Science
AAAS

Supplementary Materials for

Where and with whom does a brief social-belonging intervention promote progress in college?

Gregory M. Walton *et al*.

Corresponding author: Gregory M. Walton, gwalton@stanford.edu

Science **380**, 499 (2023)
DOI: 10.1126/science.ade4420

The PDF file includes:

Materials and Methods

Overview

The College Transition Collaborative's Social-Belonging Trial is a multi-site randomized-controlled trial using a between-subjects design with three experimental conditions at the individual level. Here we report the active control condition and the standard belonging condition, with materials constant across school contexts. A third condition, which adapted belonging materials for each campus, will be reported separately, as the present paper focuses on contextual not material heterogeneity.

The CTC trial was designed to examine heterogeneity in multiple ways. First, we randomized a large sample of students within schools, maximizing power within and across sites. Second, we used an active control condition, to isolate the treatment effect. Third, we assessed a common objective outcome, first-year full-time completion obtained from school records. Fourth, we assessed a common manipulation check across school sites, to ensure that heterogeneity in outcome effects does not reflect variability in initial uptake (*47*). Fifth, we included diverse post-secondary institutions, including broad-access public universities, public flagship universities, liberal arts colleges, and elite private universities. These institutions represent all regions of the United States, serve diverse student populations, and vary widely including in selectivity, resources, and student demographics (Table S4). Sixth, we examined heterogeneity at the level of the local-identity group (K=374). Relative to alternative approaches, such as testing college-level effects or examining static canonical groups, this approach provides a far more nuanced and precise assessment of each group's specific psychological circumstance and greater statistical power to detect sources of heterogeneity. Seventh, especially because moderation analyses can be unreliable, analyses follow a rigorous multistep preregistration

(https://osf.io/bydwf/). We also use a flexible but conservative method, Bayesian Causal Forest (BCF) to confirm heterogeneity and detect nonlinearity (*35*).

Post-secondary institutions implemented randomized treatment and control materials through existing online prematriculation processes in the summer before students entered college. These materials drew directly on past content (*8*, *10*, *12*). This was a text-based reading-and-writing task which, including manipulation check and demographic measures, lasted in total approximately 30 minutes. The following spring a subsample of participating students in each cohort completed an online follow-up survey assessing relevant psychological and behavioral measures. Academic outcomes were obtained from institutional records.

Primary intent-to-treat analyses include students who saw the first page of randomized content whether they completed the materials or not (see Tables S1-S3 for completion rates and fidelity measures). Additionally, participants had to have available outcome data from school records and necessary demographic data to be included in analyses (N=26,911; see Fig. 2). Participating students belonged to one of 374 local-identity groups, defined by race-ethnicity, first-generation-status, college, and cohort.

Immediately after the randomized content, students completed the key manipulation check: *anticipated growth in belonging*. They reported the level of belonging they anticipated experiencing at the beginning of the first year and at the end of their second year. Analyses focused on the difference score, to index anticipated growth in belonging over this period.

The primary outcome was whether students completed the first year full-time enrolled in each semester, obtained from institutional records. Analyses included two person-level covariates commonly used in higher-education, gender and high school standardized test score, from institutional records.

We expected two group-level factors would moderate results: *group historic achievement level* and *belonging affordances*. To assess group historic achievement, we examined performance along the primary outcome, first-year full-time completion rates, in each local-identity group using data provided by each partner institution over 2-4 years prior to the study (see Table S7). To assess belonging affordances, we examined responses among the subsample of control-condition students in each local-identity group who completed a spring-term survey to four belonging items.

Partnership and IRB

The College Transition Collaborative (CTC). CTC is a research-practitioner partnership that conducts research and develops and evaluates practices designed to support belonging, growth, and equity in post-secondary contexts. CTC is based at Stanford University, with researchers and staff across North America.

Partner institutions. A total of 22 US colleges participated. Sixteen colleges participated in the 2015 and 2016 cohorts (Allegheny College; Bowling Green State University; California State University, Northridge; The College of Wooster; Cornell University; Dartmouth College; DePauw University; Hope College; Indiana State University; Indiana University; Lewis & Clark College; Southern Oregon University; University of California, Santa Cruz; University of Central Arkansas; Wabash College; Yale University). An additional six colleges participated in the 2016 cohort only (Albion College; California State University, Dominguez Hills; Kalamazoo College; Ohio Wesleyan University; University of Oregon; University of Pittsburgh) (for full institutional details, see Table S4). One more institution, a selective public STEM-focused university in Canada, used a different design than the US colleges and thus will be reported separately.

In general, institutions reached out to our team expressing interest in working together following media coverage of our prior intervention research (*12*, *48*). All agreed to basic parameters of the study, including to embed intervention materials in online protocols for incoming students; to recruit incoming students to complete these materials in the summer before matriculation; to randomize students to condition; to collaborate with our team to create a campus-specific customized condition, which included a structured campus visit, student focus groups, and iterative collaborative writing and revision; to share relevant student academic record data; and to contribute funding for the project.

This partnership is one example among others of the power of "team science" and shared research infrastructure for tackling major questions critical for progress but too big for any individual researcher to answer alone. The notable features of our consortium, which are shared with others (e.g., Center for Open Science, Strengthening Democracy Challenge), is (1) the multi-institution collaboration; (2) the coordination of methods, data, and measures across institutions allowing for cross-context comparisons, which would not be possible if an ad hoc or post hoc approach was taken, as in meta-analysis, where each team designs their own study and measures; (3) complex pre-registration vetted by large teams; and (4) the creation of datasets that have many secondary analysis purposes and can lead to many more papers than a single evaluation study (e.g., *49–51*). What we add is that we are not simply testing theoretical hypotheses but evaluating a real, policy-relevant solution, in partnership with institutions who can (and in many cases did) adopt and elaborate on the solution at the end of the study. This requires an additional layer of data coordination and harmonization (to acquire data across diverse sites and clean and process such data), but it ultimately leads to a shorter path from a basic scientific hypothesis test to the real-world dissemination of a policy solution.

Consent and Institutional Review Board (IRB) clearance. As the primary purpose of the project was to understand and improve student outcomes, CTC received IRB clearance for this work as a quality improvement project through Stanford University. Later, researchers submitted IRB applications for secondary analysis of the data for research purposes, including those reported here. Each partner institution was responsible for obtaining clearance with their own IRB. Partner IRBs either accepted Stanford's clearance and submitted letters indicating that decision, applied for and received their own approval as quality improvement, or required that liaisons obtain approval as academic research.

Where required, partner institutions provided their own consent forms. For schools that considered this project to be quality improvement, CTC provided standard consent form language. The consent form assured students that there were no anticipated risks, that their responses are confidential, and that they could withdraw their consent and participation at any time. It also indicated that, "by agreeing to participate, you are consenting to the examination of your academic and other records from [school name] in conjunction with this project." Students were further invited to contact the researchers at any time if they wished not to release those records.

Pre-Registration

One contribution of the present research is our approach to pre-registration. It differs from the practices that are common for laboratory studies with simpler experimental designs, in that our approach is designed to be a transparent and public record of the development of our thinking in light of developments in theory and new findings between when the study was first planned (2014) and now (2023). Our approach was guided by four goals: (1) to publicly register the study design, manipulations, and data collected (i.e. to address the "file drawer" problem);

(2) to record our hypothesis and planned analyses at each stage of development; (3) to use good judgment to present analyses that will be most accurate and defensible and best test the theory at hand, even if they are not what was pre-registered; and (4) to transparently disclose where the plans were followed, where they were changed, why they were changed, and what impact the changes had on the results.

Guided by this philosophy, our project had two pre-registered phases (https://osf.io/bydwf/). The pre-registration for Phase 1 (PR1) was submitted on November 11, 2016 (modifications on January 12, 2017; addendum on February 15, 2018). The pre-registration for Phase 2 (PR2) was submitted on October 15, 2018 (addendum on November 9, 2018). (A thorough disclosure of what data had been analyzed, and what data had not been merged yet, appears in PR2.) As the project developed, our theoretical focus narrowed in several respects. For instance, when the project began (PR1), we had not yet developed the concept of local-identity groups, which first appeared in PR2. Therefore, the initial study design implemented stratified randomization on the level of major demographic categories (race-ethnicity, gender, and first-generation status, within college). In addition, while we originally considered several outcomes (e.g., first-year GPA), we came to focus on first-year full-time completion rates because it is the single most important indicator of academic progress. In addition, full-time status, but not GPA, remains valid across campuses that differ in both grading norms and course selection practices (e.g., if students take easier courses or drop hard courses, it could hurt their progress to degree completion but artificially inflate their GPAs). Therefore, full-time status is better-suited for the cross-campus heterogeneity analyses. Finally, in addition to pre-registering the belonging affordances measure, we pre-registered an alternative focused on self-reported levels of stereotype threat. As our theorizing developed, we moved beyond this to focus on the

interplay of belonging affordances and vulnerability to belonging concerns as assessed by group historic achievement level. Below we list key elements of the pre-registration relevant to the present paper, how they aligned with the analyses reported, and any modifications made.

Component	Pre-Registration	Present paper
Data sample: Cohorts	Cohorts 1 and 2 (PR2)	Pre-registration followed as written
Data sample: Colleges	21 colleges (16 in Cohort 1, 5 in Cohort 2) (PR2)	Pre-registration followed as written, with modification (a)
Data sample: Canadian college	Exclusion of Canadian college (PR2)	Pre-registration followed as written
Analytic sample: ITT	Participants who saw the first page of randomized content (PR1 & PR2)	Pre-registration followed as written
Analytic sample: TOT	Participants who wrote appropriate text in saying-is-believing prompt (PR2)	Pre-registration followed as written, with modification (b)
Analytic sample: ITT & TOT	Planned to conduct analyses for both ITT and TOT (PR2)	Pre-registration followed as written
Definition of Race: Categories	Defined race-ethnicity following Census categories (PR2)	Pre-registration followed as written
Definition of Race: Self-report	Privileged self-report data and used institutional data when self-report data missing (PR1 & PR2)	Pre-registration followed as written
Definition of Race: Order of classification	Order specified as Hispanic, Black, White, Asian, Native Other (PR2)	Pre-registration followed as written
Definition of first-gen.: Self-report	Privileged self-report data and used institutional data when self-report data missing (PR2)	Pre-registration followed as written
Analytic sample: Exclusion by race	Exclude students missing race-ethnicity from all sources (PR1)	Pre-registration followed as written
Analytic sample: Exclusion by first-gen. status	Specify procedures for including students with unknown first-gen. status (PR2)	Pre-registration followed as written
Definition of "disadvantaged": Canonical	Hispanic, Black, Native, or Other-race students and students of any race-ethnicity who is a first-generation college student (PR2)	Pre-registration followed as written
Discussion of canonical disadvantage	Recognized that not all groups defined as "disadvantaged" may show an achievement gap at a given institution (PR1); Discussed broader limitations (PR2)	Pre-registration followed as written
Definition of local-identity groups	Introduced "local-identity groups," defined as students of the same race-ethnicity, first-generation status, college, and cohort (PR2)	Pre-registration followed as written
Treatment coding: Customized treatment	Exclusion of customized treatment (PR2)	Pre-registration followed as written
Research question: Heterogeneity	Primary goal to understand variability in treatment effects, particularly across local-identity groups (PR2)	Pre-registration followed as written
Definition of moderators: Level	Moderators specified at the local-identity group level (within institution) (PR2)	Pre-registration followed as written
Definition of moderators: Belonging affordances and latent threat	Identified two candidate moderators, latent stereotype threat and belonging affordances (PR2)	Pre-registration followed, with modification (c)

Definition of moderators: Group historic achievement	Identified historic achievement relative to White, continuing-generation students as a measure of local disadvantage and a potential moderator (in lieu of canonical societal disadvantage) (PR1 & PR2)	Pre-registration followed, with modification (d)
Moderator construction: Belonging affordances and latent threat	Defined procedure for constructing latent group factors from spring-term survey, including use of Empirical Bayes estimates (PR2)	Pre-registration followed as written
Moderator validation	Established validity of moderators by correlating latent factors with other theory-relevant factors (Table S8) (PR2)	Pre-registration followed as written
Outcomes	Focused on full-time enrollment in the first year and first-year GPA (PR2)	Pre-registration followed, with modification (e)
Definition of outcomes	Defined continuous full-time enrollment (PR1 & PR2)	Pre-registration followed as written
Definition of manipulation check	Defined manipulation check as anticipated growth in belonging (PR1 & PR2)	Pre-registration followed as written
Hypotheses: 3-way	Is there a 3-way interaction between condition, latent-identity group belonging affordances (threat) and canonical disadvantaged status (PR2)	Pre-registration followed, with modification (f)
Hypotheses: Predicted treatment benefits	Anticipated greater benefits for students in local-identity groups with higher threat/lower belonging (water-on-parched soil) metaphor (PR2)	Pre-registration followed, with modification (g)
Analytic model: Treatment heterogeneity	Planned to use FIRC model (*52*) (PR2)	Pre-registration followed as written
Analytic model: Fixed effects	Planned to use fixed intercepts for each local-identity group (PR2)	Pre-registration followed, with modification (h)
Analytic model: Random effects	Treatment effect allowed to vary across local-identity groups (PR2)	Pre-registration followed as written
Analytic model: Dichotomous outcomes	Discussed using the linear probability model (OLS) for dichotomous outcomes (PR2)	Pre-registration followed as written
Analytic model: Functional form of interaction	Did not specify the functional form; could be linear or otherwise (PR2)	Pre-registration followed, with modification (i)
Analytic model: Selection of student covariates	Would control for student gender (PR1 & PR2), standardized test (ACT/SAT) scores (PR2), and self-perceived socioeconomic status (PR2)	Pre-registration followed, with modification (j)
Analytic model: Centering of student covariates	Person-level covariates to be centered within local-identity group	Pre-registration followed as written
Analytic model: Missing values on potential covariates	Discussed using missing value dummy indicators for missing values on potential covariates	Pre-registration followed, with modification (k)
Analytic model: Handling small groups	Collapse across cohorts when local-identity groups were small	Pre-registration followed, with modification (l)
Simple effects testing	Test simple effects at 25th, 50th, and 75th percentile among local-identity groups of belonging affordances/latent threat	Pre-registration followed, with modification (m)

Note. PR1=Pre-registration 1; PR2=Pre-registration 2.

Modifications:

a) *Data sample: Colleges*. One college in Cohort 2, which had not provided academic data at the time of PR2, was added. This decision was made without knowledge of the treatment impacts in this college and prior to conducting the primary analyses for PR2.

b) *Analytic sample: TOT*. We included participants who wrote any text whatsoever in response to the saying-is-believing prompt, as we judged this simpler and less-subjective than hand-coding each participant's writing samples and judging whether the responses were valid.

c) *Definition of moderators: Content.* At the time of pre-registration we identified another moderator besides afforded belonging (self-reported levels of stereotype threat). However, prior to conducting these analyses we became aware of findings from a similar experiment (the NSLM) (*35*) that led us to narrow our interest in the intersection of just two moderators: belonging affordances and vulnerability to belonging concerns as assessed by group historic achievement level. Therefore, we did not conduct analyses using self-reports of stereotype threat. The stereotype threat variable is included in the publicly available dataset.

d) *Definition of moderators*: *Group historic achievement.* When we conducted analyses using the pre-registered variable of disparities (or "gaps") in achievement between each local identity group and a reference group (White, continuing-generation students), we found the same results as reported in the manuscript. However, when writing the paper we identified theoretical limitations of selecting an advantaged group as the reference (e.g., it could imply that this group is the "norm"). Therefore, we modified the moderator to be each group's overall level of achievement relative to 100% full-time enrollment. As mentioned, however, the results and conclusions were the same. The relative gap variable is included in the publicly available dataset.

e) *Outcomes*. We pre-registered several outcomes, including first-year GPA. In the paper we focus on first-year full-time completion rates only because (a) this is the outcome variable in the large experiments we were seeking to replicate (i.e., (*12*), Studies 1-2); (b) it is the single most important indicator of progress toward a degree for students; (c) first-year GPA is limited for cross-group and cross-college comparisons, because of variability in grading practices, course selection, and dropping rates, which can inflate or deflate GPAs. We also note that, when we wrote PR1, we stated that full-time completion was not a relevant outcome at all institutions, because some institutions have very high full-time rates. What we failed to realize, but corrected in PR2, is that there is considerable variability within institutions among local-identity groups in full-time completion rates, and that is the variation we were trying to explain.

f) *Hypotheses: 3-way interaction (Treatment × Belonging affordances × Canonical group disadvantage)*. The pre-registration said we would test whether the interaction between treatment and belonging affordances depended on group identity. In the pre-registration we stated this would be tested with respect to canonical group status but the paper focuses on the primary variable of local-identity group historic achievement, to better reflect our hypotheses about how advantage and disadvantage as a function of group identities varies with contexts (see "Discussion of canonical disadvantage").

g) *Hypotheses: Predicted treatment benefits*. Originally, we focused on the "water-on-parched soil" metaphor; that is, we anticipated greater benefits for students in more hostile climates, following the emphasis of past empirical work (*12*, *15*). However, with deeper theorizing in time we came to focus on the "seed-and-soil" metaphor, and thus anticipated greater benefits for students in local-identity groups who reported greater belonging (*39*). (The addendum to pre-registration 2 introduces the notion of belonging affordances.) Three factors contributed to this development. First, we came to appreciate the fact that the identity-group measure of belonging was assessed in the spring-term, many months after students had entered college. It thus reflected the opportunity students in a given group had to attain belonging over the first year. Second, we learned from evidence for the importance of positive affordances or "sustaining environments" (*53*) in the National Study of Learning Mindsets (*35*, *36*). Third, we learned to distinguish vulnerability to a psychological threat (previously defined in terms of static canonical disadvantage; here defined in terms of low local-identity group historic achievement) from the opportunity students have in a setting to overcome this vulnerability (e.g., to come to belong, i.e., belonging affordances) (*39*).

h) *Analytic model: Fixed effects*. A model including a fixed effect for each of the 374 sites did not converge in the multilevel model algorithm. Thus, we used separate fixed effects for each of the 12 race-ethnicity × first-generation status groups and each of the 38 college-cohort groups to retain the hypothesized heterogeneity.

i) *Analytic model: Functional form of interaction*. We used linear functional form for the interaction (both latent belonging and historic achievement moderators were continuous) and BCF to select cut-points for simple effects testing for the belonging moderator.

j) *Analytic model: Selection of student covariates*. We excluded self-perceived socioeconomic status as a student-level covariate because it was assessed post-manipulation and thus, in theory, could have been affected by random assignment (which we failed to appreciate at the time of pre-registration). PR1 also included other measures of academic preparation as covariates (high school GPA, high school class rank)

but these were not used, as they were available for less than 25% of the sample. When we conducted analyses using the prior GPA data that we did have, the pattern of results was the same.

k) *Analytic model: Missing values on potential covariates.* We used a single structural equation model approach to impute missing values of gender and standardized test scores.

l) *Analytic model: Handling small groups*. We tried various methods of collapsing small local-identity groups but ultimately decided to preserve the integrity of each local-identity group. To increase the stability of estimates of the historic group achievement moderator for small groups, we used more years of historic data and collapsed across first-generation status where necessary (see Table S7).

m) *Simple effects testing.* We used random forest models to determine cut-points for the belonging affordance moderator; this approach became available after our pre-registration. We used 25th and 75th percentile cut-points for the historic group achievement moderator.

Procedure

Recruitment. Partner institutions were responsible for the recruitment of incoming students, with CTC providing recruitment text in line with the intended representation of the exercise to students. Every incoming first-year (and transfer) student at each partner institution was eligible to participate. As part of orientation communications, schools either emailed students invitations to participate; included a link to the activity on a checklist of tasks to be completed before students arrived on campus (e.g., pay fees, upload photo for ID card, specify health insurance, etc.); and/or called students. Thus, the activity was represented as coming from students' own institution. Students were not compensated for their participation in this portion of the project.

Students participated by clicking on a link to the "What is it like to come to [school name]?" activity, which was described as "a study of students' experiences coming to [school name]" and presented with a consent form, which described the activity in general terms: "You are invited to participate by answering questions about your perspectives about college. You will also have the chance to read about the experiences of past and current students."

Randomization. Participants were randomized within school, at the student level, stratified on race-ethnicity, gender, and first-generational status (where available) from institutional records to ensure balance on these factors.

Experimental Materials

Social-belonging treatment condition. The social-belonging treatment uses stories from older students and active reflection exercises to articulate common worries about belonging in the transition to college; to represent these as normal for all students, not a sign of nonbelonging, and as passing with time; and to describe active steps students can take to build their belonging on campus. The content of both the treatment and the control condition drew heavily on past materials (*8*, *10*, *12*). It varied by partner school only to include the school name and logo, to revise specific details that did not fit certain types of schools (e.g., removing references to "teaching assistants" at colleges without teaching assistants), and to attribute stories to a set of students whose demographic background broadly matched the school's population. For an in-depth description of the materials upon which those tested here were based, including the full-text and theoretical considerations that shaped their operationalization, see (*8*).

Introduction. Following the consent page, students in the belonging condition were presented with a "Current Students Survey: A Summary of Results." This showed that, even as students broadly "reported a positive experience in college meeting other students, taking classes, and pursuing new opportunities" they also experienced common challenges: "in general, students from different backgrounds (e.g., gender, year in school, race, social class) reported many similar challenges and experiences." For example, it indicated that "almost all" students "worried at first in college about whether they fit in and belonged…when they started college," including worrying "about whether other students would include them and take them seriously in classes and coursework," and "that other students might view their abilities negatively." Yet "with time, students came to feel that they belonged in college," including "feeling comfortable working with other students and interacting with professors," and "Feeling confident that

professors and other students viewed their abilities positively." A summary statement noted that most students worry at first about whether they belong at [school name] but in time, they overcome these concerns and come to feel at home.

Notably, these materials do not deny that students can also have different experiences in college as a function of group identity nor that students may experience biases as a function of race-ethnicity, social-class, or other factors. Rather, they simply emphasize the kinds of challenges and concerns relating to belonging that are common to students from all backgrounds and represent these as common and temporary.

Stories from upper-year students. Students then read nine stories said to "illustrate the major findings of the Current Students Survey" and to be "representative of the responses of participating students" (*M*words per story=155, SD=28.72). Students were told the stories had been "edited for clarity." Each story was attributed to an upper-year (sophomore, junior, or senior) student at the college and included the student's gender and race.

Each story illustrated a common challenge to belonging students experience in the transition to college, how the student questioned their belonging as a consequence of this challenge, and how this experience and their feelings of belonging improved with time. The first story was always attributed to a student of the same gender and race-ethnicity as the participant. It read:

> *When I got into college, I was so excited about becoming a student at such a great school. But sometimes I also worried I might be different from other students. And when I got to campus, sometimes it felt like everyone else was right at home, but I wasn't sure if I fit in. At some point, I realized that almost everyone comes to college unsure whether they fit in or not. It's ironic—everybody comes to college and feels they are different from*

everybody else when, really, in at least some ways we are all pretty similar. Since I realized that, my experience at college has been almost one-hundred percent positive.

- [matched to participating student gender/race-ethnicity]

Other stories included:

I love college and I wouldn't trade my experiences here for anything. I've met some close friends, I've had some fantastic experiences, and I've certainly learned a lot. Still, the transition to college can be difficult, and it was for me. My freshman year sometimes I didn't know what I was doing—I made a lot of casual friends at parties and other places but I avoided interacting with professors in class or going to office hours. I think I was intimidated by them. I also got some low grades early on, which stressed me out. But these things all got better over time. I began to make good friends through classes. And my grades got better as I started working in study groups and asking for help from professors. I even got involved in research with a professor. Now I am happier than I have ever been at college. It is really rewarding for me to feel like I belong in the intellectual community here.

- Junior, White female

Initially my transition to college was pretty easy. Hanging out with friends in my dorm was fun, and I met a lot of people early on. After Winter Break, things got harder because it felt like all my really good friends were at home and I didn't have friends like that at school. However, I decided to just give it time and let things fall into place. I got involved in extracurriculars, and I met people who had common interests and unique perspectives. I also got to know people in class as study partners who became close friends. I found a comfort zone by exploring my interests and taking the leap into an active life on campus.

But this took time and before I found my niche here there were times when I felt quite lonely.

- Senior, Hispanic female

When I think back to the summer before freshman year, I was incredibly excited about coming to college but I was also somewhat intimidated. Walking into classes for the first time freshman year was uncomfortable, especially small classes. I worried about whether I could hold my own with other students (some of whom were upperclassmen) let alone professors. In the beginning, sometimes class discussions felt over my head. But now I feel much more relaxed. I've realized it's not about holding your own. We all bring something to the discussion, a different perspective or new ideas. It can be easy to forget what you bring. And I saw that everybody here has a common goal—to share knowledge and to learn and grow to do cool things in the future. We are all a part of that. Now I feel much more confident participating in discussions, listening, and sharing my opinions.

- Senior, White female

Active written reflection. Finally, participants completed a "saying-is-believing" task (*8*). They were asked to reflect on the themes they had read about, and to describe how and why "worries [about fitting in and belonging]…are likely to be common when students first go to college," "why students typically feel more at home on campus with time," and "what students do to feel more at home, e.g., as they get to know friends and professors." Students were encouraged to "consider specific experiences you will have…during your first year like living in a residence hall, meeting new people, joining student groups, interacting with professors, and taking college classes." They were encouraged to "draw on your past experiences with other *transitions* (like starting high school or going to a summer program) and on the stories from the

older students you just read," which were reproduced below the essay box. Finally, students were told, "Your essay may be provided, anonymously, to incoming [school name] students in future years to help give them a better understanding of what coming to college is like. The more you can describe the challenges you anticipate facing in coming to college and how you can respond to these challenges over time, the more future students will benefit. Thank you for your time and effort."

When students completed this task, they clicked on a button that led them to the post-intervention survey.

Active control condition. The control condition included the same elements as the treatment, including an introduction, stories from upper-year students, and the active written reflection. The content also focused on growth in the transition to college but emphasized how students got used to the physical rather than social environment (*10*). For instance, one of the upper-year student stories read:

> *I'm from a big city, so [school name] was an adjustment for me. Where I'm from, there are lots of people everywhere, all the time. It is noisy most of the day (and night), and that obviously isn't true of [school name]. At first, I really noticed the difference, but I've come to appreciate the opportunity to get away from noise when I want to. I think it is good for me to go to school here because it is easier to concentrate on my work when there isn't the bustle of a big city right outside my front door.*
>
> *- Sophomore, African American female*

Critical Questions and Considerations Regarding the Belonging Intervention

Here we address three questions that can come up regarding the belonging intervention.

Question 1: What is the theoretical background of the belonging intervention, how does it address group identity, and how does it relate to other social-psychological strategies to support student success? The social-belonging intervention is one of a number of strategies that aim to create a better social-psychological (e.g., "identity safe") environment in school for students who have historically been underserved (*17*, *45*, *54*, *55*). We discuss some of these strategies below. Together with the social-belonging intervention, these strategies are diverse unto themselves and fundamentally complementary ("yes and" not "either/or") ways to help students succeed.

Within this context, the social-belonging intervention comes from a specific theoretical tradition (*8*, *19*). It brings together classic attribution theory in psychology (*56*), particularly its application in attributional-retraining interventions (*57*), with research on social-identity threat (*18*, *58*). Attributional-retraining interventions are premised on the idea that students risk making stable, internal attributions for the causes of common struggles in the transition to college, such as attributing poor grades to a lack of ability, which can undermine their motivation and achievement. These interventions then offer students unstable, external attributions for these struggles, such as the idea that struggles can come about because students are not yet used to new living environments or ways of learning in college. This can raise GPA and college persistence.

Social-identity threats, including threats to students' racial-ethnic, social-class, gender, and other identities, can further inform how students make sense of everyday challenges in school (*58*). Such threats arise from the history and reality of racism and racist exclusion in education including the presence of negative stereotypes and underrepresentation. For instance, as classic research on stereotype threat shows, when a student belongs to a group that faces a negative stereotype in a specific area (e.g., a woman in math, an African American student in

intellectual reasoning), and that student takes a test said to evaluate that ability, the student can worry that a poor performance on the test could confirm that stereotype in the minds of others. Then that test-taking situation can pose a special threat to this student, a threat that does not arise for another student to whom the stereotype does not apply (*59*). Likewise, when one's group has been excluded or devalued in college, everyday adversities, such as feeling homesick, being excluded by peers, or a brusque interaction with a professor can pose a special threat to a student in that group: these experiences can seem to imply that "people like me" don't belong here, a stable, internal attribution (*19*). History and context racialize these adversities. They imply a fixed cause, rooted in group identity, for the event: "People like me don't belong here."

To contend with this circumstance, the social-belonging intervention offers students an alternative, non-threatening attribution for everyday adversities (i.e., an unstable, external one): the idea that these experiences are common, experienced by nearly all students at one time or another, and can improve with time. If so, such experiences need not portend a global or fixed lack of belonging and there are things students can do to build their belonging on campus.

To convey this idea, the belonging intervention shares nine stories from students in diverse identity groups describing different common challenges to belonging in the transition to college and how students' experiences improved with time. These stories convey a common truth—that everyday worries about belonging are normal in the transition to college and improve with time—but with variation. Challenges to belonging are represented as experienced by students from all social groups in one way or at one time or another, even as every individual student's experience is unique.

After reading these stories, students complete the saying-is-believing task in which they write advice for future students about their own experiences of belonging in the transition to

college and how worries about belonging and social adversities are normal but improve with time. This task serves several functions: it makes the experience active not passive, enhancing learning; it helps students translate the abstract ideas and the experiences of other students into their own life and circumstance, providing an opportunity to personalize and take ownership of the key message; and it positions students as benefactors of others not beneficiaries, preventing stigma associated with the receipt of help. The task is designed so students are free to call on any aspect of their personal or group identity that might be relevant to their experience, including both their experiences of non-belonging and how their experience may change with time. Students typically bring up aspects of their backgrounds in their essays, and often reference group identities.

An important point, then, is that even as the social-belonging intervention does not focus on group-based differences in experiences, it does not deny these differences, such as that students can experience racial bias, stereotyping, and discrimination or the pride students may feel in their racial group. It is *both true* that students have similar challenges and experiences (e.g., worries about belonging) in the transition to college *and* that students in different identity groups experience distinct challenges. In emphasizing the first truth, the belonging intervention does not deny the second. What the belonging intervention addresses is everyday adversities faced by students from all backgrounds, which can take on a racialized or social-class laden meaning in higher education contexts. Understanding this psychological process has deep roots in the psychological literature. Addressing it is one approach to improving student success, one that contends with a psychological consequence and cause of inequality.

This approach also reflects the finding that racial minorities often experience a greater sense of belonging and prefer contexts that promote dual identity representations, including both

a common superordinate identity (e.g., friends, students) and their distinct racial identities, versus a superordinate identity alone (*60*, *61*). Recognizing that everyone is likely to struggle and normalizing this and making it something we all go through and can overcome can help create an identity as someone who experiences these challenges (like everyone else) and who figures out a way through (using diverse strategies and in one's own time). Integrating standard materials with the saying-is-believing essay gives students space to address their own experience of belonging, given their personal and group identity, their goals, and the opportunities available to them.

Even as it is important to forestall fixed, global attributions for everyday adversities in college (e.g., "People like me don't belong here."), which can become self-confirming, many processes contribute to identity-threatening experiences in school. Thus, other approaches are also valuable. These include approaches that surface the racialized and social-class-informed experience of higher education directly, such as (a) discussing differences in students' experience along group (e.g., social-class) lines and how these are normal and not a barrier to belonging, called difference-education interventions (*44*); (b) representing college as a complex cultural space that does not require a student to fit a narrow ideal to belong and succeed, called cultural-fit interventions (*12* Experiment 3); (c) teaching students about stereotype threat (*62*) so as to forestall the inference that threat-related anxiety is a harbinger of failure (*63*); and (d) reframing stigmatized identities in strong and positive terms among both students and instructors (*42*, *64–66*). Broader efforts may also (e) support positive racial-ethnic identity development (*23*), such as through support for student affinity and cultural groups (*21*), and race- and ethnic-studies courses (*67*). Still other approaches represent intelligence as malleable not fixed, and center this way of thinking in the structure and organization of coursework (*12*, *16*, *29*, *43*). As

noted, these approaches are diverse, important unto themselves, and fundamentally complementary ways to support student success.

Question 2: Shouldn't the belonging intervention work better in "threatening" environments? Doesn't past research find greater effects in settings in which a student's group is underrepresented? Indeed, past research finds greater treatment effects in contexts that evoke threat. However, as will be seen, the present research is fully consistent with this past research, even as it advances theory in specific ways.

The most relevant past test of contextual heterogeneity in the belonging intervention was conducted by Walton, Logel and colleagues among women as they entered twelve undergraduate engineering majors (*15*). This trial yielded benefits (e.g., higher first-year GPA) for women enrolled in the six most male-dominated majors, eliminating a gender inequality. By contrast, women enrolled in the six most gender-diverse majors performed just as well as men even in the control condition and did not benefit from the treatment.

Critically, this past trial did not consider the role of belonging affordances. Here, in articulating the "seed-and-soil" model, we emphasize the importance of a context that is supportive of (affords) the way of thinking proffered by the intervention (*39*). An alternative metaphor is that of "water-on-parched soil." This metaphor implies that a proffered way of thinking could be an asset that compensates for something lacking in the environment (for discussion, see *36*). This would imply that larger effects should arise in contexts where belonging affordances are low. This is a legitimate hypothesis. A contribution of the present paper is that it clears up which of the two hypotheses is supported by the data, at least in this case—the "seed and soil" model. It is important that the National Study of Learning Mindsets—which (a) was launched at the same time as the CTC Belonging Trial, (b) was also a massive replication of an

existing psychological intervention (growth mindset), and (c) also sought to understand contextual heterogeneity through this scale-up—is coming to the same conclusion (*35*, *36*)

At a theoretical level, we believe the seed-and-soil model works better because it distinguishes vulnerabilities to psychological threat (e.g., worries about belonging) and opportunities to overcome these worries (e.g., to come to belong). Regarding the engineering trial, we would now say that women in the male-dominated majors were *vulnerable* to worries about belonging but had the *opportunity* to belong (see also *14*). While this theorizing was not tested in that trial, it is consistent with the finding that a strong majority of local-identity groups in the present trial had minimally adequate belonging affordances (85%).

Notably, even as the only contextual variable tested in the engineering trial was a lack of representation, that paper concludes by foreshadowing the importance of *both* vulnerability to worries about belonging…

> "Only women in male-dominated majors had worse outcomes than men, and only they benefited from the interventions…If the process that an intervention targets does not serve as a barrier to achievement for a given group or in a given setting, the intervention will not affect behavior." (p. 483)

…and what we now call affordances, or opportunities to belong:

> "Contexts may also differ in the extent to which they propagate the benefits of psychological interventions or undermine their effects. For instance, if the effectiveness of the social-belonging intervention depends on the potential for students to become more integrated in a school setting, long-term effects may depend on the willingness of peers and instructors to develop positive working relationships with target students." (p. 483)

Thus, even as the critical distinction between vulnerabilities and opportunities was implied in past research, past work (a) did not test this idea and (b) did not fully articulate the critical distinction.

We see the development of more nuanced ways to understand the psychological dimensions of school contexts as a significant step for the field. Much past research has focused narrowly on underrepresentation and other factors that can trigger belonging concerns (*14*, *15*, *27*, *68*, *69*). Distinguishing circumstances that create vulnerabilities to psychological threat from opportunities (affordances) to overcome these worries is more theoretically specific and useful, for instance in predicting where psychological interventions will have persistent effects and where they will fade out (*39*, *53*)

Question 3: Hasn't the belonging intervention been replicated before? Don't we know it is more beneficial for students who are disadvantaged? What are the contributions of the present trial? Yes, the belonging intervention has been replicated many times, in both post-secondary and secondary school settings (*8*). One secondary contribution of the present trial is as a replication of its effectiveness in supporting student achievement and reducing inequality in the transition to college. It is a particularly significant replication, as the sample of both schools and students is exponentially larger than has been tested previously. Further, we quantify the generalizability sample, which has not been done before, as discussed below.

However, we see the primary contribution in terms of understanding contextual heterogeneity. This includes the development of both theory and methods to understand heterogeneity among diverse identity groups across diverse contexts and drawing relevant implications about heterogeneous effects of the belonging intervention.

First, a limitation of past research on the belonging intervention is its emphasis on effects among students with canonical group identities (e.g., African American students in predominately White institutions; women in engineering; first-generation college students) or among "disadvantaged" students broadly. Yet "disadvantage" is not part of the definition of these or any other identity group. Here, by developing the local-identity group methodology, we *develop and empirically test theory* about what makes identity groups vulnerable to belonging concerns in the transition to college. This allows us to predict who may benefit from this treatment and who may not across the diversity of post-secondary contexts.

Second, we identify a critical boundary condition on the benefits of the belonging intervention: belonging affordances. When we began this project, it was not obvious that benefits would be greatest in supportive contexts, rather than in more hostile environments (as implied by the "water on parched soil" metaphor). Demonstrating the boundary condition posed by belonging affordances is a critical contribution, for two reasons: (i) it extends conceptually similar findings for boundary conditions around the growth mindset intervention, building broader theory about the intersection of wise psychological interventions and school contexts, and (ii) it highlights the need for colleges and universities to support strong belonging affordances for all of the students they serve. This final point is critical for application and theory. It shifts the focus from students to institutions, highlighting collective responsibilities. And it points the field toward research to better understand belonging affordances.

Third, combining the generalizability and heterogeneity analyses, the present study (a) shows that benefits of a brief, scaleable intervention in the transition to college generalize to more 749 colleges and universities in the United States, which annually welcome more than 1,000,000 students to college, and (b) empirically specifies boundary conditions on these

benefits. This has never been done before for the social-belonging intervention or any other psychological intervention in the post-secondary context.

Fourth, this study develops and validates a new method, local-identity groups, by which quantitative social scientists can study the *variation* in identity-group experiences across contexts. As we move to increasingly large-scale studies and diverse contexts this is essential, for we cannot assume that a given identity group has the same meaning or the same opportunities in different contexts. We've always known identity groups are not fixed in meaning. Local-identity groups allow us to relax this assumption and, correspondingly, to develop theory about how and why identity-group experiences vary in ways that, for instance, create vulnerabilities to belonging concerns and belonging affordances. Notably, this method goes beyond work on the contextual heterogeneity in the growth-mindset intervention, which has examined vulnerability at the individual level (e.g., individual students' level of past performance) and affordances at the context level (e.g., school and classroom growth-mindset cultures) (*35*, *36*).

Why have past trials been limited in their ability to understand contextual heterogeneity in the belonging intervention? They have been constrained by sample, theory, and methods: (i) Past studies have been conducted on an institution-by-institution basis, and thus lack the sample, especially of school contexts, needed to explore contextual heterogeneity adequately (notwithstanding initial small-scale efforts, e.g., the aforementioned engineering trial, which included 228 students in 12 undergraduate majors (*15*)); (ii) past studies have not been conducted with an adequate theory of contextual variability and, therefore, (iii) have not adequately assessed vulnerability to worries about belonging on the one hand and opportunities to belong (affordances) on the other. Finally, (iv) past trials have also not implemented local-identity

groups in analyses, which account for variability in students' experiences across contexts even when they share the same racial-ethnic or social-class group.

These contributions go beyond the social-belonging intervention per se. Hardly any well-powered studies have tested for heterogeneity in the persistence of treatment effects, despite the importance of this question for the social and behavioral sciences in general (*53*). In contributing to high-level theory about this question, the present results provide evidence for the importance of what Bailey and colleagues call "sustaining environments" and we have called "affordances" in giving rise to persistence rather than fade out. In this respect it is important that, even as the results converge with research on contextual heterogeneity in the growth-mindset intervention (*35*, *36*), there are also differences between the two test cases: (a) these are different interventions, which address different belief systems, namely about the opportunity to come to belong vs. the potential for intelligence to grow; (b) contextual tests of growth-mindset interventions have focused on the transition to secondary school, whereas we focus on the transition to college; and (c) as noted, the conceptualization and measurement of both vulnerabilities and affordances differ, as we focus on identity-group level measures of vulnerability and opportunities to belong appropriate to the group-based nature of belonging concerns, whereas growth-mindset trials have focused on individual measures of vulnerability (personal past poor performance) and classroom- and school-level measures of affordances (e.g., teachers' endorsement of a growth-mindset, peers' level of challenge-seeking) (*35*, *36*). This variability illustrates the robustness of the seed-and-soil model for understanding contextual heterogeneity in whether treatment persist over time, and how this model can be used flexibly to understand persistence and fadeout in different problem spaces.

Key Definitions

Race-ethnicity. There is no simple or one best way to categorize race-ethnicity, as racial-ethnic identity is a social construction and, thus, varies over time and contexts. Therefore, our approach relied on established categories and procedures created the United States Census, and we committed to this in our pre-registration. Our measure prioritized self-reported race-ethnicity and then used school-reported race-ethnicity if it was available and self-reported race-ethnicity was missing.

Following the Census, we first defined students of Hispanic or Latinx ethnic origin, of any race. These were students who (were) identified as:

1. Of Hispanic or Latinx origin (i.e., selected one or more of these sub-categories: "Mexican American/Chicano," "Puerto Rican," "Central American," or "Other Hispanic)

Second, we classified students not of Hispanic/Latinx origin into five racial groups:

1. Black/African/African Americans ("African American/Black," "African," "Caribbean," "Other Black")
2. White/European Americans ("European/European American," "Middle Eastern/Middle Eastern American," "Other White")
3. Asian/Asian Americans ("East Asian," Southeast Asian," "South Asian," "Other Asian")
4. Native American/Native Hawaiian/Other Pacific Islander ("American Indian or Alaska Native," "Native Hawaiian or Other Pacific Islander")
5. Other ("Other")

This approach departs from Census procedures in two ways. First, we combined the Census categories "American Indian or Alaska Native" with "Native Hawaiian or Other Pacific Islander"

given the small size of these groups in our sample. Second, we did not include a category "Two or more races, non-Hispanic." Instead, while recognizing the complexity of the identities and experiences of bi- and multiracial people (*70*), we assigned people who identified with more than one racial group to a racial category in the order listed above, such that membership in a later category means that a student did not identify with an earlier category. In doing so, we assumed that this categorization would be more meaningful in predicting students' experience and outcomes better than a single bi- or multi-racial category, which could include any racial groups. The experiences of biracial Black/White students, for instance, are likely to be very different from the experiences of Asian/White students. It also reflects our judgment that creating a separate category for every bi- and multiracial combination would give rise to many small local-identity groups, which would increase error in estimates of the local-identity group moderator variables.

In creating this order, we prioritized Black/African/African American given the racialized history and segregation of bi- and multiracial people partially of African descent, particularly the practice of hypodescent (i.e., "the one-drop rule") and thus how such people are often seen and treated (*71*, *72*). It is also consistent with African Americans' own preferred ways of classifying African Americans (*73*).

We do not believe there is one best way to order the White, Asian, and Native American categories. While prioritizing the White category, as our primary approach does, may deny ways in which biracial students identify and are treated as a function of their racial-minority identity, other concerns arise if the order is inverted. For instance, swapping the positions of the White and Native American categories may dilute Asian and Native American monoracial groups and neglect the way in which biracial White/Asian and White/Native American students can be

treated as White. It treats these biracial students as Asian and as Native American, even if they appear and identify primarily as White. These concerns are not theoretical: Research finds many biracial White/Native American people report a greater connection with White than with Native Americans (*74*).

Given these complexities, and given that we pre-registered the order indicated above, we followed this order in primary analyses. However, we also conducted a robustness test using an alternative order, swapping the positions of the White and Native American groups. Doing so allows us to determine, as an empirical matter, whether this decision affects our primary results. In short, it does not, as discussed next.

Descriptive results with the alternative classification order. Because students' racial-ethnic identity is not used directly in our analyses but, rather, contributes to the construction of local-identity groups, changing the racial classification order affects not only which local-identity group some bi- and multiracial students are assigned to but, also, which local-identity groups have at least one participant in each condition and therefore are retained in the intent-to-treat sample. This shifts the sample of both students and local-identity groups.

With the alternative classification order, the number of local-identity groups in the sample increases from 374 to 400. Unsurprisingly, the additional local-identity groups are Native American/continuing-generation students (11 college cohorts), Native American/first-generation students (9), Asian/continuing-generation students (5), and Asian/first-generation students (2). One local identity-group is dropped with the alternative classification order (one college cohort of Asian continuing-generation students). Compare with Table S6.

Turning to the student sample, 97% of students in the original intent-to-treat sample are retained with the same racial-ethnic classification with the alternative classification order; 2.5%

are retained with a different classification. Just 0.001% (29 students) are dropped, and 35 students are added. Thus, the intent-to-treat sample size rises from 26,911 to 26,917. The number of students classified as Native American increases from 127 to 462. The number classified as Asian/Asian American increases from 2,565 to 2,918. The number classified as White drops from 13,833 to 13,151. Compare with Table 1.

Robustness of inferential statistics with the alternative classification order. Given the modest changes to the local-identity group and student sample, unsurprisingly the primary analyses were robust to the alternative classification order. The 3-way interaction remained significant (b=0.012 [.003, .022], se=0.005, t=2.69, p=0.007), as did the historic achievement × condition interaction among local-identity groups medium-to-high in belonging affordances (b=0.012 [.0006, .0023], se=0.006, t=2.07, p=0.038).

The simple effect of the belonging treatment also remained significant:

1. among students in local-identity groups medium-to-high in belonging affordances (b=0.012 [.001, .023], se=0.006, t=2.16, p=0.031),
2. among students in local-identity groups medium-to-high in belonging affordances and low in historic achievement (*b*=0.022 [.004, .039], *se*=0.009, *t*=2.38, *p*=0.017), and
3. among students in local-identity groups medium-to-high in belonging affordances and moderate in historic achievement (*b*=0.013 [.002, .024]. *se*=0.006, *t*=2.31, *p*=0.021).

If anything, these contrasts and effect estimates are slightly stronger than those produced using the original classification order (e.g., cf. Table S10).

Finally, the treatment effect also remained non-significant with the alternative order:

4. among students in local-identity groups medium-to-high in belonging affordances and high in historic achievement (b=-0.003, *p*=0.687) and

5. among students in local-identity groups low in belonging affordances (*p*s>0.128).

First-generation status. We defined first-generation status as having no parents or guardians with a Bachelor's degree or higher. As with race-ethnicity, we prioritized self-reported status, and then used school-reported status if it was available and self-reported status was missing. If first-generation status was unreported/unknown: (a) Students who indicated "don't know" or "doesn't apply" on the relevant self-report items (which assessed parent educational attainment) were coded as "first-generation"; (b) students missing both self-report and school first-generation status and who did not self-report "don't know" or "doesn't apply" were assigned the modal first-generation status for their race-ethnicity × college × cohort group if this was available.

Local-identity groups. Once the race-ethnicity and first generation-status of each student had been coded, students could be classified into local-identity groups (i.e., race-ethnicity × first-generation status × college × cohort groups). Though 456 total groups were possible, not all college cohorts had all race-ethnicity × first-generation-status combinations. Students in the minimally processed original sample (i.e., retaining those who started the survey but excluding those at the Canadian university, *N*=45,457; see Fig. 2) belonged to one of 435 groups. Students in the final intent-to-treat sample (*N*=26,911) belonged to one of 374 groups.

Intervention condition. The independent variable was defined as *1* if students were randomly assigned to the standard belonging-treatment condition and *0* if students were randomly assigned to the control condition and they saw the first page of randomized content for their condition. Students assigned to the customized treatment condition were excluded (defined as missing) for this variable.

Person-level covariates: Gender and standardized test score. We controlled for two person-level covariates. Gender was assessed by inviting students to self-identify as "a man," "a woman," "I prefer another term" (with a text box), or "Transgender" (2016-2017 cohort only). Because the "Transgender" label was only used in one cohort, we collapse this category with "I prefer another term" in Table 1. We note that, in retrospect, this way of assessing gender identity is not optimal and we would not use it again. It forced people to choose between identifying as male, as female, or as transgender, when a transgender man or a transgender woman might identify with two of these categories. In analyses, we treated gender as a dichotomous variable equal to *1* if the student was female and *0* if the student was not female. Thus, the small number of students (n=414, or 1.5% of the intent-to-treat sample, Table 1) who did not explicitly identify as male or female were coded as zero.

For standardized test scores, we used ACT scores obtained from institutional records (observed range=11-36, M=26.06, SD=5.83, available from N=23,261 [86%] of the analytic sample). When students had only SAT scores, these were converted to an equivalent ACT score. Each variable was mean-centered within local-identity group before analysis.

When ACT or gender were missing (14% and 0.01% of the sample, respectively), a SEM-estimated value was used. Each estimation is a prediction of the covariate based on values of the other covariate (i.e., gender or ACT) as well as race-ethnicity, first-generation status, and a self-reported social-class item ("How would you describe your family's social class?"; 1= working class, 2=lower middle class, 3=middle class, 4=upper middle class, 5=upper class). Race-ethnicity and first-generation status were controlled in the model via the dummy for each racial-ethnic and first-generation group.

High school grades were not used, as they were available for less than 25% of the sample.

Year 1 full-time completion. The primary outcome was whether students completed the first year of college full-time enrolled, obtained from school-provided administrative records. This was defined as *1* if students earned the minimum credits in the first-year required by their university to complete fall and spring semesters (or fall, winter, and spring quarters) at full-time status, and *0* otherwise.

Measures

Post-intervention survey (manipulation check). Upon completing the reading-and-writing task, students in both conditions answered brief survey questions, followed by demographic questions.

The primary measure for the purpose of the present report was a manipulation check. This measure was included to determine whether the intervention achieved its intended initial impact. This assessed *anticipated growth in belonging*. Students answered three items regarding how much they expected to feel they belonged at their institution when they arrived in the fall of their first year ("How much do you think you will feel you belong/fit in/feel at home at [school name] when you arrive on campus this fall?"; α=0.88; *M*=4.35, *SD*=1.25, N=24,585) and three parallel items regarding how much they expected to feel they belonged at the end of their sophomore year (e.g., "At the end of your sophomore year, how much do you think you will feel you belong/fit in/feel at home at [school name]?"; α=0.94; *M*=5.77, *SD*=1.06, N=24,453). All six items were on a 1*(not at all)* to 7*(an extreme amount)* scale. The three items for each timepoint were averaged to form a composite of anticipated first-year fall belonging and anticipated sophomore spring belonging, respectively. Anticipated growth in belonging was calculated as the difference between these two composite scores (possible range=-6 to 6, *M*=1.43, *SD*=1.07, N=24,449). Higher scores indicate an expectation of greater growth in belonging over time,

consistent with the message of the social-belonging intervention. This growth could be achieved through lower initial anticipated belonging (consistent with the treatment message, that belonging worries are normal at first in college) and/or higher later anticipated belonging (consistent with the treatment message, that students typically overcome these worries with time).

Spring follow-up survey. In the spring-term, participating students in each cohort at each school were invited to complete a survey on their first-year experiences. At most schools, all students were invited to participate in the spring survey. At schools with larger numbers of students, a random subsample was invited to participate, oversampling underrepresented racial-ethnic minority students to obtain adequate sample sizes for all groups. The survey included a range of measures drawing on past research (*12* Experiment 3). For the purpose of the present report, the primary measure assessed belonging. Four items formed the group belonging composite. Three positively valenced items pertained to social belonging ("I feel like I belong at [school name]," "I fit in well at [school name]," and "I feel comfortable at [school name]") (*1=strongly disagree*, *6=strongly agree*) and one negatively valenced item assessed belonging uncertainty (reverse-scored, "When you think about [school name], how often, if ever, do you wonder: 'Maybe I don't belong here?'") (*1=never*, *5=always*). Thus, the composite incorporated both level of belonging and uncertainty about belonging (*19*). All items were converted to a 10-point scale before averaging (α=0.88; *M*=7.30, *SD*=1.89, N=7,209).

Fidelity of Intervention Delivery

Completion rate of randomized content. Of the 30,701 students who opened the study materials (*Ncontrol*=15,372, *Ntreatment*=15,329), 89% of students in each condition saw randomized content (i.e., clicked to the first page with randomized content) (*Ncontrol*=13,705,

Ntreatment=13,600). Of these, 91% in the control condition and 90% in the treatment condition completed the randomized content, defined as having at least clicked through these pages (Ncontrol=12,492, *Ntreatment*=12,303).

For the percentage of students who wrote essays by condition, see Table S1. For time spent on randomized content, see Table S2. For comparison with earlier trials, see Table S3, illustrating the tradeoff between scaling and engagement.

Sample

Intent-to-treat (ITT) sample. Tables 1, S4, S5, and S6 describe the student and institutional samples, and the distribution of local-identity groups across the two moderator variables, as a function of first-generation status and racial-ethnic identity. For the study CONSORT diagram, see Fig. 2.

Treatment-on-treated (TOT) sample. In addition to meeting the criteria specified for the ITT sample, the TOT sample included only students who wrote an essay (of any length or content). See Fig. 2.

Interpretation and Calculation of the Group Historic Achievement Moderator

As noted in the main text, one moderator was the local-identity group's historic level of achievement along the primary outcome: first-year full-time completion rates, in pre-experimental cohorts. We interpret lower levels of historic achievement as a vulnerability to the possibility that belonging concerns could undermine students' persistence in college. This interpretation brings together several aspects of theorizing. First, awareness of low historic persistence in one's identity group may have a causal effect in raising belonging concerns: Not seeing members of one's group persist on campus may imply to students that "people like me" might not be able to belong and succeed there (*75*). Indeed, past research finds that the prospect

of poor achievement by a member of one's racial group can evoke psychological threat among students of color in laboratory settings (*76*). Further, in field settings, perceiving that others believe that one's group is likely to perform poorly predicts lower levels of belonging (*77*). Second, low historic persistence at the group level may function as a proxy: It may reflect the presence of other factors in the college context that undermine belonging and persistence for members of a given group. Third, by using a measure of academic persistence, we focus on a vulnerability to a lack of persistence. If students in a given local-identity group were vulnerable to feelings of nonbelonging but nonetheless persisted (e.g., due to strong institutional channels for persistence) (*78*), we would not expect the belonging intervention (or any other intervention) to increase persistence. Conversely, in consequence of the many sources of inequality that accrue into and within higher education, students from poorer performing groups have greater room for improvement.

Consistent with our focus on a measure of prior performance, past research finds that values-affirmation interventions, a different social-psychological intervention but one that also aims to mitigate psychological threat and support belonging among students who face negative stereotypes in school, had its greatest effect in middle schools that (a) had larger extant racialized disparities in achievement and (b) had smaller representations of racial-minority students (*27*). This past research, which examined heterogeneity across 11 middle schools, was not able to distinguish these factors (they correlated across these schools). Nonetheless, it is consistent with our theorizing that low levels of group achievement contribute to vulnerabilities to threat and nonbelonging, which may be remedied through intentional intervention. By contrast, students in groups that are performing well are not so vulnerable to threat, and would not be expected to benefit from interventions to mitigate this threat.

In calculating the local-identity group's historic level of achievement, we were sensitive to the fact that means for smaller race-ethnicity × first-generation status groups can be less stable. Therefore, we used more years of data (up to 4 years preceding the intervention) to create the final overall mean for smaller groups. For groups of 1 to 5 students, we addressed this issue further by using the race-ethnicity × college × cohort historical average rather than the race-ethnicity × first-generation status × college × cohort historical average. The race-ethnicity × college × cohort historical average was also used for one school, which did not have historic first-generation status available. See Table S7.

The group historic achievement moderator was standardized using the mean and standard deviation across local-identity groups.

Calculation and Validation of the Belonging Affordances Moderator

Calculation of the belonging affordances moderator. To assess *belonging affordances*, we examined responses among the subsample of control-condition students who completed the four belonging items on the spring-term survey. Among these students, we estimated a structural equation model for latent belonging, and extracted standardized factor scores for the estimated latent variable at the individual student-level. We then used a multilevel random effects model to obtain Empirical Bayes (EB) estimates of belonging for each local-identity group. When a sample includes small groups, whose raw group means vary a lot, obtaining EB estimates through multilevel modeling is recommended to shrink group means to an informative average (*79*). The model included a random effect for local-identity group and Level-2 fixed effects of race-ethnicity and first-generation status, so very small local-identity groups could "borrow" information from similar groups. Thus, the final belonging score for each local-identity group was the mean score for students from that group's same race-ethnicity and first-generation status

group across colleges and cohorts plus the deviation of the local-identity group's score from this mean. After this process the belonging score was still missing for some groups, in which case it had to be imputed.

As was the group historic achievement moderator, the belonging affordance moderator was standardized using the mean and standard deviation across local-identity groups.

Validation of the belonging affordances moderator. In Phase 1 of analyses, we validated the spring-term belonging moderator by assessing correlations between spring-term EB estimates of belonging among control-group students and factors known to relate to belonging. Belonging scores were predicted by the proportional representation of the local-identity group on campus (see Table S8), consistent with past theory and research (*15*, *69*). They further predicted measures of social and academic integration on campus assessed in the spring-term survey among control-group students, including greater mentor development, greater development of close friends on campus, less loneliness, and greater involvement in student groups (*12*).

Does local identity-group belonging affordance correlate strongly with historic local identity-group achievement, or do these reflect separate factors? They reflect separate factors. Adjusted EB estimates of local identity-group belonging in the spring of the first year did not correlate with the rate at which local-identity groups had completed the first year full-time in the control condition in the two to four years prior to the study years, r=0.062, p=0.233. Thus, the historic circumstance that, we theorize, can lead to vulnerability to worries about belonging was largely independent of the opportunity students had to belong in their college in their cohort.

Validation of the Local-Identity Group Approach

Was variation in belonging affordances fully explained by students' race-ethnicity and first-generation status or did it vary by context? It varied by context. There remained significant

variability in belonging affordances at the local-identity group level after controlling for race-ethnicity (entered into the model separately for Hispanic, Black, Asian, Native, & Other) and first-generation status (with White continuing-generation students as the referent group), *var*=0.088 [.046, .169], *se*=0.029, *t*=3.00, *p*=0.003. Thus, students with the same static group identity experienced different belonging affordances across contexts.

Benchmarks for Effect Sizes

As noted in the main text, it is important to calibrate any potential reform against other possible reforms, including in terms of the impact achieved and the cost required to achieve this impact. Past research estimates that a 1 standard-deviation improvement in teacher value-added scores in a single grade (in grades 3-8) raises the probability of college attendance at age 20 by 0.82 percentage points (*80*). In college, a $3,500 scholarship in the first-year increases rates of completing the first semester full-time by 0.3 to 1.1 percentage points and the second semester by 0.8 to 3.4 percentage points (*81*). Further, when promising interventions for adolescents and young adults have been scaled up, they have tended to show null effects on objective outcomes such as post-secondary enrollment (e.g., *82*, *83*).

Supplementary Text

Analytic Approach

As preregistered, we analyzed the data using a multilevel fixed-intercept random-coefficient model approach (*47*, *52*). This type of model specifies a fixed intercept for each site while allowing the treatment effect to vary randomly across sites. In the original formulation, sites were intended to be schools. We extended this approach by applying it with local-identity groups. That is, to capture heterogeneity in this study, our model was based on the assumption that treatment effects would vary not only by colleges and college-cohorts but across the 374 racial-ethnic × first-generation-status groups × college × cohort groups. However, a model including a fixed effect for each of the 374 sites did not converge. Thus, we used separate fixed effects for each of the 12 race-ethnicity × first-generation status groups and each of the 38 college-cohort groups to retain the hypothesized heterogeneity.

Level-1: Individual Level

$$Y_{ij} = \alpha_j + \beta_j\left(T_{ij}\right) + \sum_{k=1}^{K} \theta_k\left(X_{kij}\right) + \varepsilon_{ij}$$

Level-2: Local-Identity Group Level

$$\beta_j = \beta + \lambda_A\left(A_j\right) + \lambda_M\left(M_j\right) + r_j$$

with:

$$\varepsilon_{ij} \sim N\left(0, \sigma_T^2\right)$$

$$r_j \sim N\left(0, \tau_T^2\right)$$

Where:

- Y_{ij} is the outcome for student *i* in local-identity group *j*

- α_j is a local-identity group-specific intercept (for each of the 38 college-cohorts and for each of the 12 race-ethnicity × generation-status groups (as noted, there were not enough degrees of freedom to estimate a fixed intercept for each of the 374 local-identity groups.)
- T_{ij} is a student-level indicator of randomly assigned condition (*1* if standard condition, *0* if control)
- X_{kij} is the local-identity group-centered baseline covariate *k* for each student *i* from local-identity group *j* (covariates are gender and standardized test score)
- A_j is the Empirical Bayes estimate of group-level belonging affordances, coded continuously
- M_j is the group historic level achievement, coded continuously (primary analyses), or canonical disadvantaged status (*1* = disadvantaged status, *0* = advantaged status; robustness analyses)

Effectiveness of Random Assignment

The intervention and control groups did not differ in terms of pre-random-assignment characteristics (see Table S9).

Supplementary Heterogeneity Analyses

Heterogeneity in the manipulation check. As noted in the main text, the predicted main effect of condition on anticipated increases in belonging over time was significant, *b*=0.289 [95% CI: 0.261, 0.318], *se*=0.014, *t*=20.13, *p*<0.001, showing the intervention had its intended initial effect. There was heterogeneity in this effect across local-identity groups, *b*=0.052 [.023, .116], *se*=0.021, *t*=2.419, *p*=0.016, *Q*=358.80, *df*=49, *p*<0.001. However, the treatment effect did

not interact with local-identity groups' historic achievement level, *b*=-0.018 [-0.047, 0.011], *se*=0.015, *t*=-1.24, *p*=0.216, or belonging affordances, *b*=.016 [-.008, .040], *se*=0.012, *t*=1.32, *p*=0.187. All students were retained in both intent-to-treat and treatment-on-treated analyses regardless of manipulation check responses (see Fig. 2).

Bayesian multilevel analyses. To interpret and visualize the 3-way interaction, we estimated a flexible Bayesian multilevel model called Bayesian Causal Forest (BCF). This method became available after our pre-registration but has since won several open competitions for detecting true heterogeneity where it exists and not over-interpreting the data when it does not exist (*84*). We use BCF because it allowed us to detect the threshold at which the afforded belonging moderator made a difference, and allowed us to determine the functional form (e.g., linearity) of the achievement moderator. BCF uses machine-learning tools to detect complex and non-linear interaction effects, while applying conservative prior distributions and partial pooling to prevent small groups from distorting the overall pattern of results. Thus, BCF is flexible enough to find nonlinearity while being highly conservative.

BCF produces a posterior probability of a positive treatment effect for local-identity groups in each cell of the design. As can be seen in Table S12, Panel A, in the primary intent-to-treat sample (unweighted), students in local-identity groups medium-to-high in belonging affordances with low and with medium historic achievement show a strong probability of a positive treatment effect (>85%).

Robustness tests. We had several goals in conducting robustness tests. First, we wanted to know whether the effects were reliant on the use of local-identity groups' level of historic achievement, or if they would replicate using canonical disadvantaged status as in prior studies (*12*, *13*). Second, we tested whether the model using canonical disadvantage instead of group

historic achievement would have a higher standard error for the 3-way interaction term, consistent with our theorizing that treating social disadvantage as a static grouping variable masks contextual variability. Third, we tested whether the basic pattern of results—the 3-way interaction and the effects of condition among local-identity groups medium-to-high in belonging affordances—would persist (a) including generalizability weights and (b) in the treatment-on-treated (TOT) sample.

As will be seen, these latter robustness tests yield some negative contrasts for the effect of condition among students in local-identity groups low in belonging affordances and low in historic achievement (in the primary ITT analysis, this contrast is not significant, p=0.112; Table S10). While we are sensitive to the possibility of ironic effects, we are also attuned to the fact that with a large sample there can be false positive findings in complex multi-way interactions, on the "off diagonal" cases. This is a primary reason we fit the BCF model. It will "shrink to homogeneity" if there is a small, outlying group whose treatment effects, positive or negative, are purely due to noise. Note that this does not mean that BCF will cover up truly iatrogenic effects—if the intervention caused harm, it would find it.

In this regard, it is important that, when this conservative method was used with the primary intent-to-treat sample, we found a posterior probability of a positive treatment effect above 50% in all cells of the design (Table S12, Panel A). This indicates that the median treatment effect was positive in all cells. Further, when we used this method in robustness tests (i.e., including generalizability weights and examining the TOT sample, Table S12 Panels B and C), we found positive effects in most cells and null effects in others, and a minimal (>70%) posterior probability of a negative effect in none. For this reason, we do not interpret the negative effects, even as we report them for the sake of transparency and completeness. While iatrogenic

effects are possible and could emerge in another sample, we believe the patterns observed here are likely to be artifacts of (a) how standard regression analyses force the moderators to be linear and, thus, how positive effects for the predicted groups can force a negative relationship on the other side of the moderator, and (b) failure to regularize (i.e., shrink) in the linear model.

Using a canonical societal disadvantage classifier. Using the same model as the primary analyses but replacing the group historic achievement moderator with canonical societal disadvantage (see Table 1), we found the same three-way interaction, here between canonical disadvantaged status, belonging affordances, and condition, b=0.023 [.003, .042], se=0.010, t=2.32, p=0.020. As noted in the main text, consistent with theory these analyses had a higher standard error for the 3-way interaction term (canonical disadvantaged status: SE=0.010; historic group-level achievement SE=0.005).

Interestingly, the 2-way disadvantaged status × condition interaction was significant for groups low in belonging affordances, b=-0.060 [-.116, -.005], se=0.028, t=-2.12, p=0.034, but not for groups medium-to-high in belonging affordances, b=0.009 [-.013, .031], se=0.011, t=0.82, p=0.410. However, simple effects revealed that, as predicted, canonical disadvantaged groups medium-to-high in belonging affordances exhibited a marginally significant, positive treatment effect on Year 1 full-time completion, b=0.013, [-.002, .028], se=0.008, t=1.75, p=0.081. Simple effects for all other groups were smaller and non-significant, $|t|$s≤1.150, ps≥0.133.

Including generalizability weights. Including post-stratification generalizability weights to force the composition of our sample to resemble the 749-college population of inference (see "Development of the Generalizability Sample" below), the three-way interaction between group

historic achievement, belonging affordances, and condition was significant, *b*=0.015 [.005, .026], *se*=0.005, *t*=2.88, *p*=0.004.

Among students in local-identity groups medium-to-high in afforded belonging, we found an overall conditional average treatment effect (CATE) of 1.4 percentage points, b=0.014 [.002, .026], *se*=0.006, *t*=2.27, *p*=0.023, but a non-significant Condition × Historic Achievement interaction, *b*=0.008 [-.005, -.021], se=.007, *t*=1.18, *p*=0.240.

We conducted simple effects in these groups to test our specific hypotheses and mirror the analyses of the unweighted sample. Among students in local-identity groups medium-to-high in afforded belonging with low historic achievement, we observed a treatment effect of 1.7 percentage points, though it was marginally significant, *b*=0.017 [-.002, .036], *se*=.010, *t*=1.73, *p*=0.084. Among students in local-identity groups medium-to-high in afforded belonging with medium historic achievement, the treatment effect was 2.2 percentage points, *b*= 0.022, [.010, .035], *se*=0.006, *t*=3.63, *p*≤.001. Finally, among students in local-identity groups medium-to-high in afforded belonging with high historic achievement, the treatment effect was not significant, *p*=0.785. See Table S10, Panel B.

Among students whose local-identity groups was low in afforded belonging, there was no overall treatment effect, *b*=-.004, [-.024, .016], *se*=.010, *t*=-0.42, *p*=0.676. There was also a negative interaction with group historic achievement, *b*=-0.035 [-.053, -.017], *se*=.009, *t*=-3.78, *p*<0.001. The simple effects of treatment at medium and high levels of historic achievement were non-significant, *p*s>0.236. Among students in local-identity groups low in afforded belonging with low historic achievement, there was a negative treatment effect, *b*=-.041 [-.081, -.001], *se*=.020, *t*=-2.02, *p*=0.043.

TOT sample analyses. We conducted analyses on the TOT sample paralleling the primary analyses on the ITT sample. These analyses include only those participants who, in either condition, wrote a response (of any length or content) to the saying-is-believing essay prompt.

The three-way interaction between the group historic achievement level, belonging affordances, and condition was significant, b=0.016 [.006, .025], se=0.005, t=3.20, p=0.001.

As in the ITT sample, we conducted follow-up hypothesis tests among the majority of students and local-identity groups that were medium-to-high in afforded belonging. Within this large category, we found a marginal overall CATE of 1.0 percentage point, b=0.010 [-.0008, .021], se=0.006, t=1.82, p=0.068, qualified by a Condition × Historic Achievement interaction, b=0.011 [.00004, .023], se=.006, t=1.97, p=0.049.

Among students in local-identity groups medium-to-high in belonging affordances with lower historic achievement, we observed the largest treatment effect, 1.5 percentage points, though it was marginally significant, b=0.015 [-.003, .033], se=.009, t=1.61, p=0.108. Among students in local-identity groups medium-to-high in belonging affordances with medium historic achievement, the treatment effect was 1.2 percentage points, b=0.012 [.001, .024], se=.006, t=2.13, p=0.033. Finally, among students in local-identity groups medium-to-high in belonging affordances with high historic achievement, the treatment effect was not significant, p=0.732. See Table S10, Panel C.

Among students in local-identity groups low in belonging affordances, there was no overall treatment effect, b=-0.017 [-.043, .009], se=0.013, t=-1.30, p=0.193. Again, there was a negative interaction with historic achievement, b=-0.038 [-.062, -.014], se=0.012, t=-3.08, p=0.002. The simple effects of treatment among groups low in belonging affordances at medium and high levels of historic achievement were non-significant, ps>0.313. However, among

students in local-identity groups low in belonging affordances with low historic achievement, there was a negative treatment effect, *b*=-.037 [-.067, -.008], *se*=0.015, t=-2.47, p=0.014.

Development of the Generalizability Sample

The development of the generalizability sample followed the approach described by Tipton and colleagues (*85*, *86*)

Summary. The final population to which the 22 CTC sample colleges generalize is what we will refer to as subpopulation 4 (N=749 colleges); this is one of four subpopulations that reached the final stage of consideration. With this subpopulation, we were able to achieve a B index value ≥ 0.80 and covariate balance between the population and sample colleges for 15 key covariates constructed from data available from IPEDS. This was selected over subpopulation 3, which had a slightly higher B-index value and similar covariate balance but comprised fewer colleges (N=681) and required more maximum and minimum cutoffs to key institutional indicators. In addition, we found post-stratification (sub-classification) weights proved superior to inverse probability weights in achieving covariate balance between the population and sample colleges. We apply these post-stratification weights for subpopulation 4 to the final primary (condition × local-identity group belonging × historic group achievement) regression model to obtain estimates for relevant coefficients that apply to the population of 749 colleges.

Specification.

Defining the target population. We began with a database of colleges and institutional indicators developed from several raw data files available for download from the website of the Integrated Postsecondary Education Data System (IPEDS). The database we constructed from these raw files contained colleges that had non-missing data for at least one year from 2011 to 2017 (N=8,718 colleges). We initially formed a target population of 1331 colleges from this

population frame by removing kinds of colleges that had no representation in the CTC sample of colleges (N=22) according to specific dichotomous variables available from IPEDS during the study years (2015 and 2016). Excluded colleges included those:

1) whose value for control was "not available" or private for-profit (instead of public or private non-profit),

2) that had a level of 1-year (coded by IPEDS as less than 2 years) or 2-year (coded by IPEDS as at least two but less than 4 years) instead of 4-year (coded by IPEDS as 4 years or above),

3) had an institutional category that was not "degree-granting, primarily baccalaureate,"

4-6) had an explicit open-admissions policy or proportional admissions rate equal to 0% or 100%

7) had no endowment at the end of any fiscal year,

8) provided distance education only,

9) is/was a Historically Black College or University, and

10) does/did not provide any on-campus housing.

We applied each rule first to the relevant 2015 variable and then to the corresponding 2016 variable. See Table S13. The final target population comprised 1301 colleges. These colleges were retained (not excluded) based on the above criteria (Table S13) and had available data for 15 indicators from IPEDS that we used in the final propensity score model to predict selection into the CTC sample (Table S14).

Determining list of covariates for the propensity score model. The original database we developed from annual IPEDS raw data files contained institutional indicators from seven years, 2011 to 2017. As the study took place during the years of 2015 and 2016, we created a two-year

composite for each indicator for each institution for these years. If institutions were missing data for a given indicator during those years, we used a two-year average from 2013 to 2014 if it was available.

For the purpose of predicting selection into the CTC sample, we tested a subset of indicators that we thought might be most relevant, seeking to optimize the tradeoff between the number of variables and fit between the sample and the target population. Following these tests, we selected 12 covariates for the initial propensity score model (later a 15-covariate model will be used). We standardized each variable within the target population of 1301 colleges. For monetary and count variables, we first took the natural log of each variable (plus a constant) before standardizing them. See Table S14.

Initial computation of the B index. The generalizability index B represents how well a sample can be generalized to a population (*85*). The B index quantifies the similarity in the log-odds distribution in the population and the sample by comparing the densities of the log-odds of the population and sample probabilities with one another. The log-odds distributions are obtained by regressing a sample indicator (1=college is in sample; 0=college is not in the sample but is in the population) on a set of covariates theorized to be relevant to potential variation in treatment effects, using logistic regression (here, the 12 indicators specified in the first part of Table S14). The index takes on values between 0 and 1. Values ≥ 0.90 indicate very high generalizability (sample estimates of the treatment effect apply directly to the population) and values ≥ 0.80 indicate high generalizability (sample estimates of the treatment effect can be re-weighted to apply to the population).

For bins j= 1…k

$$B = \sum_{j=1}^{k} \sqrt{w_{pj} \times w_{sj}}$$

where w_{pj} is the proportion of the inference population in each bin and w_{sj} is the proportion of the sample (here CTC study colleges) in each bin.

To determine the optimal bin size *h*, we used the formula recommended in (*87*):

$$h = 1.06 \times s \times (N+n)^{-1/5}$$

$$s = \sqrt{\frac{(n-1)s_s^2 + (N-1)s_P^2}{N+n-2}}$$

where s is the pooled standard deviation of the propensity score log-odds across the sample and population, N is the number of units (colleges) in the population, and n is the number of units (CTC colleges) in the sample.

As a further estimate of the similarity between the population and ample distributions, we also computed the coverage rate, which refers to the proportion of population colleges that fall within the same bins as sample colleges in histograms of the log-odds distribution when bins are defined using optimal bin size (*87*).

Although we began with an initial set of 12 covariates, we added one covariate at a time to the logistic regression model through a systematic process, testing a total of 23 alternative covariates to determine which increased the B index value the most. Ultimately, we determined the best model was a set of 15 covariates, the original 12 plus 3 more that emerged from this process (Table S14). The initial inference population of 1301 colleges had a B index of 0.761 and coverage of 74.2% for the 15-covariate model.

Pruning the population. As recommended (*85*), the next step is to seek to identify one or more subpopulations containing a subset of colleges in the target population that are associated with a sufficiently high B index (at least ≥ 0.80) and hence indicate adequate fit between the population and sample. We used three strategies to identify these subpopulations: 1) applying

minimum cutoffs to certain variables within the 15-covariate model (removing from the population colleges whose value for a variable fell below the minimum for the CTC sample for that variable), 2) applying maximum cutoffs to certain variables (removing from the population colleges whose value for a variable exceeded the maximum for the CTC sample for that variable) and 3) applying both minimum and/or maximum cutoffs to one or more variables.

In each case, we tried different combinations of cutoffs for indicators to determine those that had the most impact on the B index. Several combinations (20 of 36) had B indices that exceeded the desired criterion of 0.80. This process did not yield a single best subpopulation. We selected four of these as final subpopulations as a means of accommodating both population size and similarity in population and sample distributions (e.g., B index). See Tables S15 and S16.

Reweighting the Sample to be More Similar to the Population. The next step is to compare the absolute standardized mean difference (ASMD) between the population and sample for each covariate in the propensity score model. This is computed for each covariate by computing the absolute value of the difference between the population mean and the sample mean and dividing that value by the population standard deviation. To achieve adequate covariate balance, the goal is for the magnitude of the ASMD to be < ~0.25 for continuous indicators and for the magnitude of the absolute mean difference (absolute value of the population mean – sample mean) to be < ~0.125 for proportion-based indicators. When ASMDs are of this magnitude or smaller, they tend to be within the realm of covariate adjustment using regression. In our final model, 6 indicators were continuous and 9 were proportion-based indicators.

Once B index values of ≥ 0.80 are achieved, reweighting the sample to be more similar to the population can help further reduce covariate ASMDs and achieve better balance. One method

involves post-stratification (sub-classification). To implement this method, we cut the log-odds distribution in each final subpopulation into five equal size bins/strata, so each bin had $1/5^{th}$ of the number of colleges in the population. We then calculated the re-weighted sample mean for each covariate by taking the mean of each bin mean for bins containing the log-odds values of CTC colleges.

A second method involves inverse probability weighting (IPW). An inverse-probability-weighted mean for each covariate is obtained by regressing that covariate on the sample indicator (1=college is in the sample; 0=college is in the population), using a weight of 1/propensity score (derived from the log-odds distribution) for sample colleges and a weight of 1 for population colleges. The re-weighted sample mean is then the predicted mean for this regression when the sample indicator is set to 1. For each method, the re-weighted sample mean can then be used to recalculate the ASMD for that covariate.

Selecting the final population. We then compared population and sample means for raw covariates before and after reweighting sample means using post-stratification and inverse-probability weighting, for the target population and each of the four subpopulations we identified. See Tables S17-22. This comparison revealed that subpopulations 1 and 2 did not improve covariate balance relative to the target population. Subpopulations 3 and 4 did improve covariate balance. Moreover, for these subpopulations, post-stratification was fully effective in allowing all covariate ASMDs to meet the rules of thumb magnitude guidelines. Inverse probability weighting also eliminated differences on all proportion-based variables but did so for only 67% (4 of 6) of the continuous variables. Thus, post-stratification weights were the best weights to use when weighting the final (treatment effects heterogeneity) regression models.

So, for both subpopulations 3 and 4, post-stratification yields a sample that is similar to the population. In choosing between these two subpopulations, we considered the fact that subpopulation 4 is somewhat larger (749 vs. 681 colleges) and requires only a slight sacrifice to the magnitude of the B index (82.5% vs. 83.9%; moreover, using post-stratification increases the B index of subpopulation 4 to 88.7%). Furthermore, the process for generating subpopulation 4 was simpler, requiring fewer minimum and maximum cutoffs to institutional variables. Therefore, we chose subpopulation 4 as the population to which the sample could best generalize.

Reweighting the Treatment Effect Estimates. As the last step, we formed post-stratification weights for each CTC college using the formula (1/n_bin_sample)*(1/k), where n_bin_sample is the number of CTC colleges whose predicted log-odds value is in each bin in subpopulation 4 and k is 5 for the number of bins/strata. We then executed a version of the primary regression model (FIRC model with local-identity group belonging and historic group achievement moderators) that was weighted using post-stratification weights for subpopulation 4.

Table S1. Completion of saying-is-believing essays, by condition (intent-to-treat analytic sample).

	Active Control (N=13,503)	Standard Belonging Treatment (N=13,408)
Non-blank, non-brief essay	78.5%	85%
Blank essay	13%	10.5%
Brief essay (< 40 words)	8.5%	4.5%
Total	100%	100%

Table S2. Median time spent on randomized content, by condition (intent-to-treat analytic sample).

	Overall	Active Control	Standard Belonging Treatment
Reading Randomized Content	4 min 10 sec	2 min 54 sec	5 min 35 sec
Writing Saying-Is-Believing-Essay	6 min 29 sec	5 min 31 sec	7 min 36 sec
Total	10 min 39 sec	8 min 25 sec	13 min 11 sec

Note. Calculations of time spent writing include only participants who wrote an essay (88%).

Table S3. Engagement in writing the saying-is-believing essay in belonging-treatment conditions in earlier trials and in the CTC Belonging Trial.

	Walton and Cohen, 2011 (*10*)	Yeager, Walton, Brady et al., 2016 Experiment 3 (*12*)	CTC Belonging Trial (here)
Delivery Method	In-Person One-on-One Experience	Online Pre-Matriculation Module	Online Pre-Matriculation Module
Number Colleges	1	1	22
Number Students (Intent-To-Treat Analytic Samples, All Conditions)	91	1,592	26,911
% of Students Writing Essay (Treatment)	100%	91.4%	88%
Median Time Writing Saying-Is-Believing Essay (Treatment)	28 min 53 sec	11 min, 25 sec	7 min 36 sec
Median Words Written (Treatment)	558	215	126

Table S4. Characteristics of partner schools.

Partner School	Type	Carnegie Classification	Total Under-graduates Enrolled	% Black, Latino, Native American Under-graduates	Admissions Rate	SAT Total 25th/75th Percentile Score	6-Year Bachelor's Graduation Rate	End of Fiscal Year Endow-ment (in millions)	Barron's Selectivity	Location	Year(s) of Imple-mentation
Albion College	Private	Baccalaureate Colleges: Liberal Arts	1,376	16%	72%	880/1210	72%	173.07	Competitive	Midwest	2016
Allegheny College	Private	Baccalaureate Colleges: Liberal Arts	1,931	15%	68%	n/a	78%	181.64	Highly Competitive	Midwest	2015, 2016
Bowling Green State University	Public	Doctoral / Research Universities-Intensive	14,334	14%	76%	900/1150	54%	146.74	Competitive	Midwest	2015, 2016
California State University, Dominguez Hills	Public	Masters Colleges and Universities I	12,620	74%*	48%	n/a	32%	9.01	Less Competitive	West	2016
California State University, Northridge	Public	Masters Colleges and Universities I	37,188	51%*	48%	800/1030	47%	86.32	Less Competitive	West	2015, 2016
The College of Wooster	Private	Baccalaureate Colleges: Liberal Arts	2,050	15%	58%	1070/1350	76%	260.41	Very Competitive	Midwest	2015, 2016
Cornell University	Private	Doctoral / Research Universities-Extensive	14,315	19%	14%	1330/1530	93%	4524.42	Most Competitive	Northeast	2015, 2016
Dartmouth College	Private	Doctoral / Research Universities-Intensive	4,307	18%	11%	1350/1560	95%	4474.40	Most Competitive	Northeast	2015, 2016

DePauw University	Private	Baccalaureate Colleges: Liberal Arts	2,231	11%	65%	1040/1280	80%	614.57	Very Competitive +	Midwest	2015, 2016
Hope College	Private	Baccalaureate Colleges: Liberal Arts	3,392	11%	84%	1000/1270	80%	185.87	Very Competitive	Midwest	2015, 2016
Indiana State University	Public	Doctoral / Research Universities-Intensive	11,257	22%	86%	790/1020	40%	42.21	Competitive	Midwest	2015, 2016
Indiana University	Public	Doctoral / Research Universities-Extensive	38,364	10%	79%	1060/1290	78%	991.13	Very Competitive	Midwest	2015, 2016
Kalamazoo College	Private	Baccalaureate Colleges: Liberal Arts	1,443	16%	66%	n/a	82%	206.77	Highly Competitive	Midwest	2016
Lewis & Clark College	Private	Baccalaureate Colleges: Liberal Arts	2,087	14%	55%	1190/1370	79%	204.04	Most Competitive	West	2015, 2016
Ohio Wesleyan University	Private	Baccalaureate Colleges: Liberal Arts	1,671	14%	72%	n/a	66%	201.61	Competitive +	Midwest	2016
Southern Oregon University	Public	Masters Colleges and Universities I	5,490	13%	78%	900/1130	40%	27.23	Competitive	West	2015, 2016
University of California, Santa Cruz	Public	Doctoral / Research Universities-Extensive	16,231	32%*	58%	1060/1290	77%	81.98	Competitive +	West	2015, 2016
University of Central Arkansas	Public	Masters Colleges and Universities I	9,887	22%	90%	870/1060	42%	30.05	Very Competitive	South	2015, 2016
University of Oregon	Public	Doctoral / Research Universities-Extensive	20,538	14%	78%	980/1220	69%	758.69	Competitive +	West	2016

University of Pittsburgh	Public	Doctoral / Research Universities-Extensive	18,908	9%	55%	1190/1380	82%	3501.94	Highly Competitive +	Northeast	2016
Wabash College	Private	Baccalaureate Colleges: Liberal Arts	869	15%	63%	1020 / 1230	72%	327.13	Competitive +	Midwest	2015, 2016
Yale University	Private	Doctoral / Research Universities-Extensive	5,532	20%	6%	1420 / 1600	96%	25413.15	Most Competitive	Northeast	2015, 2016

Note. Selectivity data were obtained from Barron's Profiles of American Colleges (*88*). The ordinal ranking (9 levels) of these categories is: Most competitive, highly competitive +, highly competitive, very competitive +, very competitive, competitive +, competitive, less competitive, noncompetitive. All other data were obtained from the Integrated Postsecondary Education Data System in 2016, except for endowment, which is an end of (2016) fiscal year variable obtained from IPEDS. SAT 25th and 75th percentiles are the sum of the respective Critical Reading and Math scores for each percentile (out of a total possible of 1600). * denotes Hispanic-Serving Institution (HSI) in 2016-2017 as listed by the Hispanic Association of Colleges and Universities (https://www.hacu.net/images/hacu/OPAI/2016 HSI list.pdf). All other institutions are predominately White Institutions (PWIs).

Table S5. Intent-to-treat sample by partner school.

	2015 Cohort			2016 Cohort		
Partner School	Eligible Entering Students	Intent-to-Treat Sample	Intent-to-Treat Participation Rate	Eligible Entering Students	Intent-to-Treat Sample	Intent-to-Treat Participation Rate
Albion College	-	-	-	354	137	39%
Allegheny College	512	170	33%	593	126	21%
Bowling Green State University	3,530	1,191	34%	3,384	1,650	49%
California State University, Dominguez Hills	-	-	-	6,514	1,276	20%
California State University, Northridge	5,325	1,123	21%	3,986	1,366	34%
The College of Wooster	596	311	52%	566	320	57%
Cornell University	3,218	1,829	57%	3,373	1,634	48%
Dartmouth College	1,120	418	37%	1,135	360	32%
DePauw University	617	326	53%	576	326	57%
Hope College	817	188	23%	810	155	19%
Indiana State University	3,020	508	17%	2,071	690	33%
Indiana University	7,681	1,125	15%	7,233	817	11%
Kalamazoo College	-	-	-	362	170	47%
Lewis & Clark College	709	326	46%	516	249	48%
Ohio Wesleyan University	-	-	-	515	231	45%
Southern Oregon University	1,052	324	31%	1,103	304	28%
University of California, Santa Cruz	3,499	1,886	54%	4,499	2,350	52%
University of Central Arkansas	2,062	606	29%	2,913	447	15%
University of Oregon	-	-	-	3,996	1,608	40%
University of Pittsburgh	-	-	-	3,568	780	22%
Wabash College	263	148	56	217	111	51
Yale University	1,388	635	46%	1,378	690	50%
TOTAL	35,409	11,114	31%	49,662	15,797	32%
TOTAL COMBINING COHORTS	**85,071**	**26,911**	**32%**			

Note. The participation rate is the intent-to-treat sample divided by eligible entering students.

Table S6. Distribution of local-identity groups (LIGs) by group historic achievement and belonging affordances, as a function of first-generation and racial-ethnic identity.

First-Generation Status	Race-Ethnicity		# Local-Identity Groups	Low Historic Group Achievement (0-24th percentile)		Medium Historic Group Achievement (25-74th percentile)		High Historic Group Achievement (75-100th percentile)	
				Belonging Affordance		Belonging Affordance		Belonging Affordance	
				Low	Medium/ High	Low	Medium/ High	Low	Medium/ High
First-Generation	Of Hispanic/Latinx American Origin (of any race)		37	5%	32%	11%	32%	3%	16%
	Not of Hispanic/ Latinx Origin	Black/African/African American	36	28%	14%	36%	0%	22%	0%
		Asian/Asian American	24	8%	0%	58%	0%	33%	0%
		Native American/Native Hawaiian/Other Pacific Islander	11	36%	36%	9%	9%	9%	0%
		White/European American	38	0%	13%	61%	0%	0%	26%
		Other	33	3%	36%	15%	18%	9%	18%
		TOTAL First-Generation	**179**	**11%**	**21%**	**21%**	**23%**	**12%**	**12%**
Continuing-Generation	Of Hispanic/Latinx American Origin (of any race)		35	0%	20%	0%	60%	0%	20%
	Not of Hispanic/ Latinx Origin	Black/African/African American	38	18%	16%	26%	16%	13%	11%
		Asian/Asian American	31	6%	3%	32%	23%	23%	13%
		Native American/Native Hawaiian/Other Pacific Islander	17	18%	6%	18%	47%	6%	6%
		White/European American	38	0%	8%	5%	63%	0%	24%
		Other	36	3%	14%	14%	33%	6%	31%
		TOTAL Continuing-Generation	**195**	**7%**	**12%**	**15%**	**40%**	**8%**	**18%**

Note. Values reflect row percentages (i.e., all row percentages sum to 100%). Belonging affordance cut-points determined through Bayesian Causal Forest (-0.5 SD, 36th percentile).

Table S7. Calculation of the group historic achievement moderator.

Average Local-Identity Group Size from 2011 to 2015	Average Type	Average	2015 Local-Identity Groups	2016 Local-Identity Groups
1 to 5 students	Race-Ethnicity × College × Cohort	4-year historical average of Y1 FT Completion	Average of Y1 FT Completion in 2011, 2012, 2013, 2014	Average of Y1 FT Completion in 2012, 2013, 2014, 2015
6 to 9 students	Race-Ethnicity × First-Gen. Status × College × Cohort	3-year historical average of Y1 FT Completion	Average of Y1 FT Completion in 2012, 2013, 2014	Average of Y1 FT Completion in 2013, 2014, 2015
10 or more students	Race-Ethnicity × First-Gen. Status × College × Cohort	2-year historical average of Y1 FT Completion	Average of Y1 FT Completion in 2013, 2014	Average of Y1 FT Completion in 2014, 2015

Note. Y1 FT Completion = Rate of completing the first year of college full-time, among cohorts entering in the indicated years.

Table S8. Correlations between Empirical Bayes estimates of local-identity group spring-term belonging and related group-level variables (intent-to-treat analytic sample).

Measure	Correlation
Current proportional representation of local-identity group on campus[1]	r=0.585, p<0.001
Historic (2-year prior) proportional representation of local-identity group[1]	r=0.583, p<0.001
Mentor development (any mentor)[2]	r=0.164, p=0.003
Mentor development (faculty or administrator mentor)[2]	r=0.194, p=0.004
Having close friends on campus[2]	r=0.431, p<0.0001
Level of closeness of friendships on campus[2]	r=0.192, p=0.0005
Loneliness[2]	r=-0.471, p<0.0001
Involvement in extracurricular organizations/student groups[2]	r=0.144, p=0.011

Note. All measures with [1] superscript are data provided by the participating institutions. "Current" refers to the same institutional cohort as the given randomized sample (either the 2015-2016 or 2016-2017 class year). All measures with [2] superscript are data provided by participating control-condition students during the spring survey. The latter measures are all unconditional Empirical Bayes estimates.

Table S9. Baseline equivalence across conditions (intent-to-treat sample).

			Active Control (N=13,503)	Standard Belonging Treatment (N=13,408)	p-value of difference
Standardized Test Score		ACT score	26.032 (5.82)	26.085 (5.84)	0.490
Gender		Female	58.27%	58.75%	0.430
Generation Status		First-generation status	35.28%	35.23%	0.924
Race-Ethnicity	Of Hispanic/Latinx American Origin (of any race)		19.96%	20.33%	0.958
	Not of Hispanic/ Latinx Origin	Black/African/African American	8.46%	8.54%	
		Asian/Asian American	9.49%	9.57%	
		Native American/Native Hawaiian/Other Pacific Islander	0.47%	0.48%	
		White/European American	51.70%	51.10%	
		Other	9.92%	9.98%	

Note. Columns depict means (SDs) or percents.

Table S10. Conditional average treatment effects (CATES) on the probability of completing the first-year full-time enrolled by group historic achievement rate and belonging affordance across local-identity groups, including primary test (A) and robustness tests (B and C).

A. Intent-To-Treat (ITT) analytic sample (unweighted)				
		Low Historic Achievement	Medium Historic Achievement	High Historic Achievement
Belonging Affordance	Low	b=-.023 [-.051, .005], z=-1.587, p=0.112	b=.0004 [-.019, .020], z=.042, p=0.967	b=.013 [-.014, .040], z=.928, p=0.353
	Medium/ High	b=.020 [.003, .038], z=2.265, p=0.023	b=.013 [.002, .024], z=2.271, p=0.023	b=-.003 [-.020, .013], z=-.412, p=0.681
B. Intent-To-Treat (ITT) analytic sample (weighted)				
		Low Historic Achievement	Medium Historic Achievement	High Historic Achievement
Belonging Affordance	Low	b=-.041 [-.081, -.001], z=-2.020, p=0.043	b=.002 [-.016, .021], z=.249, p=0.804	b=.016 [-.010, .042], z=1.186, p=0.236
	Medium/ High	b=.017 [-.002, .036], z=1.729, p=0.084	b=.022 [.010, .035], z=3.627, p=0.0003	b=.003 [-.016, .021] z=.273, p=0.785
C. Treatment-On-Treated (TOT) sample (unweighted)				
		Low Historic Achievement	Medium Historic Achievement	High Historic Achievement
Belonging Affordance	Low	b=-.037 [-.067, -.008], z=-2.467, p=0.014	b=-.003 [-.023, .018], z=-0.275, p=0.783	b=.015 [-.014, .043], z=1.009, p=0.313
	Medium/ High	b=.015 [-.003, .033], z=1.606, p=0.108	b=.012 [.001, .024], z=2.133, p=0.033	b=-.003 [-.020, 014], z=-0.343, p=0.732

Note. CIs are 95%; cut point for belonging is -0.5 (36^{th} percentile, based on BCF results); cut points for historic achievement are 0-24^{th} percentiles, 25^{th}-74^{th} percentile, 75-100^{th} percentile. The ITT sample is 26,911 students at 22 colleges; the TOT sample is 23,771 students at 22 colleges. The generalizability population—to which the weights applied in Panel B force the composition of our sample of schools to resemble—is 1,019,790 first-time, full-time degree seeking undergraduates at 749 colleges annually.

Table S11. Mean rates of completing the first-full full-time enrolled in the intent-to-treat sample, by condition.

Active Control Condition				
		Low Historic Achievement	Medium Historic Achievement	High Historic Achievement
Belonging Affordance	Low	0.609 (0.010)	0.812 (0.008)	0.948 (0.012)
	Medium/ High	0.572 (0.006)	0.801 (0.004)	0.971 (0.006)
Social-Belonging Treatment Condition				
		Low Historic Achievement	Medium Historic Achievement	High Historic Achievement
Belonging Affordance	Low	0.587 (.011)	0.812 (.008)	0.960 (.012)
	Medium/ High	0.593 (.008)	0.814 (.005)	0.967 (.007)

Note. Marginal means calculated using the model for the 3-way interaction, using the margins command in Stata. Standard errors indicated in parentheses. Data from the intent-to-treat analytic sample (unweighted), N=26,911, k=374.

Table S12. Posterior probability of a positive treatment effect on the rate of completing the first year full-time enrolled by group historic achievement and belonging affordance across local-identity groups.

A. Intent-To-Treat (ITT) analytic sample (unweighted)				
		Low Historic Achievement	Medium Historic Achievement	High Historic Achievement
Belonging Affordance	Low	.5905	.5835	.5485
	Medium/ High	.8755	.8655	.6825
B. Intent-To-Treat (ITT) analytic sample (weighted)				
		Low Historic Achievement	Medium Historic Achievement	High Historic Achievement
Belonging Affordance	Low	.5385	.5850	.5510
	Medium/ High	.8700	.8940	.7070
C. Treatment-On-Treated (TOT) sample (unweighted)				
		Low Historic Achievement	Medium Historic Achievement	High Historic Achievement
Belonging Affordance	Low	.3943	.4930	.4813
	Medium/ High	.7133	.8157	.6767

Note. Values derived from Bayesian Causal Forest (BCF) analyses. The greatest probabilities of a positive treatment effect are among students in local-identity groups medium-to-high in belonging affordances with low or medium historic achievement. Note that the posterior probability for all cells in both ITT samples (Panels A and B) are above .50, indicating that the median effect size in all cells was positive, and that a minimal posterior probability of a negative treatment effect (<.30, or a 70% or greater probability of a negative treatment effect) was found in no cell in any sample.

Table S13. Development of the generalizability sample: Application of rules removing kinds of colleges with no representation in the CTC Belonging Trial sample.

Filtering Rule	Filtering Rule Description	Relevant Variable	N_Include	N_Exclude
Rule0	All Colleges in IPEDS 2011 to 2017 Database		8718	
Rule1	Exclude Colleges with Not Available [-3] & Private for-profit [3]			
	Apply Rule1 to 2015 Variable	CONTROL_2015	5162	3556
	Apply Rule1 to 2016 Variable	CONTROL_2016	4996	166
Rule2	Exclude Colleges with 1-year [1] or 2-year college level [2]			
	Apply Rule2 to 2015 Variable	LEVEL_YR_2015	3401	1595
	Apply Rule2 to 2016 Variable	LEVEL_YR_2016	3387	14
Rule3	Exclude College that is not "degree-granting, primarily baccalaureate"			
	Apply Rule3 to 2015 Variable	INSTCAT_2015	1887	1500
	Apply Rule3 to 2016 Variable	INSTCAT_2016	1855	32
Rule4	Exclude College with "Yes" [1] for Open-Admission Policy			
	Apply Rule4 to 2015 Variable	OPENADMP_yn_2015	1687	168
	Apply Rule4 to 2016 Variable	OPENADMP_yn_2016	1673	14
Rule5	Exclude College with Admissions Rate of 100%			
	Apply Rule5 to 2015 Variable	ADMRATE_2015	1633	40
	Apply Rule5 to 2016 Variable	ADMRATE_2016	1623	10
Rule6	Exclude College with Admissions Rate of 0%			
	Apply Rule6 to 2015 Variable	ADMRATE_2015	1622	1
	Apply Rule6 to 2016 Variable	ADMRATE_2016	1622	0
Rule7	Exclude College with No Endowment			
	Apply Rule7 to 2015 Variable	ENDOWMENT_END_yn_2015	1464	158
	Apply Rule7 to 2016 Variable	ENDOWMENT_END_yn_2016	1463	1
Rule8	Exclude College with Distance Education Only			
	Apply Rule8 to 2015 Variable	DISTNCED_yn_2015	1463	0
	Apply Rule8 to 2016 Variable	DISTNCED_yn_2016	1463	0
Rule9	Exclude Historically Black College or University [1]			
	Apply Rule9 to 2015 Variable	HBCU_2015	1404	59
	Apply Rule9 to 2016 Variable	HBCU_2016	1404	0
Rule10	Exclude College That Does Not Provide On-Campus Housing [0]			
	Apply Rule10 to 2015 Variable	CAMPUS_HOUSING_yn_2015	1334	70
	Apply Rule10 to 2016 Variable	CAMPUS_HOUSING_yn_2016	1331	3

Table S14. Development of the generalizability sample: Covariates for the propensity score model.

Overall Category	Variable	Variable Description	LN
Initial set of 12 covariates for the propensity score model			
Admissions	ADMRATE	Admissions Rate (Number Accepted / Number of Applicants)	
Fall Enrollment	PERC_ALL_UG_FT	Percentage of All Undergraduates Who Are Full-Time Enrolled as of Oct. 15	
	FT_FALL_RET_RATE	Full-Time Fall Retention Rate	
Proportional Representation	PERC_UG_BLN	Percentage of All Undergraduates Who Are Black, Latinx, or Native	
Finance	ENDOWMENT_END	Endowment at End of Fiscal Year (in millions of dollars)	X
Institutional Priorities	PER_UG_RESEARCH	Annual Amount Spent on Research Per Undergraduate	X
	PER_UG_STUD_SERV	Annual Amount Spent on Student Services Per Undergraduate	X
Financial Aid Percentage	PCT_PELL_UG	Percentage of Undergraduates Awarded Pell Grants	
Financial Aid Average Amount	AVG_GRANT_AID_UG	Average annual amount of grant aid awarded (from any source) per undergraduate	X
Graduation	Perc_Bach_6yr_ALL	6-Year (150% Time) Bachelor's Degree Attainment Rate	
Graduation Gap	Perc_Bach_6yr_It	White/Asian - Black/Latinx/Native Difference in 6-Year Bachelor's Degree Rates	
Size	TOTAL_No_UG	Total Number of Undergraduates in Financial Aid Cohort	X
Additional 3 covariates added to form the final model			
Institutional Priorities	PER_UG_INSTRUCTION	Amount Spent on Instruction Per Undergraduate	X
Institutional Priorities	PERC_UG_STUD_SERV	Percentage Spent on Student Services Per Undergraduate vs Other Priorities	
Graduation Gap	Perc_Bach_6yr_Gap	White/Asian - Black/Latinx/Native Difference in 4-Year Bachelor's Degree Rates	

Note. All covariates represent the annual average at a given institution for the years 2015 and 2016 or 2013 and 2014, if values for 2015 and 2016 were not available, standardized within the target population of colleges (N=1331). For variables with an "X" in the LN column, the natural log of (the indicator + the indicator minimum value + 1) was taken before it was standardized.

Table S15. Development of the generalizability sample: Characteristics of final subpopulations with B indices exceeding 0.80 using the 15-covariate model.

Population or Subpopulation	# of Colleges	# of Variables Subject to Minimum Cutoffs	# of Variables Subject to Maximum Cutoffs	Bin Size	B Index	Coverage
0	1301	None	None	0.434	0.761	0.742
1	817	7	0	0.344	0.825	0.823
2	906	5	0	0.360	0.802	0.797
3	681	6	2	0.321	0.839	0.884
4	749	5	1	0.342	0.825	0.832

Note. All populations include the 12 covariates listed in Table A1.2 plus the 3 covariates in gray in Table A1.3. The coverage rate refers to the proportion of population colleges that fall within the same bins as sample colleges in histograms of the log-odds distribution when bins are defined using optimal bin size (*87*).

Table S16. Development of the generalizability sample: Variables subject to minimum and maximum cutoffs in final subpopulations.

	Type of Cutoff	Variable
Subpopulation 1	Minimum	Endowment at End of Fiscal Year
		Total Undergraduates in Financial Aid Cohort
		Amount Spent on Instruction Per Undergraduate
		Percentage of Undergraduates Awarded Pell Grants
		Amount Spent on Student Services Per Undergraduate
		White/Asian-Black/Latino/Native Gap in 6-Year Bachelor's Degree Rates
		6-Year (150% Time) Bachelor's Degree Attainment Rate
	Maximum	[none]
Subpopulation 2	Minimum	Endowment at End of Fiscal Year
		Percentage of Undergraduates Awarded Pell Grants
		Amount Spent on Student Services Per Undergraduate
		White/Asian-Black/Latino/Native Gap in 6-Year Bachelor's Degree Rates
		6-Year (150% Time) Bachelor's Degree Attainment Rate
	Maximum	[none]
Subpopulation 3	Minimum	Endowment at End of Fiscal Year
		Total Undergraduates in Financial Aid Cohort
		Amount Spent on Instruction Per Undergraduate
		Percentage of Undergraduates Awarded Pell Grants
		White/Asian-Black/Latino/Native Gap in 6-Year Bachelor's Degree Rates
		6-Year (150% Time) Bachelor's Degree Attainment Rate
	Maximum	Average annual amount of grant aid awarded per undergraduate
		White/Asian-Black/Latino/Native Gap in 6-Year Bachelor's Degree Rates
Subpopulation 4	Minimum	[same as Subpopulation 2]
	Maximum	White/Asian-Black/Latino/Native Gap in 6-Year Bachelor's Degree Rates

Table S17. Development of the generalizability sample: Target Population (N=1301) Covariates Summary

Variable	Raw POP Mean	Raw POP SD	Raw CTC Mean	Raw ASMD	Abs. Raw Mean Diff.	Subclass CTC Mean	Subclass ASMD	Abs. Subclass Mean Diff.	IPW CTC Mean	IPW ASMD	IPW Standardized Coefficient	Abs. IPW Mean Diff.
ADMRATE	64.8%	18.9%	60.8%	0.21	0.04	69.4%	0.24	0.05	74.4%	0.51	0.5	0.10
PERC_ALL_UG_FT	85.4%	13.2%	92.2%	0.52*	0.07	86.5%	0.09	0.01	79.4%	0.45	0.44	0.06
FT_FALL_RET_RATE	77.3%	10.9%	85.0%	0.71***	0.08	79.7%	0.22	0.02	76.6%	0.06	0.06	0.01
PERC_UG_BLN	20.8%	14.3%	19.7%	0.07	0.01	24.1%	0.24	0.03	25.6%	0.34	0.33	0.05
ENDOWMENT_END	$370	$1,778	$1,946	0.89***		$830	0.26		$330	0.02	0.07	
PER_UG_INSTRUCTION	$15,562	$22,184	$29,462	0.63**		$18,133	0.12		$12,053	0.16	0.34*	
PER_UG_RESEARCH	$3,365	$16,763	$9,680	0.38		$4,075	0.04		$1,522	0.11	0.26	
PER_UG_STUD_SERV	$5,001	$3,912	$8,294	0.84***		$4,858	0.04		$2,984	0.52	0.97***	
PERC_UG_STUD_SERV	17.2%	8.2%	15.1%	0.25	0.02	13.1%	0.49*	0.04	12.1%	0.62	0.61**	0.05
PCT_PELL_UG	33.6%	12.9%	27.9%	0.44*	0.06	34.6%	0.08	0.01	38.2%	0.36	0.35	0.05
AVG_GRANT_AID_UG	$15,014	$8,418	$19,564	0.54*		$14,670	0.04		$10,761	0.51	0.62*	
Perc_Bach_6yr_ALL	57.7%	17.1%	70.0%	0.72***	0.12	57.8%	0	0.00	49.2%	0.50	0.49	0.09
Perc_Bach_4yr_Gap	13.1%	10.1%	12.2%	0.08	0.01	11.6%	0.14	0.01	9.3%	0.37	0.37	0.04
Perc_Bach_6yr_Gap	12.2%	10.8%	9.7%	0.24	0.03	10.5%	0.16	0.02	7.9%	0.40	0.4	0.04
TOTAL_No_UG	5,792	7,663	10,179	0.57**		9,074	0.43*		8,904	0.41	0.7***	

Note. Columns 1 to 5 correspond to raw values of each covariate in the population and CTC sample. Column 4 is the raw absolute standardized mean difference between the raw population mean and raw CTC sample mean for each covariate. Column 5 is the absolute raw mean difference between the raw population mean and raw CTC sample mean for each covariate for proportion/percentage variables only. Column 6 corresponds to a CTC sample mean for each covariate reweighted using post-stratification/sub-classification in which the population was divided into five equal size bins by applying appropriate cutpoints to the predicted log-odds distribution that a college in the population was in the sample controlling for these 15 covariates. Columns 7 and 8 are the absolute standardized mean difference and absolute mean difference (proportion variables only) between the sub-classification CTC mean and the raw population mean. Column 9 corresponds to a CTC sample mean for each covariate reweighted using inverse probability weighting (IPW), where the weight for each sample college was 1/propensity score, the probability that a college in the population was in the sample controlling for these 15 covariates. Columns 10 and 12 are the absolute standardized mean difference and absolute mean difference (proportion variables only) between the inverse-probability-weighted CTC mean and the raw population mean. Column 1 is the regression coefficient from an inverse-probability-weighted regression model in which both population and sample covariate values were standardized in the target population (N=1331 colleges with available data for each covariate). For columns 4, 7, and 11, ***$p \leq .001$, **$p \leq .01$, *$p \leq .05$. Endowment is in millions of dollars. All other monetary variables are in their original metric.

Table S18. Development of the generalizability sample: Subpopulation 1 (N=817) Covariates Summary

Variable	Raw POP Mean	Raw POP SD	Raw CTC Mean	Raw ASMD	Abs. Raw Mean Diff.	Subclass CTC Mean	Subclass ASMD	Abs. Subclass Mean Diff	IPW CTC Mean	IPW ASMD	IPW Standardized Coefficient	Abs. IPW Mean Diff.
ADMRATE	66.0%	17.9%	60.8%	0.29	0.05	67.9%	0.11	0.02	73.2%	0.40	0.38	0.07
PERC_ALL_UG_FT	87.3%	11.3%	92.2%	0.44*	0.05	89.2%	0.17	0.02	83.5%	0.34	0.28	0.04
FT_FALL_RET_RATE	80.5%	8.3%	85.0%	0.54*	0.05	81.9%	0.17	0.01	78.1%	0.29	0.22	0.02
PERC_UG_BLN	18.9%	12.1%	19.7%	0.07	0.01	21.0%	0.17	0.02	24.4%	0.45	0.38	0.06
ENDOWMENT_END	$444	$1,763	$1,946	0.85***		$1,006	0.32		$473	0.02	0.18	
PER_UG_INSTRUCTION	$16,168	$20,548	$29,462	0.65**		$20,428	0.21		$13,903	0.11	0.32	
PER_UG_RESEARCH	$3,787	$15,817	$9,680	0.37		$4,772	0.06		$2,191	0.10	0.17	
PER_UG_STUD_SERV	$4,966	$3,532	$8,294	0.94***		$5,612	0.18		$3,553	0.40	0.84**	
PERC_UG_STUD_SERV	16.3%	7.5%	15.1%	0.15	0.01	13.7%	0.34	0.03	12.4%	0.51	0.46*	0.04
PCT_PELL_UG	30.7%	10.6%	27.9%	0.26	0.03	30.8%	0.01	0.00	36.2%	0.52	0.42	0.05
AVG_GRANT_AID_UG	$15,709	$8,374	$19,564	0.46*		$16,634	0.11		$12,286	0.41	0.51	
Perc_Bach_6yr_ALL	62.6%	14.0%	70.0%	0.53*	0.07	63.3%	0.05	0.01	53.6%	0.64	0.52	0.09
Perc_Bach_4yr_Gap	14.7%	8.0%	12.2%	0.31	0.02	12.4%	0.29	0.02	11.5%	0.40	0.32	0.03
Perc_Bach_6yr_Gap	13.2%	8.2%	9.7%	0.43*	0.04	10.5%	0.33	0.03	10.4%	0.35	0.26	0.03
TOTAL_No_UG	7,072	8,358	10,179	0.37		9,894	0.34		9,919	0.34	0.45**	

Note. Columns 1 to 5 correspond to raw values of each covariate in the population and CTC sample. Column 4 is the raw absolute standardized mean difference between the raw population mean and raw CTC sample mean for each covariate. Column 5 is the absolute raw mean difference between the raw population mean and raw CTC sample mean for each covariate for proportion/percentage variables only. Column 6 corresponds to a CTC sample mean for each covariate reweighted using post-stratification/sub-classification in which the population was divided into five equal size bins by applying appropriate cutpoints to the predicted log-odds distribution that a college in the population was in the sample controlling for these 15 covariates. Columns 7 and 8 are the absolute standardized mean difference and absolute mean difference (proportion variables only) between the sub-classification CTC mean and the raw population mean. Column 9 corresponds to a CTC sample mean for each covariate reweighted using inverse probability weighting (IPW), where the weight for each sample college was 1/propensity score, the probability that a college in the population was in the sample controlling for these 15 covariates. Columns 10 and 12 are the absolute standardized mean difference and absolute mean difference (proportion variables only) between the inverse-probability-weighted CTC mean and the raw population mean. Column 1 is the regression coefficient from an inverse-probability-weighted regression model in which both population and sample covariate values were standardized in the target population (N=1331 colleges with available data for each covariate). For columns 4, 7, and 11, ***$p \leq .001$, **$p \leq .01$, *$p \leq .05$. Endowment is in millions of dollars. All other monetary variables are in their original metric.

Table S19. Development of the generalizability sample: Subpopulation 2 (N=906) Covariates Summary

Variable	Raw POP Mean	Raw POP SD	Raw CTC Mean	Raw ASMD	Abs. Raw Mean Diff.	Subclass CTC Mean	Subclass ASMD	Abs. Subclass Mean Diff.	IPW CTC Mean	IPW ASMD	IPW Standardized Coefficient	Abs IPW Mean Diff.
ADMRATE	65.9%	17.6%	60.8%	0.29	0.05	68.8%	0.16	0.03	73.6%	0.44	0.41	0.08
PERC_ALL_UG_FT	87.1%	11.8%	92.2%	0.44*	0.05	88.4%	0.12	0.01	83.1%	0.33	0.29	0.04
FT_FALL_RET_RATE	79.6%	8.8%	85.0%	0.61**	0.05	80.4%	0.1	0.01	77.7%	0.21	0.17	0.02
PERC_UG_BLN	19.0%	12.1%	19.7%	0.06	0.01	21.4%	0.2	0.02	24.4%	0.45	0.37	0.05
ENDOWMENT_END	$407	$1,678	$1,946	0.92***		$982	0.34		$441	0.02	0.15	
PER_UG_INSTRUCTION	$15,776	$19,830	$29,462	0.69**		$19,669	0.2		$13,547	0.11	0.3	
PER_UG_RESEARCH	$3,430	$15,059	$9,680	0.42		$4,779	0.09		$2,057	0.09	0.25	
PER_UG_STUD_SERV	$5,034	$3,471	$8,294	0.94***		$5,232	0.06		$3,463	0.45	0.89**	
PERC_UG_STUD_SERV	16.8%	7.7%	15.1%	0.22	0.02	13.1%	0.49*	0.04	12.4%	0.57	0.53**	0.04
PCT_PELL_UG	31.7%	10.8%	27.9%	0.35	0.04	32.2%	0.05	0.01	36.5%	0.45	0.37	0.05
AVG_GRANT_AID_UG	$15,677	$8,245	$19,564	0.47*		$15,282	0.05		$12,093	0.43	0.54*	
Perc_Bach_6yr_ALL	61.6%	14.1%	70.0%	0.6**	0.08	60.7%	0.06	0.01	52.9%	0.61	0.5	0.09
Perc_Bach_4yr_Gap	14.8%	8.3%	12.2%	0.3	0.03	12.8%	0.24	0.02	11.3%	0.42	0.34	0.03
Perc_Bach_6yr_Gap	13.5%	8.6%	9.7%	0.44*	0.04	11.5%	0.23	0.02	10.3%	0.37	0.29	0.03
TOTAL_No_UG	6,700	8,372	10,179	0.42		10,356	0.44*		9,690	0.36	0.54***	

Note. Columns 1 to 5 correspond to raw values of each covariate in the population and CTC sample. Column 4 is the raw absolute standardized mean difference between the raw population mean and raw CTC sample mean for each covariate. Column 5 is the absolute raw mean difference between the raw population mean and raw CTC sample mean for each covariate for proportion/percentage variables only. Column 6 corresponds to a CTC sample mean for each covariate reweighted using post-stratification/sub-classification in which the population was divided into five equal size bins by applying appropriate cutpoints to the predicted log-odds distribution that a college in the population was in the sample controlling for these 15 covariates. Columns 7 and 8 are the absolute standardized mean difference and absolute mean difference (proportion variables only) between the sub-classification CTC mean and the raw population mean. Column 9 corresponds to a CTC sample mean for each covariate reweighted using inverse probability weighting (IPW), where the weight for each sample college was 1/propensity score, the probability that a college in the population was in the sample controlling for these 15 covariates. Columns 10 and 12 are the absolute standardized mean difference and absolute mean difference (proportion variables only) between the inverse-probability-weighted CTC mean and the raw population mean. Column 1 is the regression coefficient from an inverse-probability-weighted regression model in which both population and sample covariate values were standardized in the target population (N=1331 colleges with available data for each covariate). For columns 4, 7, and 11, ***$p \leq .001$, **$p \leq .01$, *$p \leq .05$. Endowment is in millions of dollars. All other monetary variables are in their original metric.

Table S20. Development of the generalizability sample: Subpopulation 3 (N=681) Covariates Summary

Variable	Raw POP Mean	Raw POP SD	Raw CTC Mean	Raw ASMD	Abs. Raw Mean Diff.	Subclass CTC Mean	Subclass ASMD	Abs. Subclass Mean Diff.	IPW CTC Mean	IPW ASMD	IPW Standardized Coefficient	Abs. IPW Mean Diff.
ADMRATE	65.0%	18.5%	60.8%	0.23	0.04	69.9%	0.27	0.05	73.7%	0.48	0.46	0.09
PERC_ALL_UG_FT	87.9%	10.9%	92.2%	0.4	0.04	86.1%	0.16	0.02	81.5%	0.59	0.48	0.06
FT_FALL_RET_RATE	81.0%	8.4%	85.0%	0.47*	0.04	80.4%	0.07	0.01	77.5%	0.42	0.32	0.04
PERC_UG_BLN	19.6%	12.4%	19.7%	0.01	0.00	20.5%	0.07	0.01	23.5%	0.31	0.27	0.04
ENDOWMENT_END	$511	$1,919	$1,946	0.75***		$875	0.19		$512	0.00	0.3	
PER_UG_INSTRUCTION	$17,000	$22,270	$29,462	0.56**		$18,377	0.06		$14,192	0.13	0.39*	
PER_UG_RESEARCH	$4,386	$17,108	$9,680	0.31		$4,208	0.01		$2,334	0.12	0.04	
PER_UG_STUD_SERV	$4,905	$3,758	$8,294	0.9***		$4,863	0.01		$3,490	0.38	0.82**	
PERC_UG_STUD_SERV	15.4%	7.5%	15.1%	0.03	0.00	13.0%	0.32	0.02	12.1%	0.44	0.39*	0.03
PCT_PELL_UG	30.6%	11.0%	27.9%	0.24	0.03	32.0%	0.13	0.01	36.2%	0.51	0.43	0.06
AVG_GRANT_AID_UG	$15,470	$8,841	$19,564	0.46*		$14,451	0.12		$11,763	0.42	0.54	
Perc_Bach_6yr_ALL	63.1%	14.5%	70.0%	0.48*	0.07	59.8%	0.23	0.03	52.2%	0.75	0.63*	0.11
Perc_Bach_4yr_Gap	12.0%	5.6%	12.2%	0.04	0.00	11.0%	0.18	0.01	10.0%	0.36	0.2	0.02
Perc_Bach_6yr_Gap	11.0%	6.4%	9.7%	0.2	0.01	8.8%	0.34	0.02	8.5%	0.39	0.23	0.02
TOTAL_No_UG	7,941	8,789	10,179	0.25		10,126	0.25		9,385	0.16	0.32*	

Note. Columns 1 to 5 correspond to raw values of each covariate in the population and CTC sample. Column 4 is the raw absolute standardized mean difference between the raw population mean and raw CTC sample mean for each covariate. Column 5 is the absolute raw mean difference between the raw population mean and raw CTC sample mean for each covariate for proportion/percentage variables only. Column 6 corresponds to a CTC sample mean for each covariate reweighted using post-stratification/sub-classification in which the population was divided into five equal size bins by applying appropriate cutpoints to the predicted log-odds distribution that a college in the population was in the sample controlling for these 15 covariates. Columns 7 and 8 are the absolute standardized mean difference and absolute mean difference (proportion variables only) between the sub-classification CTC mean and the raw population mean. Column 9 corresponds to a CTC sample mean for each covariate reweighted using inverse probability weighting (IPW), where the weight for each sample college was 1/propensity score, the probability that a college in the population was in the sample controlling for these 15 covariates. Columns 10 and 12 are the absolute standardized mean difference and absolute mean difference (proportion variables only) between the inverse-probability-weighted CTC mean and the raw population mean. Column 1 is the regression coefficient from an inverse-probability-weighted regression model in which both population and sample covariate values were standardized in the target population (N=1331 colleges with available data for each covariate). For columns 4, 7, and 11, ***$p \leq .001$, **$p \leq .01$, *$p \leq .05$. Endowment is in millions of dollars. All other monetary variables are in their original metric.

Table S21. Development of the generalizability sample: Subpopulation 4 (N=749) Covariates Summary

Variable	Raw POP Mean	Raw POP SD	Raw CTC Mean	Raw ASMD	Abs. Raw Mean Diff.	Subclass CTC Mean	Subclass ASMD	Abs. Subclass Mean Diff.	IPW CTC Mean	IPW ASMD	IPW Standardized Coefficient	Abs. IPW Mean Diff.
ADMRATE	64.9%	18.3%	60.8%	0.23	0.04	69.9%	0.27	0.05	73.7%	0.48	0.46	0.09
PERC_ALL_UG_FT	87.7%	11.5%	92.2%	0.4	0.05	86.1%	0.14	0.02	81.3%	0.56	0.48	0.06
FT_FALL_RET_RATE	80.1%	9.0%	85.0%	0.54*	0.05	80.2%	0.01	0.00	77.4%	0.31	0.25	0.03
PERC_UG_BLN	19.7%	12.5%	19.7%	0	0.00	21.5%	0.14	0.02	23.8%	0.33	0.28	0.04
ENDOWMENT_END	$471	$1,836	$1,946	0.8***		$714	0.13		$498	0.02	0.25	
PER_UG_INSTRUCTION	$16,644	$21,565	$29,462	0.59**		$17,779	0.05		$14,038	0.12	0.36	
PER_UG_RESEARCH	$3,975	$16,353	$9,680	0.35		$3,338	0.04		$2,291	0.10	0.12	
PER_UG_STUD_SERV	$5,017	$3,706	$8,294	0.88***		$4,801	0.06		$3,497	0.41	0.86**	
PERC_UG_STUD_SERV	16.0%	7.7%	15.1%	0.11	0.01	13.2%	0.37	0.03	12.2%	0.49	0.45*	0.04
PCT_PELL_UG	31.5%	11.2%	27.9%	0.32	0.04	32.7%	0.11	0.01	36.5%	0.44	0.37	0.05
AVG_GRANT_AID_UG	$15,543	$8,761	$19,564	0.46*		$15,194	0.04		$11,780	0.43	0.55	
Perc_Bach_6yr_ALL	62.1%	14.6%	70.0%	0.55*	0.08	58.9%	0.22	0.03	51.9%	0.70	0.59	0.10
Perc_Bach_4yr_Gap	12.0%	5.8%	12.2%	0.04	0.00	10.6%	0.25	0.01	9.8%	0.38	0.22	0.02
Perc_Bach_6yr_Gap	11.1%	6.5%	9.7%	0.22	0.01	8.4%	0.42	0.03	8.3%	0.43	0.26	0.03
TOTAL_No_UG	7,402	8,679	10,179	0.32		7,914	0.06		9,134	0.20	0.42**	

Note. Columns 1 to 5 correspond to raw values of each covariate in the population and CTC sample. Column 4 is the raw absolute standardized mean difference between the raw population mean and raw CTC sample mean for each covariate. Column 5 is the absolute raw mean difference between the raw population mean and raw CTC sample mean for each covariate for proportion/percentage variables only. Column 6 corresponds to a CTC sample mean for each covariate reweighted using post-stratification/sub-classification in which the population was divided into five equal size bins by applying appropriate cutpoints to the predicted log-odds distribution that a college in the population was in the sample controlling for these 15 covariates. Columns 7 and 8 are the absolute standardized mean difference and absolute mean difference (proportion variables only) between the sub-classification CTC mean and the raw population mean. Column 9 corresponds to a CTC sample mean for each covariate reweighted using inverse probability weighting (IPW), where the weight for each sample college was 1/propensity score, the probability that a college in the population was in the sample controlling for these 15 covariates. Columns 10 and 12 are the absolute standardized mean difference and absolute mean difference (proportion variables only) between the inverse-probability-weighted CTC mean and the raw population mean. Column 1 is the regression coefficient from an inverse-probability-weighted regression model in which both population and sample covariate values were standardized in the target population (N=1331 colleges with available data for each covariate). For columns 4, 7, and 11, ***$p \leq .001$, **$p \leq .01$, *$p \leq .05$. Endowment is in millions of dollars. All other monetary variables are in their original metric.

Table S22. Development of the generalizability sample: Effectiveness of post-stratification (sub-classification) and inverse probability weighting (IPW) in achieving covariate balance.

Target Population (N=1301)	Raw	Subclass	IPW
Total Covariates	15	15	15
Proportion	9	9	9
Continuous	6	6	6
Number ASMD Continuous ≤ 0.25	0	4	3
Perc ASMD Continuous ≤ 0.25	0.0%	66.7%	50.0%
Number ASMD Proportion ≤ 0.125	9	9	9
Perc ASMD Proportion ≤ 0.125	100.0%	100.0%	100.0%

Subpopulation 1 (N=817)	Raw	Subclass	IPW
Total Covariates	15	15	15
Proportion	9	9	9
Continuous	6	6	6
Number ASMD Continuous ≤ 0.25	0	4	3
Perc ASMD Continuous ≤ 0.25	0.0%	66.7%	50.0%
Number ASMD Proportion ≤ 0.125	9	9	9
Perc ASMD Proportion ≤ 0.125	100.0%	100.0%	100.0%

Subpopulation 2 (N=906)	Raw	Subclass	IPW
Total Covariates	15	15	15
Proportion	9	9	9
Continuous	6	6	6
Number ASMD Continuous ≤ 0.25	0	4	3
Perc ASMD Continuous ≤ 0.25	0.0%	66.7%	50.0%
Number ASMD Proportion ≤ 0.125	9	9	9
Perc ASMD Proportion ≤ 0.125	100.0%	100.0%	100.0%

Subpopulation 3 (N=681)	Raw	Subclass	IPW
Total Covariates	15	15	15
Proportion	9	9	9
Continuous	6	6	6
Number ASMD Continuous ≤ 0.25	0	6	4
Perc ASMD Continuous ≤ 0.25	0.0%	100.0%	66.7%
Number ASMD Proportion ≤ 0.125	9	9	9
Perc ASMD Proportion ≤ 0.125	100.0%	100.0%	100.0%

Subpopulation 4 (N=749)	Raw	Subclass	IPW
Total Covariates	15	15	15
Proportion	9	9	9
Continuous	6	6	6
Number ASMD Continuous ≤ 0.25	0	6	4
Perc ASMD Continuous ≤ 0.25	0.0%	100.0%	66.7%
Number ASMD Proportion ≤ 0.125	9	9	9
Perc ASMD Proportion ≤ 0.125	100.0%	100.0%	100.0%

Note. ASMD values for continuous covariates ≤ ASMD > and AMD values for proportion-based covariates ≤ 12.5 are within the realm of covariate adjustment using regression. Using these rules of thumb, the sample and population are very different in raw values for six variables. Using sub-classification (post-stratification) eliminates differences on all variables for subpopulations 3 and 4. Subpopulations 1 and 2 do not improve covariate balance relative to the target population. Using IPW (inverse probability weighting) eliminates differences on all proportion-based variables and 67% (4 of 6) continuous variables for subpopulations 3 and 4. Both subpopulation 3 and 4 achieve similar covariate balance, especially with post-stratification. Subpopulation 3 has a slightly higher B index (83.9% vs. 82.5%). Subpopulation 4 is somewhat larger in size (749 vs. 681 colleges). Using post-stratification increases the B index to 92.5% and 88.7% for subpopulations 3 and 4, respectively. For both subpopulations 3 and 4, post-stratification yields a sample that is similar to the generalizability population.

References and Notes

1. R. Chetty, J. N. Friedman, E. Saez, N. Turner, D. Yagan, "Mobility report cards: The role of colleges in intergenerational mobility" (NBER Working Paper 23618, National Bureau of Economic Research, 2017); http://www.nber.org/papers/w23618.

2. J. Rothwell, "What colleges do for local economies: A direct measure based on consumption," Brookings Institution, 17 November 2015; https://www.brookings.edu/research/what-colleges-do-for-local-economies-a-direct-measure-based-on-consumption/.

3. E. Aucejo, Z. Tobin, "Assessing racial disparities in postsecondary education," Federal Reserve Bank of Boston, 5 October 2021; https://www.bostonfed.org/-/media/Documents/events/2021/racial-disparities-in-todays-economy/Assessing-Racial-Disparities-in-Postsecondary-Education.pdf?la=en.

4. P. Tough, *The Inequality Machine: How College Divides Us* (Mariner Books, 2021).

5. J. A. Berlin, Invited commentary: Benefits of heterogeneity in meta-analysis of data from epidemiologic studies. *Am. J. Epidemiol.* **142**, 383–387 (1995). doi:10.1093/oxfordjournals.aje.a117645 Medline

6. B. B. McShane, J. L. Tackett, U. Böckenholt, A. Gelman, Large-scale replication projects in contemporary psychological research. *Am. Stat.* **73** (suppl. 1), 99–105 (2019). doi:10.1080/00031305.2018.1505655

7. C. J. Bryan, E. Tipton, D. S. Yeager, Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nat. Hum. Behav.* **5**, 980–989 (2021). doi:10.1038/s41562-021-01143-3 Medline

8. G. M. Walton, S. T. Brady, "The social-belonging intervention" in *Handbook of Wise Interventions: How Social Psychology Can Help People Change*, G. M. Walton, A. J. Crum, Eds. (The Guilford Press, 2021), pp. 36–62.

9. D. A. Prentice, D. T. Miller, Pluralistic Ignorance and the Perpetuation of Social Norms by Unwitting Actors. *Adv. Exp. Soc. Psychol.* **28**, 161–209 (1996). doi:10.1016/S0065-2601(08)60238-5

10. G. M. Walton, G. L. Cohen, A brief social-belonging intervention improves academic and health outcomes of minority students. *Science* **331**, 1447–1451 (2011). doi:10.1126/science.1198364 Medline

11. S. T. Brady, G. L. Cohen, S. N. Jarvis, G. M. Walton, A brief social-belonging intervention in college improves adult outcomes for black Americans. *Sci. Adv.* **6**, eaay3689 (2020). doi:10.1126/sciadv.aay3689 Medline

12. D. S. Yeager, G. M. Walton, S. T. Brady, E. N. Akcinar, D. Paunesku, L. Keane, D. Kamentz, G. Ritter, A. L. Duckworth, R. Urstein, E. M. Gomez, H. R. Markus, G. L. Cohen, C. S. Dweck, Teaching a lay theory before college narrows achievement gaps at scale. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E3341–E3348 (2016). doi:10.1073/pnas.1524360113 Medline

13. M. C. Murphy, M. Gopalan, E. R. Carter, K. T. U. Emerson, B. L. Bottoms, G. M. Walton, A customized belonging intervention improves retention of socially disadvantaged students

at a broad-access university. *Sci. Adv.* **6**, eaba4677 (2020). doi:10.1126/sciadv.aba4677 Medline

14. K. R. Binning, N. Kaufmann, E. M. McGreevy, O. Fotuhi, S. Chen, E. Marshman, Z. Y. Kalender, L. B. Limeri, L. Betancur, C. Singh, Changing social contexts to foster equity in college science courses: An ecological-belonging intervention. *Psychol. Sci.* **31**, 1059–1070 (2020). doi:10.1177/0956797620929984 Medline

15. G. M. Walton, C. Logel, J. M. Peach, S. J. Spencer, M. P. Zanna, Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. *J. Educ. Psychol.* **107**, 468–485 (2015). doi:10.1037/a0037461

16. M. Broda, J. Yun, B. Schneider, D. S. Yeager, G. M. Walton, M. Diemer, Reducing inequality in academic success for incoming college students: A randomized trial of growth mindset and belonging interventions. *J. Res. Educ. Eff.* **11**, 317–338 (2018). doi:10.1080/19345747.2018.1429037

17. J. M. Harackiewicz, S. J. Priniski, Improving student outcomes in higher education: The science of targeted intervention. *Annu. Rev. Psychol.* **69**, 409–435 (2018). doi:10.1146/annurev-psych-122216-011725 Medline

18. C. M. Steele, A threat in the air: How stereotypes shape intellectual identity and performance. *Am. Psychol.* **52**, 613–629 (1997). doi:10.1037/0003-066X.52.6.613 Medline

19. G. M. Walton, G. L. Cohen, A question of belonging: Race, social fit, and achievement. *J. Pers. Soc. Psychol.* **92**, 82–96 (2007). doi:10.1037/0022-3514.92.1.82 Medline

20. N. M. Stephens, S. A. Fryberg, H. R. Markus, C. S. Johnson, R. Covarrubias, Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first-generation college students. *J. Pers. Soc. Psychol.* **102**, 1178–1197 (2012). doi:10.1037/a0027143 Medline

21. T. N. Brannon, A. Lin, "Pride and prejudice" pathways to belonging: Implications for inclusive diversity practices within mainstream institutions. *Am. Psychol.* **76**, 488–501 (2021). doi:10.1037/amp0000643 Medline

22. S. Murrar, M. R. Campbell, M. Brauer, Exposure to peers' pro-diversity attitudes increases inclusion and reduces the achievement gap. *Nat. Hum. Behav.* **4**, 889–897 (2020). doi:10.1038/s41562-020-0899-5 Medline

23. D. L. Gray, E. C. Hope, J. S. Matthews, Black and belonging at school: A case for interpersonal, instructional, and institutional opportunity structures. *Educ. Psychol.* **53**, 97–113 (2018). doi:10.1080/00461520.2017.1421466

24. M. Gopalan, S. T. Brady, College students' sense of belonging: A national perspective. *Educ. Res.* **49**, 134–137 (2020). doi:10.3102/0013189X19897622

25. E. Goffman, *Stigma: Notes on the Management of Spoiled Identity* (Prentice-Hall, 1963).

26. M. Cikara, J. E. Martinez, N. A. Lewis Jr., Moving beyond social categories by incorporating context in social psychological theory. *Nat. Rev. Psychol.* **1**, 537–549 (2022). doi:10.1038/s44159-022-00079-3

27. G. D. Borman, J. Grigg, C. S. Rozek, P. Hanselman, N. A. Dewey, Self-affirmation effects are produced by school context, student engagement with the intervention, and time: Lessons from a district-wide implementation. *Psychol. Sci.* **29**, 1773–1784 (2018). doi:10.1177/0956797618784016 Medline

28. K. Deaux, N. Bikmen, A. Gilkes, A. Ventuneac, Y. Joseph, Y. A. Payne, C. M. Steele, Becoming American: Stereotype threat effects in Afro-Caribbean immigrant groups. *Soc. Psychol. Q.* **70**, 384–404 (2007). doi:10.1177/019027250707000408

29. E. A. Canning, K. Muenks, D. J. Green, M. C. Murphy, 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). doi:10.1126/sciadv.aau4734 Medline

30. J. G. Starck, S. Sinclair, J. N. Shelton, How university diversity rationales inform student preferences and outcomes. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2013833118 (2021). doi:10.1073/pnas.2013833118 Medline

31. A. M. Locks, S. Hurtado, N. A. Bowman, L. Oseguera, Extending notions of campus climate and diversity to students' transition to college. *Rev. Higher Educ.* **31**, 257–285 (2008). doi:10.1353/rhe.2008.0011

32. N. J. Shook, R. Clay, Interracial roommate relationships: A mechanism for promoting sense of belonging at university and academic performance. *J. Exp. Soc. Psychol.* **48**, 1168–1172 (2012). doi:10.1016/j.jesp.2012.05.005

33. R. M. Carey, N. M. Stephens, S. S. M. Townsend, M. G. Hamedani, Is diversity enough? Cross-race and cross-class interactions in college occur less often than expected, but benefit members of lower status groups when they occur. *J. Pers. Soc. Psychol.* **123**, 889–908 (2022). doi:10.1037/pspa0000302 Medline

34. K. L. Milkman, M. Akinola, D. Chugh, What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *J. Appl. Psychol.* **100**, 1678–1712 (2015). doi:10.1037/apl0000022 Medline

35. D. S. Yeager, P. Hanselman, G. M. Walton, J. S. Murray, R. Crosnoe, C. Muller, E. Tipton, B. Schneider, C. S. Hulleman, C. P. Hinojosa, D. Paunesku, C. Romero, K. Flint, A. Roberts, J. Trott, R. Iachan, J. Buontempo, S. M. Yang, C. M. Carvalho, P. R. Hahn, M. Gopalan, P. Mhatre, R. Ferguson, A. L. Duckworth, C. S. Dweck, A national experiment reveals where a growth mindset improves achievement. *Nature* **573**, 364–369 (2019). doi:10.1038/s41586-019-1466-y Medline

36. D. S. Yeager, J. M. Carroll, J. Buontempo, A. Cimpian, S. Woody, R. Crosnoe, C. Muller, J. Murray, P. Mhatre, N. Kersting, C. Hulleman, M. Kudym, M. Murphy, A. L. Duckworth, G. M. Walton, C. S. Dweck, Teacher mindsets help explain where a growth-mindset intervention does and doesn't work. *Psychol. Sci.* **33**, 18–32 (2022). doi:10.1177/09567976211028984 Medline

37. M. A. Kraft, Interpreting effect sizes of education interventions. *Educ. Res.* **49**, 241–253 (2020). doi:10.3102/0013189X20912798

38. S. Woody, C. M. Carvalho, J. S. Murray, Model interpretation through lower-dimensional posterior summarization. *J. Comput. Graph. Stat.* **30**, 144–161 (2021). doi:10.1080/10618600.2020.1796684

39. G. M. Walton, D. S. Yeager, Seed and soil: Psychological affordances in contexts help to explain where wise interventions succeed or fail. *Curr. Dir. Psychol. Sci.* **29**, 219–226 (2020). doi:10.1177/0963721420904453 Medline

40. N. M. Stephens, H. R. Markus, S. A. Fryberg, Social class disparities in health and education: Reducing inequality by applying a sociocultural self model of behavior. *Psychol. Rev.* **119**, 723–744 (2012). doi:10.1037/a0029028 Medline

41. N. A. Bowman, N. Denson, Institutional racial representation and equity gaps in college graduation. *J. Higher Educ.* **93**, 399–423 (2022). doi:10.1080/00221546.2021.1971487

42. D. M. Silverman, R. J. Rosario, I. A. Hernandez, M. Destin, The ongoing development of strength-based approaches to people who hold systemically marginalized identities. *Pers. Soc. Psychol. Rev.* 10.1177/10888683221145243 (2023). doi:10.1177/10888683221145243 Medline

43. K. M. Kroeper, A. C. Fried, M. C. Murphy, Towards fostering growth mindset classrooms: Identifying teaching behaviors that signal instructors' fixed and growth mindsets beliefs to students. *Soc. Psychol. Educ.* **25**, 371–398 (2022). doi:10.1007/s11218-022-09689-4

44. N. M. Stephens, M. G. Hamedani, M. Destin, Closing the social-class achievement gap: A difference-education intervention improves first-generation students' academic performance and all students' college transition. *Psychol. Sci.* **25**, 943–953 (2014). doi:10.1177/0956797613518349 Medline

45. E. N. Smith, D. S. Yeager, C. S. Dweck, G. M. Walton, An organizing framework for teaching practices that can "expand" the self and address social identity concerns. *Educ. Psychol. Rev.* **34**, 2197–2219 (2022). doi:10.1007/s10648-022-09715-z

46. J. P. Goyer, D. S. Yeager, G. M. Walton, C. Logel, M. C. Murphy, College Transition Collaborative, Does a Social-belonging Intervention Reduce the Effects of Social Identity Threat on Group-based Inequalities in Academic Outcomes? Results from a Large, Multi-site, Randomized Trial, Open Science Framework (2020); https://doi.org/10.17605/OSF.IO/ZT653.

47. M. J. Weiss, H. S. Bloom, T. Brock, A conceptual framework for studying the sources of variation in program effects. *J. Policy Anal. Manage.* **33**, 778–808 (2014). doi:10.1002/pam.21760

48. P. Tough, "Who gets to graduate?," *The New York Times*, 18 May 2014; https://www.nytimes.com/2014/05/18/magazine/who-gets-to-graduate.html.

49. E. R. Carter, S. T. Brady, L. A. Murdock-Perriera, M. K. Gilbertson, T. Ablorh, M. C. Murphy, The racial composition of students' friendship networks predicts perceptions of injustice and involvement in collective action. *J. Theor. Soc. Psychol.* **3**, 49–61 (2019). doi:10.1002/jts5.27

50. C. Logel, J. M. Le Forestier, E. B. Witherspoon, O. Fotuhi, A social-belonging intervention benefits higher weight students' weight stability and academic achievement. *Soc. Psychol. Personal. Sci.* **12**, 1048–1057 (2021). doi:10.1177/1948550620959236

51. J. LaCosse, E. A. Canning, N. A. Bowman, M. C. Murphy, C. Logel, A social-belonging intervention improves STEM outcomes for students who speak English as a second language. *Sci. Adv.* **6**, eabb6543 (2020). doi:10.1126/sciadv.abb6543 Medline

52. H. S. Bloom, S. W. Raudenbush, M. J. Weiss, K. Porter, Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *J. Res. Educ. Eff.* **10**, 817–842 (2017). doi:10.1080/19345747.2016.1264518

53. D. H. Bailey, G. J. Duncan, F. Cunha, B. R. Foorman, D. S. Yeager, Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. *Psychol. Sci. Public Interest* **21**, 55–97 (2020). doi:10.1177/1529100620915848 Medline

54. L. Hernández, L. Darling-Hammond, "Creating identity-safe schools and classrooms" (Learning Policy Institute, 2022); https://doi.org/10.54300/165.102.

55. D. M. Steele, B. Cohn-Vargas, *Identity Safe Classrooms, Grades K-5: Places to Belong and Learn* (Corwin Press, 2013).

56. B. Weiner, An attributional theory of achievement motivation and emotion. *Psychol. Rev.* **92**, 548–573 (1985). doi:10.1037/0033-295X.92.4.548 Medline

57. T. D. Wilson, M. Damiani, N. Shelton, "Improving the academic performance of college students with brief attributional interventions" in *Improving Academic Achievement*, J. Aronson, Ed. (Academic Press, 2002), pp. 89–108.

58. C. M. Steele, S. J. Spencer, J. Aronson, Contending with group image: The psychology of stereotype and social identity threat. *Adv. Exp. Soc. Psychol.* **34**, 379–440 (2002). doi:10.1016/S0065-2601(02)80009-0

59. C. M. Steele, J. Aronson, Stereotype threat and the intellectual test performance of African Americans. *J. Pers. Soc. Psychol.* **69**, 797–811 (1995). doi:10.1037/0022-3514.69.5.797 Medline

60. J. F. Dovidio, S. L. Gaertner, T. Saguy, Another view of "we": Majority and minority group perspectives on a common ingroup identity. *Eur. Rev. Soc. Psychol.* **18**, 296–330 (2007). doi:10.1080/10463280701726132

61. J. F. Dovidio, S. L. Gaertner, T. Saguy, Commonality and the complexity of "we": Social attitudes and social change. *Pers. Soc. Psychol. Rev.* **13**, 3–20 (2009). doi:10.1177/1088868308326751 Medline

62. A. Ben-Zeev, Y. Paluy, K. L. Milless, E. J. Goldstein, L. Wallace, L. Márquez-Magaña, K. Bibbins-Domingo, M. Estrada, 'Speaking truth' protects underrepresented minorities' intellectual performance and safety in STEM. *Educ. Sci. (Basel)* **7**, 65 (2017). doi:10.3390/educsci7020065 Medline

63. C. S. Rozek, G. Ramirez, R. D. Fine, S. L. Beilock, Reducing socioeconomic disparities in the STEM pipeline through student emotion regulation. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 1553–1558 (2019). doi:10.1073/pnas.1808589116 Medline

64. C. A. Bauer, R. Boemelburg, G. M. Walton, Resourceful actors, not weak victims: Reframing refugees' stigmatized identity enhances long-term academic engagement. *Psychol. Sci.* **32**, 1896–1906 (2021). doi:10.1177/09567976211028978 Medline

65. I. A. Hernandez, D. M. Silverman, M. Destin, From deficit to benefit: Highlighting lower-SES students' background-specific strengths reinforces their academic persistence. *J. Exp. Soc. Psychol.* **92**, 104080 (2021). doi:10.1016/j.jesp.2020.104080

66. D. M. Silverman, I. A. Hernandez, M. Destin, Educators' beliefs about students' socioeconomic backgrounds as a pathway for supporting motivation. *Pers. Soc. Psychol. Bull.* **49**, 215–232 (2023). doi:10.1177/01461672211061945 Medline

67. S. Bonilla, T. S. Dee, E. K. Penner, Ethnic studies increases longer-run academic engagement and attainment. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2026386118 (2021). doi:10.1073/pnas.2026386118 Medline

68. S. Cheryan, S. A. Ziegler, A. K. Montoya, L. Jiang, Why are some STEM fields more gender balanced than others? *Psychol. Bull.* **143**, 1–35 (2017). doi:10.1037/bul0000052 Medline

69. M. C. Murphy, C. M. Steele, J. J. Gross, Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychol. Sci.* **18**, 879–885 (2007). doi:10.1111/j.1467-9280.2007.01995.x Medline

70. C. S. Rozek, S. E. Gaither, Not quite White or Black: Biracial students' perceptions of threat and belonging across school contexts. *J. Early Adolesc.* **41**, 1308–1337 (2021). doi:10.1177/0272431620950476

71. A. K. Ho, J. Sidanius, D. T. Levin, M. R. Banaji, Evidence for hypodescent and racial hierarchy in the categorization and perception of biracial individuals. *J. Pers. Soc. Psychol.* **100**, 492–506 (2011). doi:10.1037/a0021562 Medline

72. F. J. Davis, *Who Is Black? One Nation's Definition* (Pennsylvania State Univ. Press, 1991).

73. S. O. Roberts, C. Bareket-Shavit, M. Wang, The souls of Black folk (and the weight of Black ancestry) in U.S. Black Americans' racial categorization. *J. Pers. Soc. Psychol.* **121**, 1–22 (2021). doi:10.1037/pspa0000228 Medline

74. K. Parker, J. M. Horowitz, R. Morin, M. H. Lopez, "Multiracial in America: proud, diverse, and growing in numbers" (Pew Research Center, 2015); https://www.pewresearch.org/social-trends/2015/06/11/multiracial-in-america/.

75. G. M. Walton, S. T. Brady, "The many questions of belonging" in *Handbook of Competence and Motivation: Theory and Application*, A. J. Elliot, C. S. Dweck, D. S. Yeager, Eds. (Guilford Press, ed. 2, 2017), pp. 272–293.

76. G. L. Cohen, J. Garcia, "I am us": Negative stereotypes as collective threats. *J. Pers. Soc. Psychol.* **89**, 566–582 (2005). doi:10.1037/0022-3514.89.4.566 Medline

77. C. Good, A. Rattan, C. S. Dweck, Why do women opt out? Sense of belonging and women's representation in mathematics. *J. Pers. Soc. Psychol.* **102**, 700–717 (2012). doi:10.1037/a0026659 Medline

78. J. P. Goyer, G. M. Walton, D. S. Yeager, The role of psychological factors and institutional channels in predicting the attainment of postsecondary goals. *Dev. Psychol.* **57**, 73–86 (2021). doi:10.1037/dev0001142 Medline

79. S. W. Raudenbush, A. S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods* (Sage Publications, 2002).

80. R. Chetty, J. N. Friedman, J. E. Rockoff, Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *Am. Econ. Rev.* **104**, 2633–2679 (2014). doi:10.1257/aer.104.9.2633

81. S. Goldrick-Rab, R. Kelchen, D. N. Harris, J. Benson, Reducing income inequality in educational attainment: Experimental evidence on the impact of financial aid on college completion. *Am. J. Sociol.* **121**, 1762–1817 (2016). doi:10.1086/685442

82. K. A. Bird, B. L. Castleman, J. T. Denning, J. Goodman, C. Lamberton, K. O. Rosinger, Nudging at scale: Experimental evidence from FAFSA completion campaigns. *J. Econ. Behav. Organ.* **183**, 105–128 (2021). doi:10.1016/j.jebo.2020.12.022

83. O. Gurantz, J. Howell, M. Hurwitz, C. Larson, M. Pender, B. White, "Realizing your college potential? Impacts of College Board's RYCP campaign on postsecondary enrollment" (EdWorkingPaper 19-40, Annenberg Institute at Brown University, 2019); https://doi.org/10.26300/nqn3-sp29.

84. V. Dorie, M. Harada, N. B. Carnegie, J. Hill, A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Stat. Med.* **35**, 3453–3470 (2016). doi:10.1002/sim.6973 Medline

85. E. Tipton, L. Fellers, S. Caverly, M. Vaden-Kiernan, G. Borman, K. Sullivan, V. Ruiz de Castilla, Site selection in experiments: An assessment of site recruitment and generalizability in two scale-up studies. *J. Res. Educ. Eff.* **9** (suppl. 1), 209–228 (2016). doi:10.1080/19345747.2015.1105895

86. E. Tipton, R. B. Olsen, A review of statistical methods for generalizing from evaluations of educational interventions. *Educ. Res.* **47**, 516–524 (2018). doi:10.3102/0013189X18781522

87. E. Tipton, How generalizable is your experiment? An index for comparing experimental samples and populations. *J. Educ. Behav. Stat.* **39**, 478–501 (2014). doi:10.3102/1076998614558486

88. *Profiles of American Colleges 2016* (Barron's Educational Series, ed. 32, 2015).