once all data had been cleaned and processed, and the pre-registrations were submitted and frozen on osf.io, did the principal investigators access the data for analysis.

Treatment (Culture-of-Learning) and Control (Cognitive-Science-of-Learning) FUSE programs.

The various touchpoints for the treatment and control programs are depicted in Figure 2 in the main paper. The treatment and control activities are summarized in Extended Data Table 2. An overview of each condition, and screen captures and quotes from the modules, appear in the Supplemental Online Materials, Supplemental Figures 1 to 12. ***Elements in common across conditions:*** All FUSE teachers (treatment and control) were treated as "fellows" to honor them and increase participation with the program. Synchronous activities (in both conditions) were informed by research on how best to move novices toward greater expertise. For example, novices can more readily develop expertise when they invent rules by comparing contrasting cases (i.e. seeing examples and non-examples side by side)[6], and so FUSE uses "inventing with contrasting cases," which involves non-examples alongside better examples of practices. Both treatment and control synchronous sessions were facilitated by other public school math teachers in the same state because prior research found that teachers learn pedagogical skills in a more enduring way when receiving coaching from experienced, near-peer classroom teachers [45]. These "near-peer coaches," continued to facilitate monthly coaching sessions throughout the year. Those coaching sessions were held in small groups, in order to create a peer network of support throughout the year, because research has found that teachers learn better from a peer community in which to share ideas about their instructional practices [71]. Thus, FUSE connects teachers in a distributed network of peer "fellows" who share novel ideas and tips. During the fall and spring PLI events (each lasting one day) both treatment and control teachers received reports on their student data, and they could discuss ongoing adjustments to their instruction based on the results. Finally, at the end of the year, both treatment and control teachers made a virtual presentation on the changes to instruction they made that year, and their recorded presentations were entered into a "library of practices" for use by future FUSE fellows. ***Elements that differed across conditions:*** During the summer professional learning institute (PLI), which kicked off the fellowship, treatment teachers attended a different keynote session, which delivered the values-aligned message concerning the adolescent motivation for status and respect, relative to control teachers, whose keynote explained how the brain forms and retains memories. Then treatment and control teachers completed different asynchronous modules on the key practices relevant to their respective experimental conditions, followed by synchronous activities to put the modules' recommendations into practice.

Outcome Measures

Teacher self-reports: Manipulation checks.

"Net Promoter" item. On the post-PLI survey, teachers answered this item on an 11-point scale: *On a scale of 0 to 10, how likely would you be to recommend the FUSE fellowship to a friend or colleague?*

Teacher fixed mindsets. Teacher fixed mindset was measured on the pre-PLI survey and the post-PLI survey by taking the unweighted average of agreement with five items: (1) *There's a lot of talk about things like grit or growth mindset, but deep down an experienced teacher knows that some kids have the ability to excel and others don't;* (2) *A student who starts the beginning of the year near the bottom of the class rarely ever has the potential to become a high performer;* (3) *People have a certain amount of intelligence, and they really can't do much to change it;* (4) *Being a "math person" or not is something about you that you really can't change;* (5) *Students have a certain amount of math ability and they really can't do much to change it.* Each mindset item (here and below) was rated on a scale from 1 = *Strongly disagree* to 6 = *Strongly agree*.

Teacher enforcer mindsets. Using the same rating scale as the fixed mindset items, participants rated their agreement with three enforcer mindset items[17], which were turned into a composite by taking the unweighted average of the ratings: (1) *Most adolescents are incapable of behaving correctly unless there are severe consequences for disobedience;* (2) *Most adolescents will take* *advantage of their teachers if they don't fear punishment for immature behavior;* (3) *Most adolescents lack the self-discipline to care more about their schoolwork than their social lives.*

***Teacher mentor mindsets*.** Using the same rating scale as the fixed mindset items, participants rated their agreement with two mentor mindset items[17], which were turned into a composite by taking the unweighted average of the ratings: (1) *Adolescents can display remarkable self-discipline if the classroom environment properly motivates them;* (2) *Adolescents can overcome even high levels of stress and frustration if a teacher properly supports them.*

***Treatment (culture of learning) practices*.** For the "intended" treatment practices composite, teachers rated how likely they would be (on a five-point scale) to use each of four treatment practices on the summer post-PLI survey. The stem was: *How likely are you to use this teaching practice at least once during this 2024-2025 school year?* The items were: (1) *Spend class time encouraging students to reveal and discuss their mistakes with the teacher or classmates;* (2) *From the start of the year, spend time explaining to students how your class is designed to help them be successful when they are struggling;* (3) *Encourage every eligible student to earn points back on an exam or quiz by correcting missed questions or retaking the exam;* (4) *Spend in-class or out-of-class time reviewing or re-teaching concepts students missed after each summative exam.* For the "recalled" treatment practices composite (on the Spring post-PLI survey) teachers reported how often they actually used each of the practices, stem: *How often, if ever, did you use this teaching practice so far during the 2024-25 school year?*

***Control (cognitive science of learning) practices*.** Teachers also rated their likelihood (on a five-point scale) of using the control group practices on the Summer post-PLI survey and their retrospective use of the practices on the Spring post-PLI survey, in response to the same stems. Items: (1) *Assign activities that encourage students to make predictions, which means making an educated guess about the solution to a problem before trying to answer it;* (2) *Design assignments and quizzes so that concepts are mixed up and spaced out over multiple weeks or months, rather than focusing on practicing and mastering one skill at a time;* (3) *Increase the frequency of activities in which students have to retrieve math knowledge from memory (e.g., more low-stakes quizzes or filling out a visual organizer without consulting course materials) so that students have to practice remembering the course content.*

Culture of learning language (open-ended responses to mistakes, post-PLI). To assess possible changes in teachers' natural language in response to mistakes, on the post-PLI survey administered at the end of the first three days of the fellowship, teachers were asked to free-write how they would respond to three common student math mistakes. An example stimulus appears in Extended Data Figure 5, and all stimuli appear in the supplemental online materials. The mistake stimuli that teachers responded to were selected by Dr. William Schmidt's research team at the Center for the Study of Curriculum, using the following criteria: (a) they were common mistakes on previously-released standardized test items; (b) they assessed the same content as the relevant Texas state standards (i.e. TEKS); and (c) the mistakes could plausibly have resulted from different ways of student thinking. These criteria were selected so that the task could elicit teachers' authentic language concerning mistakes that their students would be likely to make during the upcoming year, while also allowing for potential meaningful variation in how teachers responded.

Teachers' open-ended responses were coded by independent coders, blind to experimental condition, into two major categories: *culture of learning* language and *culture of judgment and evaluation* language. (see Extended Data Figure 5 and the supplemental online material). The *culture of learning* language included two verbal behaviors that teachers were trained on in the FUSE fellowship: surfacing student thinking (i.e. asking students to explain their rationale for their answers) and validating students' correct thinking (i.e. noting the steps that students took correctly). The *culture of judgment and evaluation* category included autonomy threatening language (i.e. language that feels controlling and implies students cannot think on their own) and judgmental language (i.e. aggressive or accusatory questions that shame or blame students for their mistakes). Each of the three teacher responses received a score of 1 if the teacher's response included the relevant language and 0 if it did not. Codes were averaged across the three stimuli, for each of the two composites (culture of learning; culture of judgment and evaluation). A final measure used for analysis was calculated by taking the difference between the two composites, yielding a final scale in which -1 corresponded to exclusively culture of judgment and evaluation responses and +1 was exclusively culture of learning responses.

Teacher fixed mindset language (speech ratings, Fall). On the Fall PLI survey, teachers completed a task in which they evaluated various "speeches" that a teacher could give in a math class, in which they explain their philosophy of teaching. Participants evaluated six different mini-speeches, each on a different topic (e.g. the meaning of struggle, the possibility for improvement after a low grade, how the teacher views questions). For each topic, participants were randomly assigned to read and evaluate a "fixed mindset" speech text or a "false growth mindset" speech text. Ratings for all the different speech stimuli were combined into a single composite by taking their unweighted average.

Classroom culture assessments: Primary outcome. At three measurement occasions (SEMs 1, 2, and 2), students rated their teachers with this item: *Students in my math class feel like our teacher treats them with respect* (1 = *Not at all true;* 5 = *Extremely true*).

Classroom culture assessments: Secondary outcomes.

Culture of learning. At two measurement occasions (SEMs 2 and 3) students completed a more comprehensive assessment of the classroom culture of learning. The nine items in this assessment, each rated on a five-point scale, were developed over a five-year period and subjected to extensive testing of reliability and validity, including exploratory and confirmatory factor analysis and concurrent and longitudinal associations with validity criteria. Collectively, the items captured the theorized changes to the classroom culture that the research team expected would be experienced by students if their teachers successfully created a culture of learning. Prior to this study, however, the items had never been administered in full in a randomized trial, and so they were pre-registered as a secondary outcome. The ratings of the nine items were combined into a composite by taking their unweighted average. The items were: (1) *My math teacher believes that some students are smart at math and others are not. (reversed);* (2) *In my math class, a lot of students hope that they will not be called on, because they are afraid they might say something wrong. (reversed);* (3) *My math teacher cares more about whether I learn and improve in class than whether I get the highest grade on my first try.;* (4) *My math teacher cares more about whether students get the right answer to a math problem on the first try, not about whether we improve with time. (reversed);* (5) *My math teacher thinks that mistakes or struggles in math are bad and should be avoided. (reversed);* (6) *My math teacher makes students feel embarrassed or ashamed when they make a lot of mistakes in* *math. (reversed);* (7) *My math teacher gets upset or frustrated when students make a lot of mistakes in math. (reversed);* (8) *My math teacher encourages students to discuss our mistakes with them or each other so that we can fix our mistakes in the future.;* (9) *My math teacher asks students questions to uncover what they were thinking when they made a mistake on a math problem.*

Learning from mistakes. On an exploratory basis, to assess which of the items appeared to show the greatest impact of the treatment program, another classroom culture composite was created exclusively from the items examining teachers' responses to mistakes (items 5, 6, and 7 above).

Classroom math performance assessment: The PROM/SE short form. On two occasions (in early Fall and late-Winter/early-Spring), teachers' classrooms completed was a short form of the established PROM/SE math assessment. The short form was designed to be suitable for administration at scale as a part of regular instruction in 10 minutes, which was the time constraint determined for this study, while also maintaining reliability and construct validity at the classroom level. The PROM/SE measure was created independently from the FUSE research team by Dr. William Schmidt, a global leader in mathematics curriculum and assessment[72], as a part of a large NSF-supported project, "Promoting Rigorous Outcomes in Mathematics/Science Education (PROM/SE)" (award #0314866). The assessment measures five core knowledge domains that are essential for success in pre-Algebra and Algebra I: *ratios and proportions; equations and expressions; linear functions; inequalities; other expressions and inequalities.* (See the Supplemental Online Materials for sample items). These five areas were selected by Schmidt's cross-comparison of the content standards in the Texas state standards (TEKS); that is, the test was normed to the Texas standards, where the current research was conducted. The specific items in the PROM/SE test were selected or adapted by a panel of mathematics curriculum experts and educators from previously-released standardized test item banks as a part of NSF award #0314866. The source item banks included the Trends in International Mathematics and Sciences Study (TIMSS), the Programme for International Student Assessment (PISA), and state-approved achievement tests. Within each domain, twenty candidate items were selected, which varied in their difficulties from 30% correct to 70% correct. Once the expert panel selected candidate items, extensive piloting and cognitive pretesting was conducted as a part of PROM/SE. Participant completed "think-aloud" protocols, and then researchers edited items with language that caused confusion or equivocation, to reduce differential item functioning.

Random selection of items (i.e. "matrixing") was used for the PROM/SE short form. That is, each participant was presented with ten items, two from each knowledge domain, randomly selected without replacement from a pool of 20 items per domain. According to experts in educational measurement, the goal of matrixing "is the reduction in the total amount of time needed for testing while still obtaining group-level estimates of student performance. This can save assessment administration time and scoring time, and reduce the costs of assessment without adding much to the data analysis and reporting tasks."[73] Also, "a matrix sampling assessment design could give enough of the benefits of a full-length assessment without the significant drawbacks that long testing time has on students."[74] Because the goal of the present study was to provide reliable teacher-level estimates for all five content areas, and because the teacher was the level of random assignment and the unit of analysis, then matrixing was used to minimize respondent time (and burden) and therefore maximize response rates. This reduced missing data serves to improve the validity of the study's causal inferences.

To reduce random variance introduced by matrix sampling (i.e. random differences across students in the difficulty levels of the items provided) each item was scored using Rash scaling, which equated items according to their "difficulty" parameters, yielding a single performance score for each student on a common metric. The final measure was calculated on a z-score scale (i.e. *m.* of zero and *s.d.* of 1). Thus the raw analysis scores can be interpreted in terms of student-level standard deviations, making the unstandardized results conform to What Works Clearinghouse standards for student-level performance effect size reporting[57]. To reduce the possibility of bias in the analysis or reporting of the math tests, all scaling of items and calculations of final performance scores were conducted by an independent, contracted analysis team who was blind to all information about students and teachers, including experimental condition assignments.

The validity for the PROM/SE measure is supported, in part, by an analysis of how teachers' scores were related to their students' state standardized test scores. In a pilot sample, 32 teachers' students took both the PROM/SE short form and the Texas STAAR math test (the state's official test), which is scored in terms of four categories: 0 = *failure, or does not approach grade level*, 1 = *approaches grade level expectations*, 2 = *meets grade level expectations*, 3 = *masters grade level expectations*. Teachers' students' average scores on the PROM/SE were correlated at $r = .65$ with the average mastery level of students in their classrooms on the STAAR test. Note that the PROM/SE was designed to measure more conceptual mathematical reasoning, while the STAAR was designed to measure more algorithmic or rote mathematical knowledge, and so the two should be highly (but not perfectly) correlated. A scatterplot depicting the associated between the two measures appears in Extended Data Figure 6. Also, at the teacher level in the current study, we found a meaningful correlation between teachers' classroom-average test scores in the control condition and teachers' self-reported levels of trust, $r = .31$, $t = 3.74$, $p<.001$, replicating prior research[75]. In terms of individual differences validity at the student level, in the current data we also found a meaningful validity correlation between test scores and expectancies for success, $r = .40$, $t = 54.31$, $p<.001$, replicating past studies[13].

Validity evidence also comes from prior evidence that districts with higher-quality curriculum (i.e. more grade-level-appropriate math content) also scored higher on the PROM/SE math assessment (r=.39). A subsequent unpublished validation study led by Dr. Schmidt's research team with >5,000 6th to 8th grade students showed almost identical grade-level progress on the PROM/SE measure with expected grade-level progress (according to Lipsey's benchmarks[76]). Grade 6 to 7: .30*SD* (expected) vs. .31*SD* (PROM/SE); Grade 7 to 8: .32*SD* (expected) vs. .30*SD* (PROM/SE). The current study's data with >20,000 students completing the short form PROM/SE showed a similar trend (e.g. growth of .34*SD* from 7th to 8th grade). Overall, the PROM/SE measure was a valid assessment of 6th to 9th graders' mathematical knowledge.

Teacher self-reports: Primary outcomes.

Burnout. Teacher burnout was assessed on the Fall and Spring post-PLI surveys via a composite of two items that were adapted from the nationally-representative RAND survey of educators[77]: (1) *How confident are you that you could handle the stresses of your job right now? (reversed);* (2) *The stress and disappointments involved in teaching aren't really worth it.* Each item was rated on a five-point scale, and the composite was created by taking their unweighted average. Secondary analyses dichotomized the burnout variable to make the results more interpretable, which yielded an "elevated burnout symptoms" group (> 0.5 *s.d.* above the mean) and a "non-elevated burnout symptoms" group (all others). Note that dichotomization did not influence the model fit (or the posterior probabilities of a difference) because it was done after a model fit that used the full, continuous burnout measure.

Attrition intentions. Teachers' attrition intentions were measured on the Spring post-PLI survey via an item that was also adapted from the nationally-representative RAND survey of educators[77]. Participants rated this statement *What is the likelihood that you will leave your job at your school by the end of the current school year (2024-2025)?* On a four-point scale (1 = *Very unlikely*; 4 = *Very likely*).

Teacher self-reports: Secondary outcomes.

Satisfaction with Life Scale (SWLS). On the Spring post-PLI survey, teachers rated five items from the well-validated SWLS (each on a five-point scale), and their ratings were combined into a composite by taking their unweighted average. The items were: (1) *In most ways my life is close to my ideal;* (2) *The conditions of my life are excellent;* (3) *I am satisfied with my life;* (4) *So far I have gotten the important things I want in life;* (5) *If I could live my life over, I would change almost nothing.*

School context moderator: School rating level. To conduct exploratory analyses by school context, we used Texas' official rating system for public schools. All schools in Texas are rated on a 0 to 100 point scale that is a weighted composite of two values: the greater of either the aggregated student achievement (i.e. overall performance) or school progress (i.e. improvement from one year to the next) scores, and the a score representing the closing of gaps (i.e. reduced disparities in performance between demographic groups). Higher ratings reflect purportedly higher-functioning schools. Moderation analyses were conducted using overall ratings from the year prior to the study (the 2023-2024 school year).

Statistical analysis approach. BCF modeling was the pre-registered analysis approach for two reasons. First, BCF can flexibly incorporate many possible covariates using adaptations of the Bayesian Additive Regression Tree (BART) algorithm. This allows the model to control for any chance imbalances between conditions at each measurement occasion. This was important in the current study because (a) the number of teachers is adequate but modest (*N* = 154), and so chance differences can occur even after randomization, and (b) if there are even small amounts of missing data (e.g. if a given teacher is chronically ill or on leave), it can exacerbate chance imbalances. Therefore, we include many potential covariates that predict the outcome both in the primary statistical model. We also estimate and include a propensity score for missingness. The machine-learning algorithm then makes decisions about how best to incorporate those covariates into the model fits. (The propensity score is also used for inverse probability weighting of results after fitting the model).

Second, BCF is designed to model the moderators of effects separately from the covariates that reduce bias in the treatment contrast. BCF can do this in a relatively conservative way, by incorporating prior expectations of null or weak moderation, and "shrinking to homogeneity." BCF can furthermore estimate a model with multiple, possibly non-linear moderators at once (with a penalty term to avoid over-fitting), to avoid the high false-positive rates that come from re-fitting a linear interaction term model several times. BCF was therefore ideal for both estimating a true average causal effect while also exploring the possible impact of the pre-registered moderators.

For teacher-level models, a single-level regression was estimated. When classroom-level outcomes were analyzed (e.g. student reports or performance scores), a multi-level model with a teacher-level random intercept was estimated, and results were summarized at the teacher level.

Calculation of years of learning. Following the published benchmarks [58], one year of learning for 9th graders in math was .25 *s.d.*; for 8th graders, .22 *s.d.*; for 7th graders, .32 *s.d.*; for 6th graders, .30 *s.d.*

DATA AND CODE AVAILABILITY

Because data involve educational records from minors, all data and syntax for processing, analyzing, and reporting results are available on the secure server at the Texas Behavioral Science and Policy Institute, in deidentified form. Data can be accessed with necessary approvals (e.g. IRB approval and signed documentation of agreement with data security and privacy principles). Contact: txbspidata@prc.utexas.edu.

ACKNOWLEDGEMENTS

The ideas, interventions, and pilot materials for the present study were developed initially in partnership with OnRamps, led by Dr. Jennifer Porter, contributed greatly to the initial phases of the FUSE project. Kumar Garg helped shape the vision of the initiative. The theoretical foundation of the FUSE program was developed via the “Global Mindset Initiative,” which involved working papers and conferences supported by the Yidan Prize Foundation and the Programme for International Student Assessment (PISA). The FUSE program was developed and launched initially by Andrew Kim, who co-founded FUSE and recruited all of the school partners for the current study. Matt Forte conducted the randomization. The team is grateful to the large number of scholars who contributed to the various meetings and convenings that led to the launch of FUSE, and to the educators and facilitators, including: Sergio Estrada; Charles House; Krisstian Langford, and more.

FUNDING

Financial support to develop the intervention’s theoretical model, measurement, and empirical basis came from Schmidt Futures, the National Science Foundation (HRD-1761179; DUE-2201928; DUE-2243530; DRL-2322330; DRL-2509858), the William and Melinda Gates Foundation (now the Gates Foundation), an Advanced Research Fellowship from the Jacobs Foundation (to D. Yeager), the William T Grant Foundation (to D. Yeager), and the Yidan Prize Foundation (to D. Yeager). The Texas Behavioral Science and Policy Institute is supported by the Raymond Dickson Centennial Professorship (to D. Yeager) and by the Population Research Center’s NICHD grant (P2CHD042849). The Alpha Motivation Lab received financial support from the Future of Education, LLC.

COMPETING INTERESTS

None of the authors owns a financial stake in any growth mindset product or service for students or teachers. The FUSE program is now available (after evaluation) for purchase at cost to school districts via the University of Texas at Austin, a non-profit academic institution. None of the authors receives pay, owns equity, or has any arrangement for future profit from the FUSE program, and all funds raised from the FUSE program go to support the research and training mission at UT Austin.

References

1. Anderson, J. & Winthrop, R. *The Disengaged Teen: Helping Kids Learn Better, Feel Better, and Live Better*. (Penguin Random House, 2025).
2. Winthrop, R., Shoukry, Y. & Nitkin, D. *The Disengagement Gap: Why Student Engagement Isn't What Parents Expect*. https://eric.ed.gov/?id=ED663848 (2025).
3. Wong, Z. Y., Liem, G. A. D., Chan, M. & Datu, J. A. D. Student engagement and its association with academic achievement and subjective well-being: A systematic review and meta-analysis. *J. Educ. Psychol.* **116**, 48–75 (2024).
4. Rege, M. *et al.* How can we inspire nations of learners? An investigation of growth mindset and challenge-seeking in two countries. *Am. Psychol.* https://doi.org/10.1037/amp0000647 (2020) doi:10.1037/amp0000647.
5. Schleicher, A. Why schools need to reengage with bored learners. *OECD Education and Skills Today* https://oecdedutoday.com/why-schools-need-to-reengage-with-bored-learners/ (2025).
6. Doan, S., Steiner, E. D. & Pandey, R. *Teacher Well-Being and Intentions to Leave in 2024: Findings from the 2024 State of the American Teacher Survey*. https://www.rand.org/pubs/research_reports/RRA1108-12.html (2024).
7. Yeager, D. S., Lee, H. Y. & Dahl, R. E. Competence and motivation during adolescence. in *Handbook of competence and motivation: Theory and application* (eds Elliot, A. J., Dweck, C. S. & Yeager, D. S.) 431–448 (Guildford Press, New York, NY, 2017).
8. Sims, S. *et al.* Effective teacher professional development: New theory and a meta-analytic test. *Rev. Educ. Res.* **95**, 213–254 (2025).
9. Visscher, A. J., Dmoshinskaia, N., Pellegrini, M. & Naizaque, A. R. (When) Do Teacher Professional Development Interventions Improve Student Achievement? A Meta-analysis of 128 High-Quality Studies. *Educ. Res. Rev.* 100742 (2025) doi:10.1016/j.edurev.2025.100742.

10. Ng, A. Ed-Tech Usage Continues to Rise, Despite District Focus on Tougher Standards. *EdWeek* https://marketbrief.edweek.org/education-market/ed-tech-usage-continues-to-rise-despite-district-focus-on-tougher-standards/2024/06 (2024).
11. Buchholz, J., Cignetti, M. & Piacentini, M. *Developing Measures of Engagement in PISA*. https://dx.doi.org/10.1787/2d9a73ca-en (2022).
12. Murphy, M. C. *Cultures of Growth*. (Simon and Schuster, New York, NY, 2024).
13. Dweck, C. S. & Yeager, D. Global Mindset Initiative Introduction: Envisioning the Future of Growth Mindset Research in Education. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.3911564 (2021).
14. Murphy, M., Fryberg, S., Brady, L., Canning, E. & Hecht, C. Global Mindset Initiative Paper 1: Growth Mindset Cultures and Teacher Practices. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.3911594 (2021).
15. Hecht, C. A., Yeager, D. S., Dweck, C. S. & Murphy, M. C. Beliefs, affordances, and adolescent development: Lessons from a decade of growth mindset interventions. in *Advances in Child Development and Behavior* vol. 61 169–197 (JAI, 2021).
16. Hecht, C. A., Dweck, C. S., Murphy, M. C., Kroeper, K. M. & Yeager, D. S. Efficiently exploring the causal role of contextual moderators in behavioral science. *Proc. Natl. Acad. Sci.* **120**, e2216315120 (2023).
17. Yeager, D. S. *10 to 25: The Science of Motivating Young People*. (Avid Reader Press, New York, NY, 2024).
18. TNTP. *The Mirage: Confronting the Hard Truth about Our Quest for Teacher Development*. http://tntp.org/assets/documents/TNTP-Mirage_2015.pdf (2015).

19. Study: Billions of dollars in annual teacher training is largely a waste. *TNTP* https://tntp.org/news/study-billions-of-dollars-in-annual-teacher-training-is-largely-a-waste/ (2015).
20. Dweck, C. S. *Mindset: The New Psychology of Success*. (Random House, New York, NY, 2006).
21. Nadella, S., Shaw, G. & Nichols, J. T. *Hit Refresh: The Quest to Rediscover Microsoft's Soul and Imagine a Better Future for Everyone*. (HarperCollins, United States, 2017).
22. Canning, E. A. *et al.* Cultures of genius at work: Organizational mindsets predict cultural norms, trust, and commitment. *Pers. Soc. Psychol. Bull.* **46**, 626–642 (2020).
23. Trzesniewski, K. *et al. Global Mindset Initiative Paper 3: Measuring Growth Mindset Classroom Cultures*. https://papers.ssrn.com/abstract=3911591 (2021).
24. Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N. & Brzustoski, P. Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science* **324**, 400–403 (2009).
25. Sheeran, P. & Webb, T. L. The intention–behavior gap. *Soc. Personal. Psychol. Compass* **10**, 503–518 (2016).
26. Walton, G. M. & Crum, A. J. *Handbook of Wise Interventions*. (Guilford Publications, 2020).
27. Yeager, D. S. & Walton, G. M. Social-psychological interventions in education: They're not magic. *Rev. Educ. Res.* **81**, 267–301 (2011).
28. Aronson, E. The power of self-persuasion. *Am. Psychol.* **54**, 875–884 (1999).
29. Cialdini, R. B. & Goldstein, N. J. Social influence: Compliance and conformity. *Annu. Rev. Psychol.* **55**, 591–621 (2004).
30. Tormala, Z. L. & Rucker, D. D. Attitudes: Form, function, and the factors that shape them. in *The handbook of social psychology* (Situational Press, 2025).

31. Bryan, C. J. Values-alignment interventions: An alternative to pragmatic appeals for behavior change. in *Handbook of Wise Interventions 259–285* (2020).

32. Tipton, E., Hedges, L., Yeager, D., Murray, J. & Gopalan, M. *Global Mindset Initiative Paper 4: Research Infrastructure and Study Design*. https://papers.ssrn.com/abstract=3911643 (2021).

33. K-12 Education: Why Math, Why Now? https://usprogram.gatesfoundation.org/news-and-insights/articles/why-math-why-now-factsheet (2022).

34. Sun, K. L. The role of mathematics teaching in fostering student growth mindset. *J. Res. Math. Educ.* **49**, 330–355 (2018).

35. The Math Narrative Project. *Math Narrative Project Messaging Guide*. https://www.mathnarrative.org/wp-content/uploads/2024/12/Math-Narrative-Project_-Messaging-Guide-2024-1.pdf (2024).

36. Meyer, M., Cimpian, A. & Leslie, S.-J. Women are underrepresented in fields where success is believed to require brilliance. *Front. Psychol.* **6**, 235 (2015).

37. Chestnut, E. K., Lei, R. F., Leslie, S.-J. & Cimpian, A. The myth that only brilliant people are good at math and its implications for diversity. *Educ. Sci.* **8**, 65 (2018).

38. 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).

39. Crosnoe, R., Lopez-Gonzalez, L. & Muller, C. Immigration from Mexico into the math/science pipeline in American education. *Soc. Sci. Q.* **85**, 1208–1226 (2004).

40. Carroll, J. M. *et al.* Mindset × Context: Schools, Classrooms, and the Unequal Translation of Expectations into Math Achievement. *Monogr. Soc. Res. Child Dev.* **88**, 7–109 (2023).

41. Brown, P. C., Roediger III, H. L. & McDaniel, M. A. *Make It Stick: The Science of Successful Learning*. (Harvard University Press, 2014).

42. Toppo, G. ‘Cognitive Science,’ All the Rage in British Schools, Fails to Register in U.S. https://www.the74million.org/article/cognitive-science-all-the-rage-in-british-schools-fails-to-register-in-u-s/ (2025).
43. Boser, U. *Learn Better: Mastering the Skills for Success in Life, Business, and School, or, How to Become an Expert in Just About Anything*. (Harmony/Rodale, 2017).
44. Blazar, D., McNamara, D. & Blue, G. Instructional Coaching Personnel and Program Scalability. *EdWorkingPaper 21-499* https://doi.org/10.26300/2DES-S681 (2022) doi:10.26300/2DES-S681.
45. Blazar, D. Teacher Coaching to Improve Instruction at Scale: Opportunities and Challenges in Policy Contexts. *Teach. Coll. Rec. Voice Scholarsh. Educ.* **122**, 1–9 (2020).
46. Kraft, M. A., Blazar, D. & Hogan, D. The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Rev. Educ. Res.* **88**, 547–588 (2018).
47. Bryk, A. S., Gomez, L. M. & Grunow, A. Getting ideas into action: Building networked improvement communities in education. in *Frontiers in sociology of education* (ed. Hallinan, M.) (Springer, New York, 2010).
48. Hecht, C. A., Bryan, C. J. & Yeager, D. S. A values-aligned intervention fosters growth mindset–supportive teaching and reduces inequality in educational outcomes. *Proc. Natl. Acad. Sci.* **120**, e2210704120 (2023).
49. Herren, D., Hahn, R., Murray, J., Carvalho, C. & He, J. stochtree: Stochastic Tree Ensembles (XBART and BART) for Supervised Learning and Causal Inference. (2025).
50. Hahn, P. R., Murray, J. S. & Carvalho, C. M. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Anal.* https://doi.org/10.1214/19-BA1195 (2020) doi:10.1214/19-BA1195.

51. Yeager, D. S. *et al.* A national experiment reveals where a growth mindset improves achievement. *Nature* **573**, 364–369 (2019).

52. Yeager, D. S. *et al.* A synergistic mindsets intervention protects adolescents from stress. *Nature* **607**, 512–520 (2022).

53. Yeager, D. S. *et al.* Teacher mindsets help explain where a growth-mindset intervention does and doesn't work. *Psychol. Sci.* **33**, 18–32 (2022).

54. McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. Abandon Statistical Significance. *Am. Stat.* **73**, 235–245 (2019).

55. Gelman, A. Bayesian statistics Then and now. *Stat. Sci.* **25**, 162–165 (2010).

56. van de Schoot, R. *et al.* Bayesian statistics and modelling. *Nat. Rev. Methods Primer* **1**, 1–26 (2021).

57. What Works Clearinghouse. *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0.* https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5.0-0-508.pdf (2022).

58. 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).

59. Bryan, C. J., Tipton, E. & Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nat. Hum. Behav.* **5**, 980–989 (2021).

60. Canning, E. A., Muenks, K., Green, D. J. & Murphy, M. C. STEM faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes. *Sci. Adv.* **5**, eaau4734 (2019).

61. Bryan, C. J., Yeager, D. S. & Hinojosa, C. P. A values-alignment intervention protects adolescents from the effects of food marketing. *Nat. Hum. Behav.* **3**, 596–603 (2019).

62. Diener, E., Emmons, R. A., Larsen, R. J. & Griffin, S. The satisfaction with life scale. *J. Pers. Assess.* **49**, 71–75 (1985).

63. Bryan, C., Hecht, C., Blazar, D., Kraft, M. & Solheim, O. Global Mindset Initiative Working Paper 2: Designing an Intervention to Motivate Growth Mindset-Supportive Teaching Practices. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.3911995 (2021).

64. Chang, E. H. *et al.* The mixed effects of online diversity training. *Proc. Natl. Acad. Sci.* **116**, 7778–7783 (2019).

65. Swami, V. *et al.* Life satisfaction around the world: Measurement invariance of the Satisfaction With Life Scale (SWLS) across 65 nations, 40 languages, gender identities, and age groups. *PLoS One* **20**, e0313107 (2025).

66. Diener, E. New findings and future directions for subjective well-being research. *Am. Psychol.* **67**, 590–597 (2012).

67. Andersen, N. K., Wimmelmann, C. L., Mortensen, E. L. & Flensborg-Madsen, T. Longitudinal associations of self-reported satisfaction with life and vitality with risk of mortality. *J. Psychosom. Res.* **147**, 110529 (2021).

68. Folk, D. & Dunn, E. How can people become happier? A systematic review of preregistered experiments. *Annu. Rev. Psychol.* **75**, 467–493 (2024).

69. Folk, D. & Dunn, E. A systematic review of the strength of evidence for the most commonly recommended happiness strategies in mainstream media. *Nat. Hum. Behav.* **7**, 1697–1707 (2023).

70. Tipton, E., Spybrook, J., Fitzgerald, K. G., Wang, Q. & Davidson, C. Toward a System of Evidence for All: Current Practices and Future Opportunities in 37 Randomized Trials. *Educ. Res.* **50**, 145–156 (2021).

71. Stigler, J. W. & Hiebert, J. Lesson study, improvement, and the importing of cultural routines. *ZDM* **48**, 581–587 (2016).

72. William Schmidt: The scholar who changed math education | College of Education - Michigan State University. *Michigan State University* https://education.msu.edu/new-educator/2025/2025/03/William-schmidt-the-scholar-who-changed-math-education.

73. Roeber, E. *What Does It Mean to Use Matrix Sampling Instudent Assessment?* https://michiganassessmentconsortium.org/wp-content/uploads/ThinkPoint_MatrixSampling3.pdf.pdf.

74. Future of Testing in Education: The Way Forward for State Standardized Tests. *Center for American Progress* https://www.americanprogress.org/article/future-testing-education-way-forward-state-standardized-tests/ (2021).

75. Bryk, A. S. & Schneider, B. *Trust in Schools: A Core Resource for Improvement*. (Russell Sage Foundation, New York, NY, 2002).

76. Lipsey, M. W. *et al. Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms*. https://ies.ed.gov/ncser/pubs/20133000/ (2012).

77. Doan, S., Steiner, E. D., Woo, A. & Levine, P. R. State of the American Teacher Survey: 2025 Technical Documentation and Survey Results. https://www.rand.org/pubs/research_reports/RRA1108-17.html (2025).

78. Crum, A. J., Salovey, P. & Achor, S. Rethinking stress: The role of mindsets in determining the stress response. *J. Pers. Soc. Psychol.* **104**, 716–733 (2013).

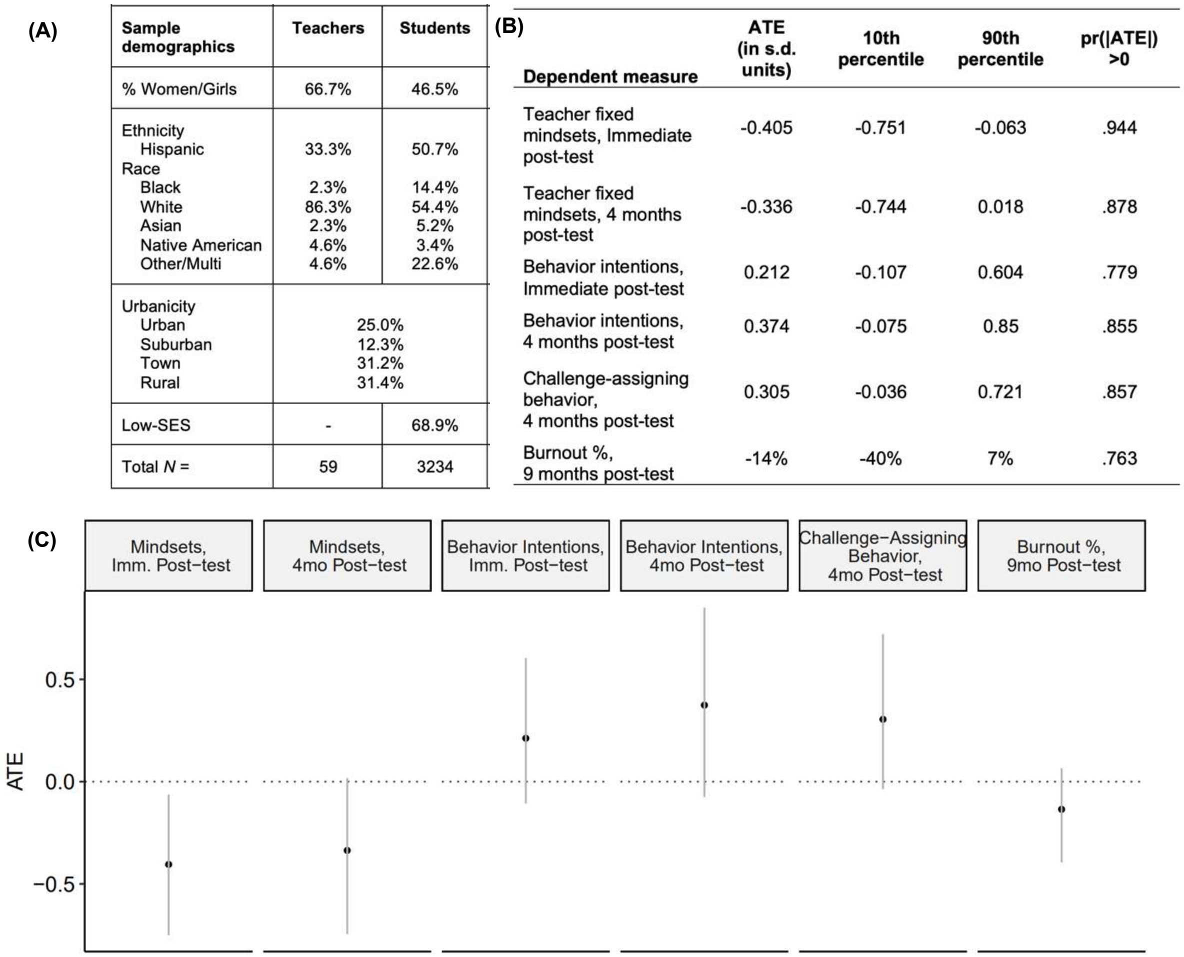
79. Kraft, M. A. Interpreting Effect Sizes of Education Interventions. *Educ. Res.* **49**, 241–253 (2020).

80. Pane, J. F., Doss, C., Todd, I. & Seaman, D. Efficacy of Zearn Math over two years in grades 3 to 5: An experiment in Texas. http://www.edworkingpapers.com/ai25-1211 (2025).

Extended Data Tables and Figures

Correlate of Teacher Fixed Mindsets in a National Sample	**Example Item**	***r* =**	**Interpretation**
Mindset "meaning system"			
"Ability" attributions	A student "is always going to struggle because they just don't have enough math ability."	.39	Fixed mindset (FM) teachers explain student struggles in terms of a lack of ability
"Support" attributions	"With the right support, the student can learn better ways to study and do well in the class."	-.40	FM teachers are unlikely to explain/understand student struggles in terms of a lack of teacher support
Effort beliefs	"If a student has to try really hard in math, it means they can't be good at that subject."	.43	FM teachers interpret a student's struggle or need for high effort as a sign that they lack ability.
Self-efficacy	"I probably can't motivate everyone of my students to work hard and learn all the material in my class."	.23	Because FM teachers explain student struggle in terms of fixed qualities, they are less confident in their ability to influence student motivation.
Behavior intentions			
Fixed mindset speeches	"Students who do the best at the beginning of the year are typically the same ones who do well at the end."	.29	FM teachers think that students' performance, like their ability, will remain constant through the year .
Intention to enact culture-of-learning practices	Likelihood of explaining that mistakes are "opportunities for the class to learn more."	-.29	FM teachers are less likely to say to their classrooms that mistakes are helpful learning opportunities.
Perceived barriers to enacting culture-of-learning	How often will something get in the way of making the culture-of-learning practices one of your top priorities?	.29	FM teachers think that it would be infeasible to, for example, allow students to revise and resubmit work, or to troubleshoot their mistakes.
Reluctance to change			
Misunderstanding of growth mindset	Growth mindset "causes teachers to lower their standards for academic rigor."	.28	FM teachers think that implementing growth mindset-supportive practices means they have to betray their standards of rigor.
Skepticism of growth mindset	Skeptical of "emphasizing your support for students' growth mindsets in your own teaching."	.40	FM teachers are generally skeptical of culture of learning practices.

Extended Data Table 1. Results of a survey of teacher beliefs and culture-building practices in a nationally-representative sample (the RAND American Educator Panel). Note. To inform our intervention approach, we conducted a survey of a U.S. representative sample of secondary math educators (*N* = 980). We examined teachers' *fixed mindsets* (their beliefs that students' abilities were fixed and cannot change) through items such as the classic item "Students have a certain amount of intelligence, and they really can't do much to change it," and via new items, such as "There's a lot of talk about things like grit or growth mindset, but deep down an experienced teacher knows that some kids have the ability to excel and others don't." Overall, this national survey suggested that a successful intervention to promote culture-of-learning practices would need to speak to teachers' internalized mindset beliefs and meaning-making tendencies that could otherwise make teachers reluctant to implement the culture of learning practices.



(A)

Sample demographics	Teachers	Students
% Women/Girls	66.7%	46.5%
Ethnicity		
Hispanic	33.3%	50.7%
Race		
Black	2.3%	14.4%
White	86.3%	54.4%
Asian	2.3%	5.2%
Native American	4.6%	3.4%
Other/Multi	4.6%	22.6%
Urbanicity		
Urban	25.0%	
Suburban	12.3%	
Town	31.2%	
Rural	31.4%	
Low-SES	-	68.9%
Total *N* =	59	3234

(B)

Dependent measure	ATE (in s.d. units)	10th percentile	90th percentile	pr(|ATE|) >0
Teacher fixed mindsets, Immediate post-test	-0.405	-0.751	-0.063	.944
Teacher fixed mindsets, 4 months post-test	-0.336	-0.744	0.018	.878
Behavior intentions, Immediate post-test	0.212	-0.107	0.604	.779
Behavior intentions, 4 months post-test	0.374	-0.075	0.85	.855
Challenge-assigning behavior, 4 months post-test	0.305	-0.036	0.721	.857
Burnout %, 9 months post-test	-14%	-40%	7%	.763

(C)
Mindsets, Imm. Post−test
Mindsets, 4mo Post−test
Behavior Intentions, Imm. Post−test
Behavior Intentions, 4mo Post−test
Challenge−Assigning Behavior, 4mo Post−test
Burnout %, 9mo Post−test
ATE
0.5
0.0
−0.5

Extended Data Figure 1. A pilot experiment (*N*= 59 teachers, *N* = 3,234 students) demonstrated the impact of the FUSE program on teachers' mindsets, behavior intentions, behavior and (potentially) burnout. Note: In the summer of 2021, 8th and 9th grade math teachers and their Algebra I students in public schools in Texas were randomized (at the teacher level) to either the treatment (culture-of-learning) or control (cognitive-science-of-learning) versions of the FUSE program. Panel (A): Demographics of teachers and students. Low-SES = Low-socioeconomic status, indexed by having a non-college-educated mother. Panel (B): Average Treatment Effects (ATEs) estimated in a Bayesian Causal Forest (BCF) statistical model, scaled in teacher-level standard deviation (*s.d.*) units, 10th-90th percentiles of the posterior distributions, and posterior probabilities that the absolute value of the ATE is greater than zero. Panel (C): Plots of results listed in Panel (B).

Element	Justification	Implementation
Honorific, non-stigmatizing framing to reduce reactance	People tend to reject invitations to change that come across as trying to "fix" someone's flaws[1].	FUSE is framed as a high-status "fellowship" that teachers are invited to by district authorities.
Values-alignment	People tend to reject behaviors that are misaligned with their motivational priorities[2]	FUSE is framed as aligned with teachers' #1 value: *Being the kind of teacher who inspires enthusiastic learning and compliance.*
Social influence	People tend to reject behaviors that are "out of step" with what they perceive to be normal[3].	Thus FUSE utilizes "social proof," in the form of testimonials from current Texas math teachers who have used the practices.
Autonomy-supportive explanations	People react negatively when they are commanded to adopt a given behavior without being granted agency[4].	Therefore the FUSE materials carefully explain the rationale behind each suggested practice, by explaining how the new language taps into adolescent values.
Self-persuasion	People internalize behaviors more readily when they have freely advocated to others that they should do the behavior[5].	FUSE asks fellows to publicly advocate for the recommended practices, for example by presenting to peers, and coaching sessions with peers in which they describe progress.
Stress-can-be-enhancing mindset	People tend to perceive stress as debilitating of their performance and outcomes, and this mindset undermines effective coping with stress[52,78]	FUSE asks teachers to complete the "synergistic mindsets" intervention, which trains them on a stress-can-be-enhancing mindset.[52]

Extended Data Table 2. Social-psychological elements of the treatment (culture-of-learning) group in the Fellowship Using the Science of Engagement (FUSE) program.

(A)

Outcome	CATE (in *s.d.* units)	CATE Lower Bound	CATE Upper Bound	pr(ATE)>0	***Teacher N=***
Fixed Mindset	-0.319	-0.475	-0.167	.997	142
Enforcer Mindset	-0.345	-0.512	-0.180	.996	141
Mentor Mindset	0.326	0.138	0.515	.985	141
Intended Treatment Practices	0.488	0.290	0.687	>.999	140
Recalled Treatment Practices	0.203	0.009	0.401	.911	126
Fixed Mindset Language	-0.139	-0.322	0.049	.817	132
Culture of Learning Responses to Mistakes	0.511	0.294	0.718	>.999	136
Intended Control Practices	-1.346	-1.562	-1.142	>.999	139
Recalled Control Practices	-0.374	-0.575	-0.167	.989	126

(B)

Outcome	ATE (in *s.d.* units)	ATE Lower Bound	ATE Upper Bound	pr(ATE)>0	***Teacher N=***
Respect	0.163	-0.017	0.353	.875	152
Culture of Learning	0.160	0.028	0.297	.938	138
Learning from Mistakes	0.281	0.079	0.483	.961	138
Classroom Fixed Mindset	-0.233	-0.401	-0.071	.967	152
Classroom Engagement	0.447	0.253	0.646	.997	132

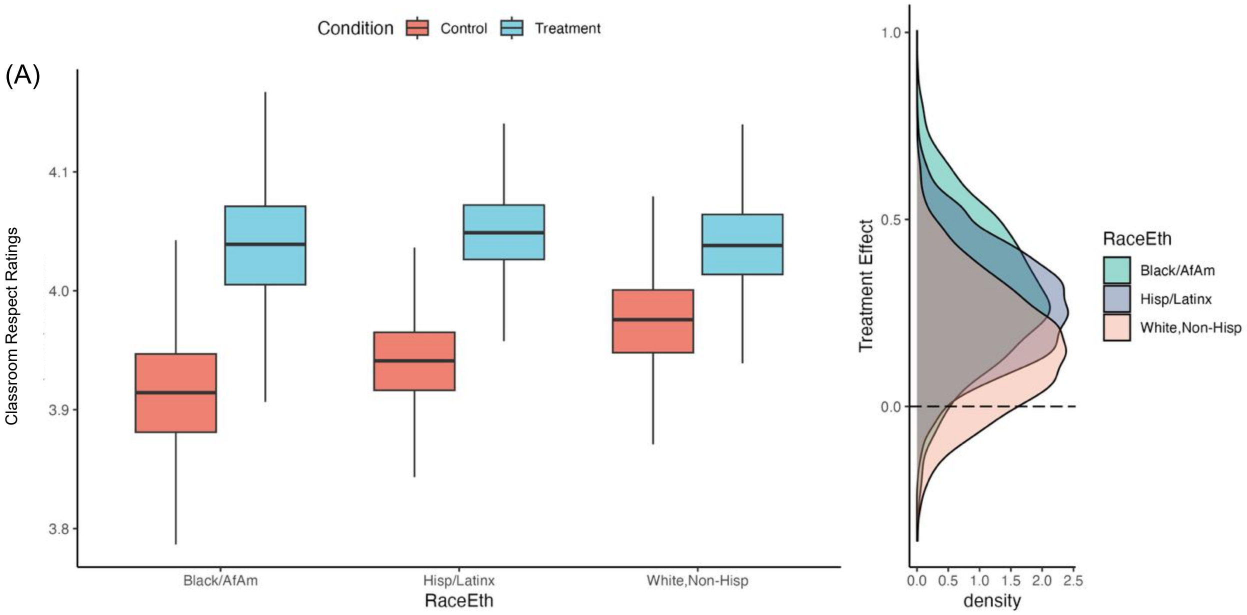
(C)

Outcome	CATE (in *s.d.* units)	CATE Lower Bound	CATE Upper Bound	pr(ATE)>0	***Teacher N=***
Burnout (Fall)	-0.240	-0.410	-0.062	.955	133
Burnout (Spring)	-0.111	-0.300	0.078	.777	126
Satisfaction with Life	0.278	0.075	0.477	.959	126
Attrition Intentions	-0.044	-0.256	0.170	.611	132

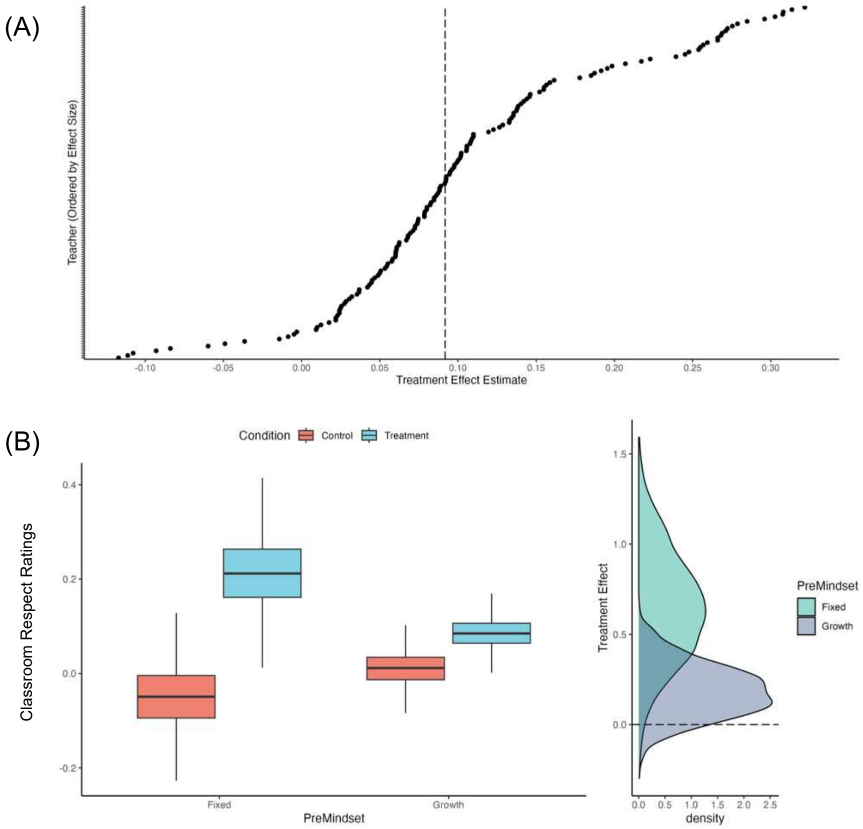
Extended Data Table 3. Point estimates and Average treatment effects (ATEs) for (A) teacher beliefs and behaviors; (B) classroom-aggregate outcomes, and (C) teacher burnout and well-being. Note: Results from BCF models. Posterior probabilities correspond to the probability of an effect in the expected direction.

Moderator Subgroup	Control Group *m*	Treatment Group *m*	CATE (in scale units)	CATE (in *s.d.* units)	CATE Lower Bound	CATE Upper Bound	pr(CATE)>0
All teachers and students	-0.167	-0.046	0.122	0.253	0.105	0.405	.989
Student Race/ Ethnicity							
Black/African-American	-0.153	0.047	0.200	0.417	0.165	0.664	.988
White, Non-Hispanic	0.008	0.148	0.140	0.292	0.112	0.479	.981
Hispanic/ Latinx	-0.185	-0.065	0.120	0.249	0.098	0.403	.984
Asian	0.229	0.411	0.182	0.380	0.057	0.700	.936
Other	-0.372	-0.312	0.060	0.125	0.329	0.782	.782
Teacher Mindset							
Fixed	-0.006	0.198	0.203	0.424	0.121	0.738	.961
Growth	0.001	0.105	0.104	0.217	0.061	0.371	.962
Trust							
Low	-0.132	-0.017	0.230	0.480	0.247	0.729	.997
High	0.130	0.154	0.024	0.050	-0.182	0.283	.606

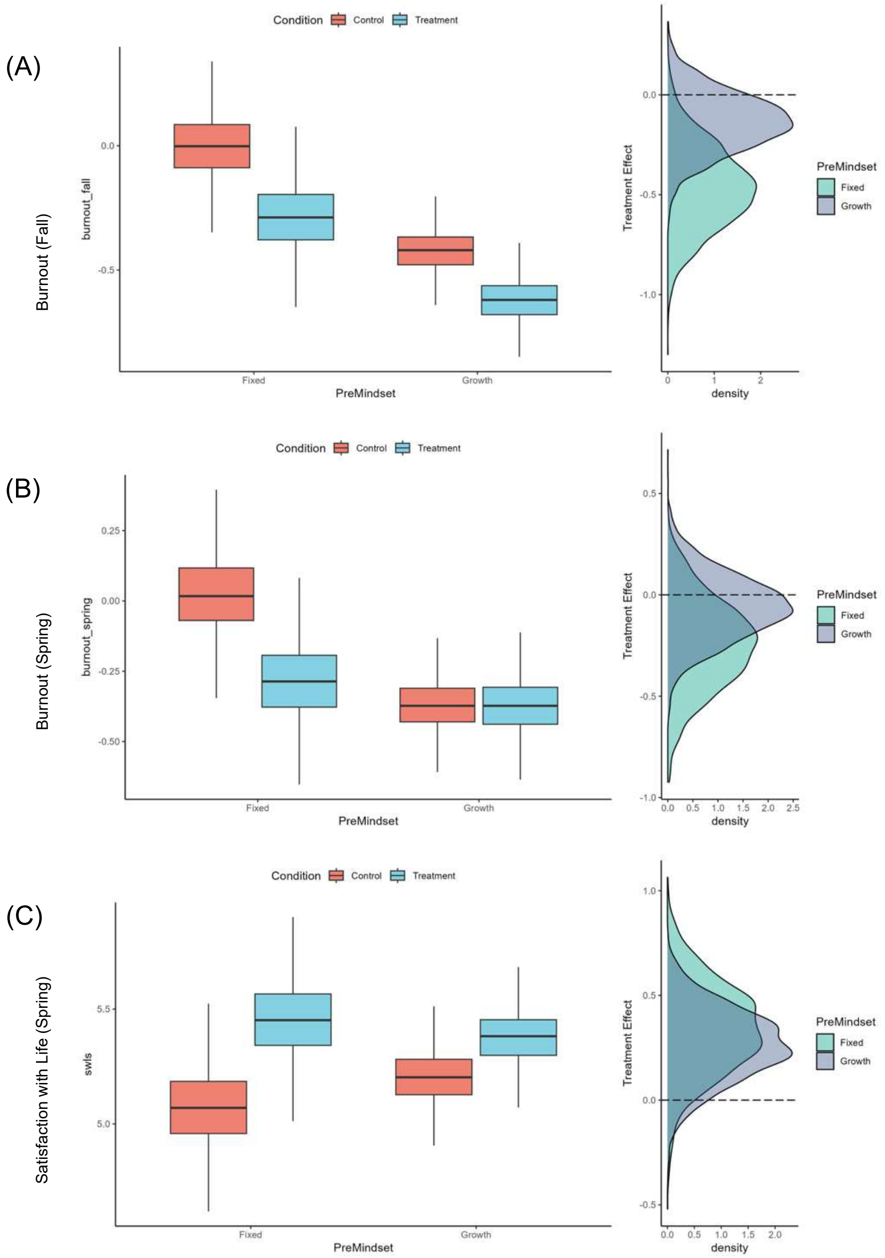
Extended Data Table 4. Classroom math test performance average treatment effects (ATEs) and conditional average treatment effects (CATEs) overall and across subgroups. Note: Results from multilevel BCF models, which were fit by nesting students within teachers and estimating a random intercept for each teacher. During summarization and analysis of race/ethnicity CATES, model fits were aggregated to "local identity groups" (i.e. race/ethnicity groups within teachers) and then averaged to yield teacher-level treatment impacts. For analysis of between-teacher moderators (fixed mindset and trust), results were first aggregated to yield estimates of each teacher's treatment impact and then summarized across moderators. Note that moderators were included as fully continuous variables in the model fit, to avoid potentially arbitrary choices about subgroup cut-points for the moderators. Correlations of estimated effect sizes with the full, continuous moderators are reported in the main text. Here, the cut-point for fixed mindset was a score of 3 or higher (out of six) because that represents the point at which a teacher does not mostly "agree" with the fixed mindset statements, following prior research. Because the measure of trust was new and adapted for the present research, no prior cut-point was available; the subgroups for high and low trust were therefore defined by the top and bottom terciles. For moderation by race/ethnicity: pr($\text{Diff}_{\text{CATEs Black vs. White}}$>0) = .755; pr($\text{Diff}_{\text{CATEs Hispanic vs. White}}$>0) <.75; pr($\text{Diff}_{\text{CATEs Asian vs. White}}$>0) <.75; pr($\text{Diff}_{\text{CATEs Other vs. White}}$<0) = .855; For moderation by teacher mindset: pr($\text{Diff}_{\text{CATEs}}$>0)=.787; For moderation by trust: pr($\text{Diff}_{\text{CATES}}$)=.934. ***A note about effect sizes**:* CATEs in scale units are on the outcome measure's z-score scale, with a mean of 0 and an SD of 1, and therefore are comparable to student-level standardized effect sizes that are common in the literature. However as noted in the main text, the present study was not designed to estimate student-level effect sizes due to random error variation in the student-level test scores, which was included by design. Nevertheless, the effect sizes presented in column 3 are large compared to benchmarks in the literature. The median effect size in the literature on secondary students' academic achievement is 0.05 *s.d.* [79]. Even large and comprehensive reforms, such as the implementation of new math teaching platforms that students use for hours every week for years, tend to yield effects around .05 *s.d.*[80]. Anything over 0.20 *s.d.* is considered a large effect size[80]. Indeed, an entire year of learning in a typical school tends to be around 0.20 *s.d.* for secondary math students [76]. And the present study found credible effect sizes of .20 *s.d.* in a pre-registered, conservative heterogeneity analysis, for Black students, students with fixed mindset teachers, and students with low-trust teachers—all groups who tended to have overall higher levels of risk for poor outcomes. Thus the effect sizes presented in column 3 are large compared to benchmarks, especially for vulnerable groups.



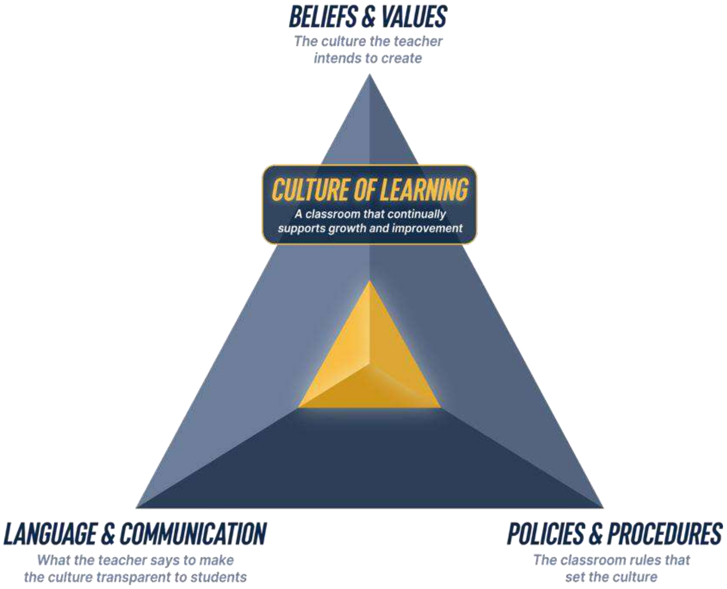
Extended Data Figure 2. Heterogeneity in the impacts of the FUSE culture-of-learning program on classroom culture of respect across student racial and ethnic groups. Note: Results from multilevel BCF models, which were fit by nesting students within teachers and estimating a random intercept for each teacher. During summarization and analysis of race/ethnicity CATES, model fits were aggregated to "local identity groups" (i.e. race/ethnicity groups within teachers) and then averaged to yield teacher-level treatment impacts. Posterior probabilities of interaction effects: pr($\text{Diff}_{\text{CATEs Black vs. White}}$>0) = .842; pr($\text{Diff}_{\text{CATEs Hispanic vs. White}}$>0) =.874.



Extended Data Figure 3. Between-teacher heterogeneity in the impacts of the FUSE culture-of-learning program on classroom culture of respect (A), as a function of teacher mindset (B). In (A), each dot is a teacher's estimated treatment impact in scale units and the dashed line is the median of the distribution of unstandardized treatment impacts (i.e. the median treatment effect). For (B), the posterior probability of a difference in conditional average treatment effects (CATEs) across teacher mindset subgroups, pr(Diff$_{\text{CATES}}$) = .949; for (C), pr(Diff$_{\text{CATES}}$) < .75. The distribution of subgroup treatment effects in the right panel is in *s.d.* units.

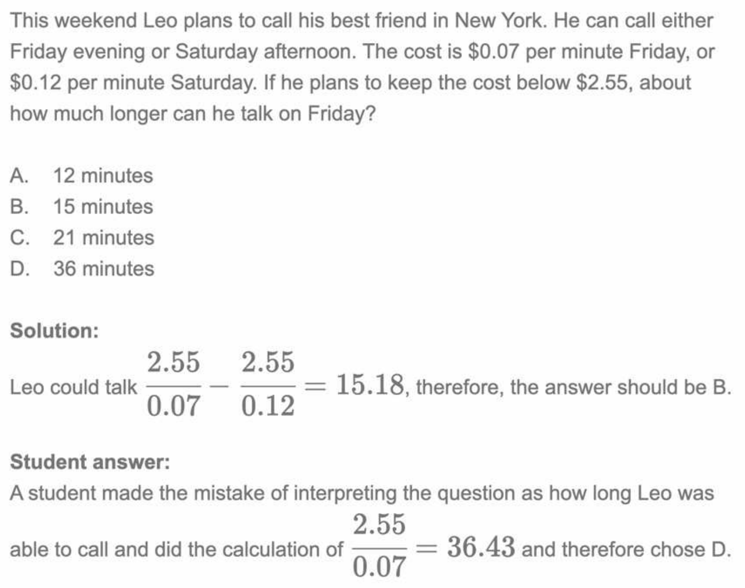


Extended Data Figure 4. Subgroup differences in the impacts of the FUSE culture-of-learning program as a function of teacher fixed mindset, for (A) reduced teacher burnout in the Fall, (B) reduced teacher burnout in the Spring, and (C) greater satisfaction with life in the Spring. In terms of subgroup differences, teachers with prior fixed mindsets in the control condition reported substantially greater burnout in the fall (October), Difference in means (Diff$_m$) = 0.61 *s.d.,* pr(Diff$_m$>0)>.999, and spring (March), Diff$_m$ = 0.44 *s.d.,* pr(Diff$_m$>0)=.993. Subgroup differences in treatment impacts on burnout: Fall CATE$_{\text{Fixed}}$ = -0.56 *s.d.*, pr(CATE<0)=.991, CATE$_{\text{Growth}}$ = -0.18, pr(CATE<0)=.898, pr(Diff$_{\text{CATEs}}$)=.958; Spring CATE$_{\text{Fixed}}$ = -0.30 *s.d.*, pr(CATE<0)=.884, CATE$_{\text{Growth}}$ = -0.08 *s.d.*, pr(CATE<0)=.702, pr(Diff$_{\text{CATEs}}$)=.820. Subgroups cut at the same cutpoint for SWLS were not meaningfully different, (pr(Diff$_{\text{CATEs}}$)<.75, but as shown in Figure 8 in the main paper, there was a strong moderation for prior mindset.


BELIEFS & VALUES
The culture the teacher intends to create
CULTURE OF LEARNING
A classroom that continually supports growth and improvement
LANGUAGE & COMMUNICATION
What the teacher says to make the culture transparent to students
POLICIES & PROCEDURES
The classroom rules that set the culture

Extended Data Figure 5. The FUSE "culture of learning" pyramid. This is the primary graphic used in the FUSE program to convey the three cornerstones on which a culture of learning is built. First, teachers are induced to examine their beliefs about students—specifically, whether they believe that all can grow and learn in a properly-supportive environment. Second, teachers examine their classroom policies which either do or do not manifest those beliefs about student growth and potential, such as whether students can earn points back for revising or retaking assessments. Third, teachers draft language and communication styles that advertise for their pro-growth and pro-learning policies, and that transparently express teachers' beliefs about students.

(A)



This weekend Leo plans to call his best friend in New York. He can call either Friday evening or Saturday afternoon. The cost is \$0.07 per minute Friday, or \$0.12 per minute Saturday. If he plans to keep the cost below \$2.55, about how much longer can he talk on Friday?

A. 12 minutes
B. 15 minutes
C. 21 minutes
D. 36 minutes

Solution:

Leo could talk $\frac{2.55}{0.07} - \frac{2.55}{0.12} = 15.18$, therefore, the answer should be B.

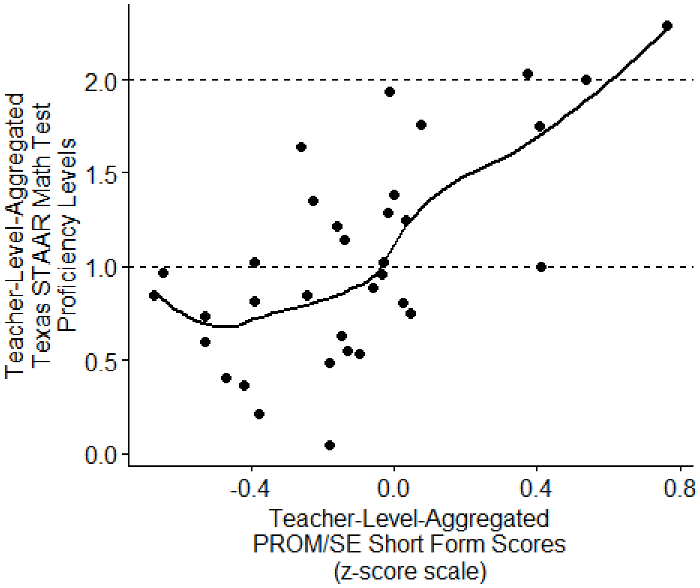
Student answer:

A student made the mistake of interpreting the question as how long Leo was able to call and did the calculation of $\frac{2.55}{0.07} = 36.43$ and therefore chose D.

(B)

Teacher language in response to mistakes	**Definition of content analysis code**	**Example of teacher response**
Culture of Learning		
Surface student's mistaken thinking	Asks questions or lets the student explain their thinking, with a genuine intent to understand how they solved the problem.	"Can you tell me what you were thinking when you set up this fraction? Why did you put 2.55 on top? Why did you divide it by 0.07?"
Validate student's correct thinking	Validate what the student did right or note their progress. The validation should feel sincere, respectful, and confidence-building, showing students they're not starting from scratch but have something to build on. Praise should go beyond just recognizing that the student attempted the problem.	"With the equation of 3k, you've correctly identified the rate of change. They are adding 3 match sticks each time, but what happens if you plug a 1 into your equation?"
Culture of Judgment and Evaluation		
Autonomy threat	The teacher gives directive or overly detailed instructions, telling students what to do and leaving little room for independent thinking. This may include using a strict or pressuring tone, giving step-by-step commands, or doing the reasoning for the student instead of supporting their own thinking.	I would say for the student, that he should read the question 2-3 times. Also, in the last sentence, key word was longer, showing that it has to be a subtraction problem between the 2 days(Friday and Saturday) and then find the total number of minutes for each day.
Harsh judgment	Language that focuses on what the student did wrong, in a way that can make them feel criticized or at fault rather than supported in learning. This can include statements or questions that emphasize mistakes or carelessness, or use a tone that could feel sarcastic, dismissive, or like a "gotcha."	So, when you answered it you did set it up using a fraction bar but did you think unit rate? (no) Did you use CUBES? (no).

Extended Data Figure 6. A measure of teachers' responses to mistakes: (A) Example stimulus; (B) Coding scheme. Note: At each post-randomization measurement occasion, teachers were presented with three of the stimuli shown above, each of which has a math problem, a correct solution, and a student mistake. After each one, teachers were asked "As the student's teacher, what would you say to them?" Teachers' free responses were coded into *culture of learning practices* (surface student's mistaken thinking; validate student's correct thinking) and *culture of judgment and evaluation practices* (autonomy threat; harsh judgment). The full codebook, with additional examples and justifications, and all additional stimuli, appear in the online supplemental materials.



Extended Data Figure 7. Validation of the PROM/SE math assessment relative to the Texas STAAR test. Note: The overall correlation between teachers' average scores on the two measures was *r* = .65. Consistent with the validity claim, 80% of teachers whose students scored above the median on the PROM/SE also averaged at least "approaches grade level" on the STAAR test for their classes (i.e. average of 1 or greater). Among those with lower aggregate scores on the PROM/SE, just 33% averaged at least "approaches grade level."

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementforYeageretalFUSEevaluation.docx