Values affirmation replication at the University of Illinois

Brianne Gutmann[1,*] and Tim Stelzer[2,†]

[1]*Texas State University, San Marcos, Texas 78666, USA*
[2]*University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*

(Received 17 December 2019; revised 6 March 2021; accepted 16 July 2021; published 20 September 2021)

Values affirmation exercises have been implemented in many contexts to combat stereotype threat in students from marginalized populations; the exercises are intended to fortify students by prompting them to self-affirm their values in short writing activities. Within the physics education research community, the style of intervention was underlined by a positive result from the University of Colorado Boulder; researchers were able to use the intervention to minimize the achievement gap between men and women in an introductory physics course. These results inspired a replication experiment in two physics courses at the University of Illinois in Urbana-Champaign, and this article provides some history and context of these interventions, describes our specific implementation, and reveals that we were unable to reproduce the positive results despite thorough attention to the details of the replication. Our findings suggest that the values affirmation exercises are not understood at a level where they should be considered a positive intervention to help marginalized populations.

DOI: 10.1103/PhysRevPhysEducRes.17.020121

I. INTRODUCTION

In the last decades, researchers have repeatedly seen the importance of students' attitudes and affect when engaging with physics material. Students' attitudes about physics and sense of belonging in the field are influenced by their identities and experiences around their identities within the field and beyond, particularly for students from marginalized populations [1–15]. These differences between majority and minority populations in physics can manifest in students' sense of self-efficacy and physics identity, which tend to persist from elementary school [10,16] to high school [17,18], introductory courses [1,2,7,10,19], upper-level courses [2], graduate school [2], and eventual careers in physics. In particular, students' sense of physics identity is a strong predictor of their persistence in the field [11,13,20–25]. Thus, attention to factors which affect students' confidence and self-efficacy is extremely relevant to their success.

In 2006, Cohen *et al.* piloted a tool, the values affirmation, as a way to counter the detrimental effects of negative experiences and expectations around students' minoritized identities [26]. Values affirmation exercises typically consist of a short writing activity about students' self-identified values, which is meant to protect minoritized students against threats to their identity and in turn bolster their self-efficacy and performance. The intervention was repeated specifically in the realm of physics education at the University of Colorado, Boulder in 2010, and the study received considerable attention for successfully reducing performance differences between men and women in an introductory physics course [27]. This paper documents an attempt to replicate Colorado's implementation at the University of Illinois at Urbana-Champaign, a similar institution in size, funding, and student demographics.

A. "Achievement gaps" in physics

Many studies that document student differences between majority and minority populations within academic fields cite performance in the form of achievement gaps with respect to gender and race [7,8,19,28]. This method of comparing groups' performance has been criticized or cited with caution by some education researchers [4,8,20,29–31]. Gutiérrez in particular cautions against studies that only document the existence of the gaps, asserting that these studies provide a static picture and do little to provide insight into addressing the complex problems [29]. Additionally, comparing majority and minority population students inherently reinforces a deficit model; differences between groups are framed as inadequacies in the minority populations [4,8,29] and assume that acting like the majority group is the goal. For example, Traxler *et al.* explain that gender gap analyses often explicitly or implicitly ask the question, "why can't women be more like

[*] brianne.gutmann@txstate.edu, she/her/hers
[†] tstelzer@illinois.edu, he/him/his

men?" [8]. Often comparisons are made by performance on standardized measures that have been validated on physics student populations which are primarily white[1] men; there has been little effort to validate these measures with attention to marginalized students [8]. As an alternative, Gloria Ladson-Billings offers a reframing of the "achievement gap" as an "education debt" to emphasize that differences between students' performance and self-efficacy are not faults of the minoritized students, but rather an accumulation of historical, economic, sociopolitical effects [31].

Furthermore, achievement gap research tends to erase the unique experiences of those at intersections of identities by treating all women or all students of color, for example, as homogeneous groups [8,11,29]. The theory of intersectionality, coined by Crenshaw, cites the damage of considering the consolidation of people along a single axis of identity by highlighting instances where the practice has hurt Black women [38]. She cites examples where initiatives and policies intended to help women ended up only serving white women, and policies intended to help Black people served to only help Black men [38]. In physics education, the majority of research is conducted at large primarily white four-year institutions within mostly male classes, and thus we are at risk of also oversimplifying and prescribing interventions which overlook more marginalized people [39]. Similarly, when considering gender gaps, researchers often treat gender as a purely binary variable [8], based on admissions information, which erases the experiences of gender nonbinary students and misrepresents transgender students who may be on record with the university as their wrong gender.

Alternately, Lubienski stresses that it is "irresponsible" not to investigate achievement gaps [40], which is highlighted by Traxler *et al.*, who cite that some thoughtfully conducted gap analyses have helped the education community understand mechanisms that contribute to performance differences [4]. Gutiérrez advises that it is more useful to focus on improving learning within marginalized populations without needing to compare them to majority students, and whatever benefit gap analyses bring come at a cost [29]. Thus, it is important to be mindful when citing and studying performance gaps.

B. Stereotype and identity threat

Of the dangers of "gap" literature, one is its oversimplification of many entangled contributions to individuals' experiences and performance. However, education research has attempted to isolate variables that affect students' performance and senses of efficacy within classrooms; there is strong research that documents the effects of identity threat and stereotype threat, over 300 studies in peer reviewed journals [41–47].

The effects of stereotype threat were named and introduced by Steele and Aronson in 1995 [48] in studies that saw Black middle school students scored below their ability on tasks when they perceived a threat to reinforce negative stereotypes. The studies saw that all participants scored similarly when the task was framed as a nondiagnostic study of problem solving, but the group that had the task framed as a diagnostic of intellectual ability saw differences in performance between white and Black students, which the authors attribute to the threat on Black participants to be concerned about stereotypes of intellect and reinforce those negative stereotypes [48]. In one of the studies, they asked participants to fill in letters to complete words; Black students in the group that had the task framed as a diagnostic intellectual task were more likely to complete the words into ideas related to negative stereotypes and self-doubt. For example, the prompt L A _ _ could be completed to LAZY, or F L _ _ _ to FLUNK [48]. The authors saw that these words appeared drastically more often for Black participants in the stereotype threat condition, but not in the nondiagnostic condition, which they use to justify their assertion that the threat was preoccupying cognitive resources [48].

These results have been replicated along many axes of stereotype in academic settings, including women on math tasks [49–53], students from low socioeconomic backgrounds [54,55], and Hispanic or Latino students [56–58]. The effects are not confined to academia, and have also affected memory tasks in the elderly [59], anxiety in gay men providing childcare [60], white men in sports [61], and women in negotiations [62] and driving [63]. In two meta-analyses which combine data from nearly 19 000 students, Walton and Spencer quantify the degree of underperformance by stereotyped students to conservatively be near 0.2 standard deviations, which they assert contributes a large part of the "gaps" seen in literature [42].

The mechanisms that drive stereotype threat are often cited as being cognitive. Increased anxiety and distraction by the threat can use up cognitive resources which would otherwise be focused on the performance tasks [42,43,48,64,65]. In particular, Krendl *et al.* used MRI scans to verify that women doing math in a stereotype condition had less activation in their brains in areas related to problem solving than women in the nonthreat condition, and saw higher brain activity in a region associated with processing negative information [66]. Likewise, Croizet

[1] The authors note that there is disagreement around respectful practices regarding the capitalization of "white" when referring to race [32–36]. Some journalists and scholars argue that not capitalizing the term (in contrast to other capitalized terms, such as Black, Hispanic, etc.) sets whiteness as an invisible default and allows white people to distance themselves from their racial identities. Others note that the capitalization of white has troubling connections to white nationalism and supremacy and do not wish to lend the term more power. In this paper, we follow the convention used by our colleagues [37] and do not capitalize the term.

et al. measured heart rate fluctuations associated with mental workload and also discovered students facing stereotype threat were experiencing higher mental workload which taxes students' limited working memory [67]. In a synthesis of research of stereotype threat related psychology, Steele *et al.* expand this explanation by citing the cyclic nature of stereotype threat; students who perform poorly are treated differently by teachers, adaptively adjust their self-image to reflect the failures, and different opportunities are available to them, which reinforce the stereotype and threat [43]. This can affect motivation and students' performance expectations [68–70], which in turn influences students' behavior.

Steele *et al.* cite a number of factors and strategies that can temper the effects of stereotype threat [64]. Many of the strategies are relational; students are better situated to buffer against stereotype threat if they have support systems to directly or indirectly counter stereotypes. Steele *et al.* argue that having friends within their identity group can reduce students' endorsement or anxiety around stereotypes simply by having more interactions with people who can contradict stereotypes and that the endorsement of stereotypes influences how students are affected by stereotype threat [64]. Graham *et al.* speak particularly to the threat to Black students in predominantly white institutions (PWIs); they found that students who had friend groups that were made of primarily Black students in high school showed more ease in transitioning academically and socially into college (particularly into a PWI), even controlling for socioeconomic status and neighborhood [71]. Likewise, seeing successful people of their identity can reduce the threat of stereotype by providing examples of success [64]. Mentors or teachers, regardless of the alignment of their identity to students can also reduce threat by valuing students' contributions and especially by speaking explicitly about stereotypes [64]. Even the act of explicitly citing an assessment as being neutral to stereotypes was shown to reduce stereotype threat for women on math tasks [49]. Within students themselves, the belief of malleability of intelligence such as Dweck's growth mindset, can reduce the impact of stereotype threat by rejecting "innate" qualities of intelligence associated with stereotypes [64,72–74].

C. Self and values affirmation history and theory

With the adverse affects of stereotype threat explicitly documented, many education researchers have prioritized energy towards addressing the threat. These efforts include social psychological interventions that target students' thoughts, feelings, and beliefs about themselves. Values affirmation exercises have received considerable attention, in part because the exercises have shown long lasting effects at minimizing gaps with only brief interventions. These interventions usually consist of short writing exercises where students identify values that are important to them and reflect on their personal importance.

Values affirmation exercises are a type of self-affirmation, a method described by Steele in which people respond to a threat to their integrity by fortifying another aspect of the self, which does not necessarily need to be related to the threatened domain [75]. He cites a study of housewives asked to help on a community project: some were told that it was known within the community that they were not generally "cooperative" in community projects, and these women were more likely to volunteer in a followup by a completely different person in a separate encounter. When their self-perception of themselves as helpful people was threatened, the women were more likely to seek out another venue to restore their sense of self and integrity [76]. In the same vein, students who are feeling threatened in their academic environment can bolster their sense of integrity by validating aspects of themselves that are valuable to them [77]. In their review of self-affirmation literature, Sherman and Cohen speak to this with direct application to an educational environment [78]:

> *When global perceptions of self-integrity are affirmed, otherwise threatening events or information lose their self-threatening capacity because the individual can view them within a broader, larger view of the self. People can thus focus not on the implications for self-integrity of a given threat or stressor, but on its informational value. When self-affirmed, individuals feel as though the task of proving their worth, both to themselves and to others, is "settled." As a consequence, they can focus on other salient demands in the situation beyond ego protection.*

Values affirmation exercises do not necessarily alleviate specific threats, but instead reduce the power of the threat to a person's adequacy [43,45,56], freeing up cognitive load to focus on the tasks at hand. Shnabel *et al.* examined different types of emphases in values affirmation exercises and discovered the strongest effects in students who were prompted to write specifically about social belonging. They prompted these reflections by asking students to write about experiences which made them "feel closer and more connected with people" [46] which aligns with Steele *et al.*'s emphasis on relational counters to stereotype threat [64].

Some of the earliest values affirmation experiments were conducted by Cohen *et al.* [26], who saw a reduction by 40% in the achievement gap for Black students compared to their white classmates. In their follow-up study, they tracked the same students over two years and saw that the effects resulted in higher GPA (0.24 grade points higher than students who did not receive the affirmation intervention) and a lower rate of remediation or grade repetition (5% compared to 18%) [79]. In the same way that stereotype threat can compound, they stress that values affirmations have the potential to persist recursively; students who feel valued may perform better, which affects how they see themselves and how they are treated, which can amplify the effect with each new validating event [79]. Values affirmation interventions are intended to disrupt the cycle of stereotype threat, either slowing its effects or setting in motion a positive counter cycle [43,45,79].

Because of their recursive nature, the benefits of values affirmation exercises are affected by their timing; earlier interventions can disrupt and bolster students against recurring interactions which may threaten their self-worth [43]. Sherman *et al.* suggest that administering interventions during times of transition are especially effective (such as students entering middle school, high school, or college) because the nature of these transitions often includes increased academic expectations, flux of identity, and disrupted support systems for students [56].

Additionally, Yeager and Walton notice that values affirmation exercises are most effective when students are unaware of their true intention [45]. By labeling interventions as "interventions," students are receiving a stigmatizing message that they are in need of help (which actually plays into stereotype threat). Instead, students who feel control over the activity are more able to take ownership of their success rather than attribute it to a "heavy-handed intervention," amplifying the recursive nature of growing self-efficacy [45]. Sherman *et al.* gave the additional rationale for subtlety in delivery: students who are aware that the intervention is meant to bolster self-worth will be using cognitive resources to consider the benefits of the intervention as a means to an end, rather than fully engaging in the activity and focusing inwards [80]. Their study explicitly measuring how awareness affects values affirmations showed empirically that awareness attenuates the positive effects of values affirmations [80].

D. Values affirmation experiments and replication

As mentioned previously, the earliest values affirmation experiments were conducted by Cohen *et al.* [26,79], which showed promising results in elevating Black middle school students' performance. Since then, others have implemented values affirmations exercises in their labs or classrooms with success. Sherman *et al.* found that a values affirmation intervention improved grades for Latino American students who participated in the writing activity [56]. Another study intending to replicate Cohen *et al.*'s studies was done in a predominantly Black and Latino middle school, with 80% of the students enrolled in a free lunch program. Values affirmation exercises were distributed to half the students through 24 homeroom teachers, and results showed that students who participated in the affirmation exercise outperformed students in the control condition. This study did not include a control for white, middle class students because the sample was too small, but instead focused on the ability of the affirmation to lift performance of stereotyped students [81] as suggested by Gutiérrez [29]. Values affirmation exercises were also successful for ethnic minority medical students [82], first generation students in biology [83], LGBTQ students (in conjunction with another activity) [84], and women in science [27,85].

Alongside successful replications, the literature also includes studies with mixed or null results [86]. When delivered to students in three St. Paul middle schools, values affirmation exercises did not statistically affect students' performance on a Minnesota state standardized test, except for women on their math scores. The researcher expected to see an improvement for racially minoritized students but was surprised to see no effect [87]. Similarly, when the intervention was given to students in six middle schools in the Philadelphia area, the author cites that the replication failed to have significant results, and actually hurt the performance of female students [88]. Another study across eleven middle schools saw some improvement, but cited that the results were much smaller than predicted by Cohen *et al.*'s results [89], which they later cite to possibly be due to poor implementation fidelity when the activity was scaled up [90]. In the college-level domain, an affirmation study in introductory biology courses marginally helped racially minoritized students but the researchers also saw that the intervention affected white women slightly negatively, which contradicts other studies [91]. In physics, biology, and biochemistry courses, a replication of a successful gender oriented values affirmation saw no difference between treatment and control conditions for men or women, though they attribute the null result to limited initial differences between men and women at their institution [92].

Most of these published works showed some mixed results, and it is possible that replications that saw purely null results were not published. Franco *et al.* followed 221 funded social science studies and discovered that the majority of null results are not even written up (over 60%), and those that are written are less likely to be accepted for publication than stronger results [93,94]. This publication bias has been persistently increasing across fields and across different countries [95], especially in the social sciences and biomedical research [95–97]. The exclusion of null results skews literature by selectively amplifying positive results and dismissing null results that contradict them [94,97,98]. In the extreme case when one is using p values of 0.05 as significance, if 5% of "chance" positive results are published and 95% of null results are not, our understanding is extremely skewed [97,98]. In fact, a synthesis of studies by Sterling *et al.* saw that over 95% of studies in biomedical and social science journal articles cited statistically significant results [99]; either the proportion of significant studies has increased or bias against nonsignificant results in publishing is driving these numbers up [98]. In the case of values affirmation interventions, the experiments are relatively quick, allowing for easy repetition and replication; statistically, more research groups attempting to replicate a result will increase the chances of statistical fluctuations that can be published as positive results.[2]

Within physics education, a notable positive result by Miyake *et al.* at the University of Colorado, Boulder showed statistically significant results in reducing the gender gap in introductory physics, in both grades and scores on the Force and Motion Conceptual Evaluation (FMCE) [27]. The result was published in 2010 in the journal *Science* and has been cited over 700 times at the time of this paper. A subset of the authors of the original paper failed to replicate their findings; they saw reduction in differences between men and women on exams but the interaction of the treatment and gender was not statistically significant and the results of the FMCE showed women performing worse with the affirmation treatment [101]. In contrast to the publicity of the first study, this contradictory replication has been cited on the order of 30 times. The original study is often cited in physics and STEM education literature without qualification or as a strong result to inform teaching strategies [6,102–110]. Some researchers acknowledge the mixed results [8,28,44,111,112] or caution about their context dependency [86,92], but simply by virtue of numbers, the first study is often cited without reference to the less successful replication. The importance of subtle contexts which affect the implementation of values affirmation studies are asserted by Bradley *et al.* [86], and Conlin *et al.* stress that the absence of attention to null results deprives the research community of opportunities to explore how these contexts can be refined [98]. As a large R1 university with high-enrollment physics classes (and thus, statistical power), the University of Illinois at Urbana-Champaign is well positioned to either replicate or define boundaries of values affirmation interventions.

II. METHODS

Our values affirmation replication was implemented in two introductory physics courses at the University of Illinois at Urbana-Champaign. The values affirmation activity was given to students in Physics 100, a preparatory mechanics course for the calculus-based physics sequence, and Physics 212, a calculus-based introductory electromagnetism course.

The preparatory course, Physics 100, is recommended to students in the engineering track who score below a threshold on a conceptual physics diagnostic test that they take before arriving at the university, and is intended to supplement students' different experiences with physics and math reasoning in high school. Students are not required to take Physics 100; the suggestion is made by their advisors, and students who score above the threshold are also allowed to enroll if they are concerned about their preparedness. The course is taught each fall with around 500 students, primarily first-term freshmen.

In contrast, Physics 212 is typically taken by engineering students in their second year, and the course typically has around 1200 students in the fall semesters. These students have already successfully completed Physics 211, the introductory mechanics course targeted by the preparatory course, meaning they are overall likely stronger students as they have the experience of completing at least one other physics course and are selectively the students who were able to succeed in that course. (The engineering physics introductory sequence population does suffer attrition as the courses get more advanced.) There is variation of experience within all course populations, but we expected these two courses to be enlightening populations to check the effects of values affirmation exercises.

The breakdown of self-identified student demographics in each of these courses is summarized in Table I, and both professors teaching Physics 100 and Physics 212 were white men. The reported student demographics are from the university itself, so students' options for self-identifying were limited to choices from the University of Illinois, which obscure some nuance in students' identities. For example, demographic information for international students does not allow for more specific designation than "international," nor does "multirace" give an option to elaborate on students' multiple racial identities. It is unclear whether international students are more likely to choose their racial identity or their identity as an international student as their demographic. Additionally, the gender options for self-reporting were limited to binary options, which forces nonbinary students to misrepresent themselves, and potentially misrepresents transgender students who may not be public about their gender.

Although the focus of the replication study is specifically on gender, we recognize that the racial and ethnic identities of the students may also interact with the intervention, and that the representations of students in the courses are relevant. Historically, values affirmations have been used to support students from minoritized populations, and Physics 100 tends to have greater representation of people of color and white women than Physics 212. Therefore, we felt it was important to acknowledge these differences, even though the majority of the analysis will be specifically focused on gender to replicate the analysis of the original study.

The timing and implementation of the values affirmation exercises at the University of Illinois mirrored the implementation of the exercises at the University of Colorado. The first author was in contact with Professor Miyake, the first author of the Colorado affirmation paper, and received an extensive implementation guide for replication which

[2] Since this paper was originally submitted, Wu *et al.* published a meta-analysis of values affirmation experiments, which also included several published null results [100].

TABLE I. Percentage of self-identified demographics for students in Physics 100 and Physics 212. The percentages are out of 399 enrollments for Physics 100 and 1012 enrollments for Physics 212. (NHPI stands for Native Hawaiian and Pacific Islander. American Indian and Alaska Native (AIAN) was also given as an option for students, but no students in either course selected this identity).

Physics 100	M	F		Physics 212	M	F	
Asian	12.3%	10.0%	22.3%	Asian	18.4%	6.5%	24.9%
Black or African American	3.8%	1.5%	5.3%	Black or African American	0.8%	0.0%	0.8%
Hispanic	12.5%	2.0%	14.5%	Hispanic	6.2%	1.5%	7.7%
NHPI	0.3%	0.3%	0.5%	NHPI	0.1%	0.1%	0.2%
White	25.6%	18.3%	43.9%	White	29.1%	12.8%	41.8%
Multirace	3.3%	2.5%	5.8%	Multirace	2.8%	1.2%	4.0%
International	3.3%	3.0%	6.3%	International	15.7%	4.0%	19.7%
Unknown	1.3%	0.3%	1.5%	Unknown	0.8%	0.2%	1.0%
	62.2%	37.8%			73.8%	26.2%	

TABLE II. Activities associated with values affirmation replication: Timing, context of implementation, and notes for clarification.

Timing	Activity	Context	Notes
Week 0	Professor primes students to expect activity	Lecture	Professor emphasizes the importance of science communication and tells students to expect writing exercise in discussion.
Week 1	First values affirmation writing activity	Discussion section	Activity and TA scripts identical to CO implementation. Consent form adapted from CO implementation.
Week 2	Attitudes survey to test stereotype endorsement	Online checkpoint	Questions to test stereotype endorsement embedded in the CLASS. Full CLASS and 3 questions to test endorsement for Physics 100. A shortened version of the CLASS and 2/3 of endorsement questions for Physics 212.
Week prior to Exam 1	Second values affirmation writing activity	Online homework	Activity identical to CO implementation, delivered online. Week 10 for Physics 100. Week 4 for Physics 212.
Exam week	First exam	Proctored exam	Week 11 for Physics 100. Week 5 for Physics 212.

was followed as closely as possible. The timing and activities are outlined in Table II, and each activity will be elaborated in the following sections.

A. Writing activity

To maintain the "stealth" aspect that has historically improved the effectiveness of values affirmation exercises [80], the values affirmation exercises were emphasized as being "writing" exercises. The exercise is first introduced in students' first lecture and is stressed as being important for science communication, as recommended by the implementation guide.

The writing exercise itself was given to students in their first discussion (recitation) section by their teaching assistants (TAs). Prior to the first week, the first author met with the TAs of the two courses to distribute the exercises, which were paper and pencil packets in plain envelopes, one for each student. The TAs were given the number of envelopes needed for their sections, which included a random sampling of control and treatment condition exercises that could be handed out randomly. TAs were not told the nature of the experiment, but were given a script with answers to potential student questions; most importantly, they were told to ensure students knew that the activity was not graded (to reduce stress that could compromise the activity) and that people associated with the course (themselves and the professor) would not read their responses. The teaching assistants were not told the true nature of the experiment to reduce the chance that their distribution and responses to students would unintentionally give students clues to its purpose. This was suggested by the implementation guide from Colorado and the script was identical to the one used in their experiment. The decision to have TAs distribute the exercise was also suggested, as it reduces emphasis that the activity is an experiment and allows students to focus on the activity. As required by our Institutional Review Board, our consent forms attached to the activity were slightly more specific about the research aspect than the provided template, but also did not hint at the goal of the activity to measure or influence gender differences.

Students were given 15 min to complete either a control or treatment writing exercise, which had three parts:

Part One: Students were given a list of personal values and asked to circle two or three that are *most* (treatment condition) or *least* (control condition) valuable to them. The options listed were identical to the Colorado implementation [27] and Cohen *et al.*'s original implementation [26]. The listed values were

–Being good at art
–Learning and gaining knowledge
–Relationships with family and friends
–Government or politics
–Independence
–Creativity

–Athletic ability
–Music
–Belonging to a social group (such as your community, racial group, or school club)
–Spiritual or religious values
–Sense of humor
–Career

Part Two: Students in the treatment condition were asked to think about the values and a time that the values were or would be important to them. Students in the control condition were asked to think about why the values might be important to someone else. All students were encouraged to focus on their thoughts, rather than grammar, spelling, or quality of writing, and write a few sentences about their selected values' importance.

Part Three: Students were asked to look over their values chosen and reasons, then list each of their values again with the top two reasons the values are important to them (treatment) or the top two reasons they could be important to someone else (control).

The envelopes were collected by TAs and returned to the researcher. An outside researcher tabulated each student's condition (treatment or control) and consent given, which was used to generate a list of email addresses to ensure that students would receive the same condition on the second iteration of the writing activity.

B. Stereotype endorsement

About a week after the writing activity, students completed an online learning attitudes assessment with questions embedded to gauge their endorsement of the stereotype that men do better than women in physics. This needed to be done after the affirmations activity but as early as possible to avoid activating stereotype threat immediately before the exam. The learning attitudes assessment used was the Colorado Learning Attitudes about Science Survey (CLASS) [113], but the assessment was primarily a vehicle to distract from the questions about gender. All the items in the survey are given as agreement with statements on a 5-point Likert scale between "strongly disagree" and "strongly agree." The stereotype endorsement statements embedded into the CLASS are listed as follows:

1. According to my own personal beliefs, I expect men to generally do better in physics than women.
2. According to general beliefs in society, men are expected to be better at physics than women.
3. I think my physics teachers expect women to do better than men in physics.

All 42 questions on the CLASS plus the three additional endorsement statements were given to Physics 100 students, and a subset of the CLASS (18 statements) and two stereotype endorsement statements were given to Physics 212. The modified version was given to Physics 212 students because of length concerns by the professor of the course. The study by Mikake *et al.* [27] found question 1 (about personal beliefs) to be most useful for their

TABLE III. Number of male (M) and female (F) participants in each condition, by course, for values affirmation experiment.

	Physics 100 ($N = 389$)			Physics 212 ($N = 1012$)		
	Treatment	Control	Sensitivity	Treatment	Control	Sensitivity
M	121	120	0.36	366	381	0.21
F	75	73	0.46	143	122	0.34

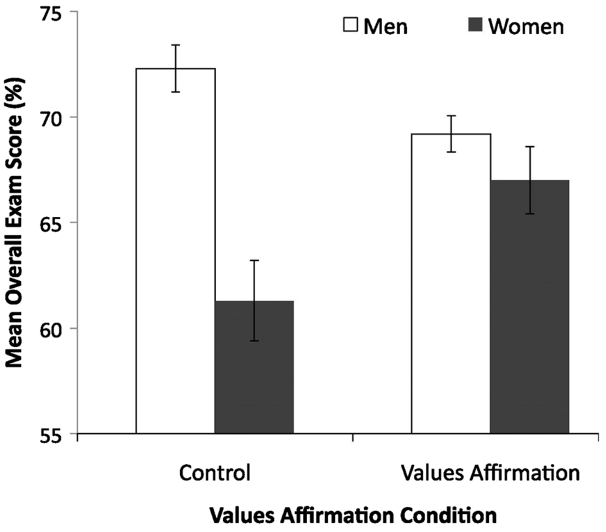
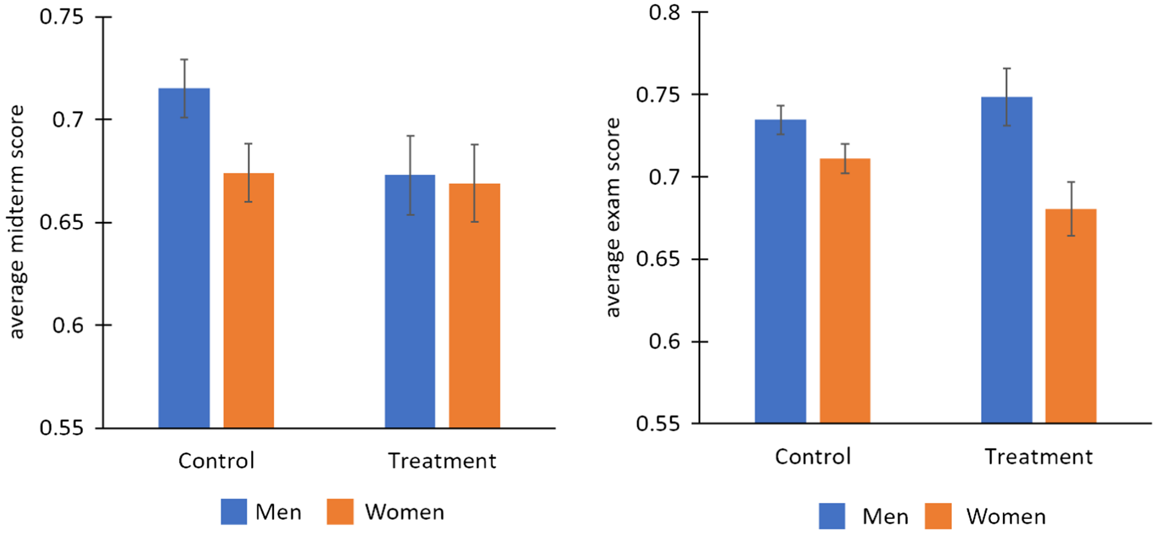


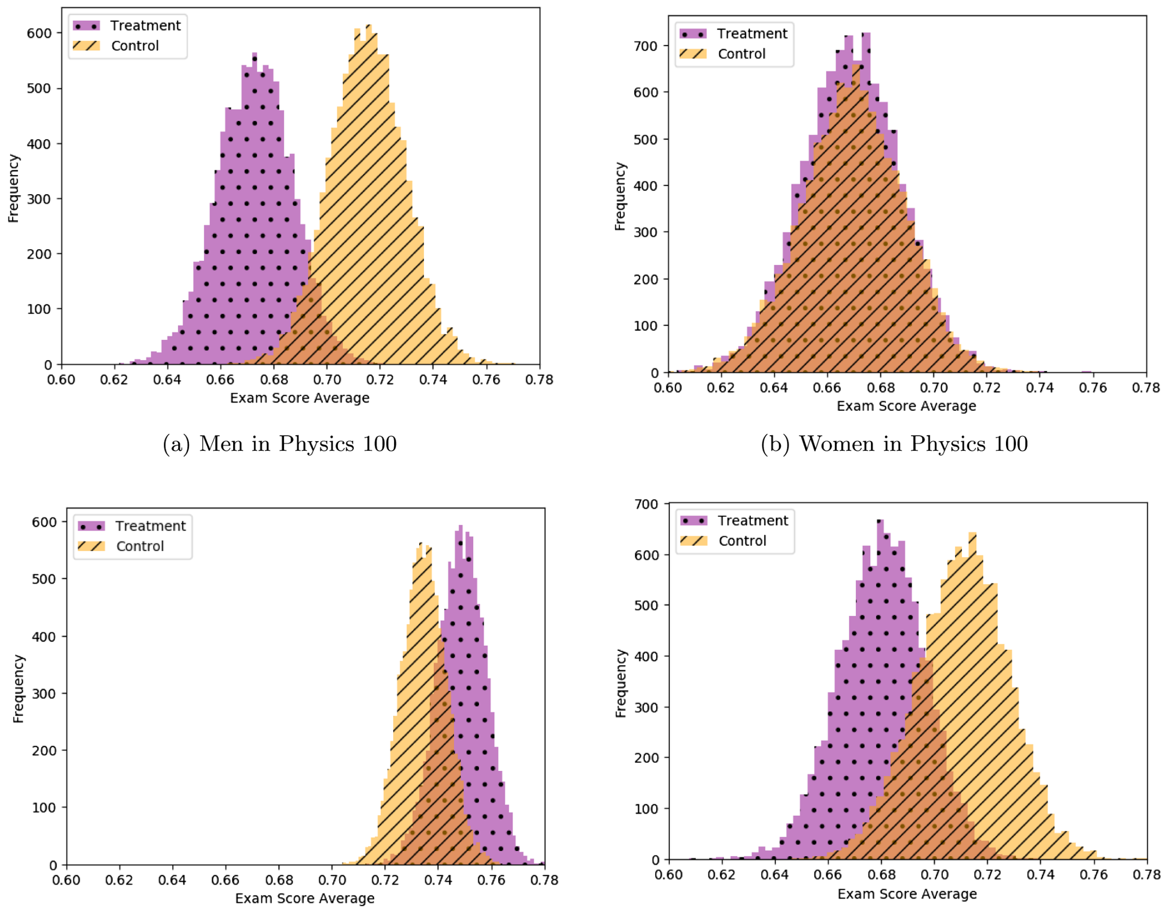
FIG. 1. Adjusted exam scores for students in the Miyake *et al.* study, by gender and treatment condition [27]. Note that the vertical axis begins at 55%. Reprinted with permission from AAAS.



(a) Physics 100: Midterm Exam Score

(b) Physics 212: Hour Exam 1 Score

FIG. 2. Average exam scores for men and women in treatment (values affirmation) and control conditions for the first exam in (a) Physics 100 and (b) Physics 212. Note that the vertical axis begins at 0.55.



(a) Men in Physics 100

(b) Women in Physics 100

(c) Men in Physics 212

(d) Women in Physics 212

FIG. 3. Histogram of bootstrapped exam average ranges for men and women in treatment (values affirmation) and control conditions in Physics 100 and 212. In all of these plots, the control condition is light orange with diagonal hatching up and to the right (//) and the treatment condition is magenta with spots; dark orange with both patterns shows their overlap.

analysis, so the first two endorsement questions were the ones included in the subset survey. Students received points for completing the learning attitudes survey but not for any specific answers.

C. Writing activity repetition and exam

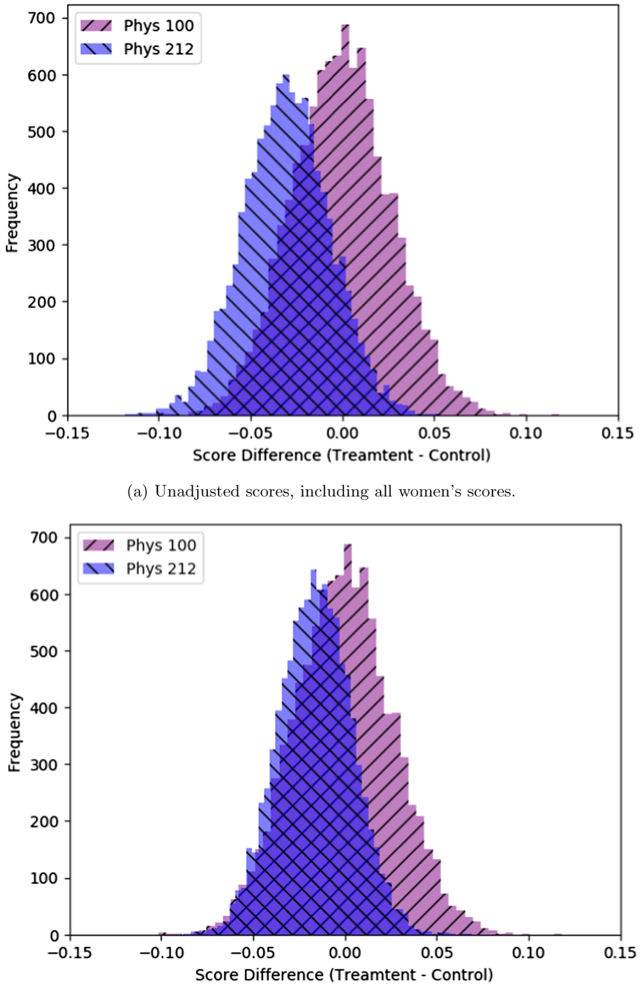
The second iteration of the writing activity was done online, the same as in the Colorado implementation. Two online forms were created with the same parts as in the paper and pencil implementation, one for control (asking students to reflect on values which are not important to them) and one for the treatment (asking students to reflect on their own values). The forms were sent to participants to match their condition in the first implementation; students received points for completing the activity, but not for any specific answers. Students who did not participate in the first activity (who enrolled late or missed the first discussion section) were randomly assigned one of the two conditions.

The timing of the second implementation was one week before each course's first exam. Since Physics 212 has three exams and a final, this came early, in week 4, for Physics 212 students. Physics 100 has only a midterm before the final exam, and their second writing activity was in their homework in week 10. The corresponding exams were given in week 5 and week 11, respectively.

D. Intended analysis

The analysis done by Colorado included students' exam grades, students' FMCE scores, final grades in the course, and both exam grades and FMCE scores as a function of high or low endorsement of gender stereotypes in physics, as measured by embedded questions on the CLASS. In our analysis, we will repeat each of these measures except for FMCE scores, which were not included in our courses. The FMCE would only be relevant for content in Physics 100, not for Physics 212, so it was not included.

The rest of the analysis was repeated, and the exam scores as a whole and by gender stereotype endorsement were additionally refined using bootstrap analyses [114]. The bootstrap analyses is a useful technique for better understanding and visualizing the statistical sensitivity of the collected data. To the extent that the student performance distributions are normal, the bootstrap analyses should be consistent with standard analysis techniques. Each bootstrap analysis calculated the size of the improved score (treatment condition–control condition) for each group of students simulating 10 000 trials. From the bootstrap, we calculated confidence intervals to give boundaries to the effect of the intervention.



(a) Unadjusted scores, including all women's scores.

(b) Unadjusted scores for Physics 100, and adjusted scores for Physics 212. The bootstrapped data set for Physics 212 includes scores from the subset of women who also took the previous physics course at the University of Illinois: 109/143 women in the treatment condition, and 85/122 women in the control condition.

FIG. 4. Histogram of bootstrapped difference in exam scores (treatment–control average) for women due to the affirmation activity, from (a) unadjusted scores and (b) adjusted scores in Physics 212. The Physics 100 histogram is shown in reddish purple with diagonal hatching up and to the right (//) and the Physics 212 histogram is bluish purple with diagonal hatching down and to the right (\\); dark purple with cross hatching shows their overlap.

III. RESULTS

Tabulating responses after the writing activity, we had 389 participants from Physics 100 and 1012 participants from Physics 212, which are reflective of the different sizes of the courses. Randomly distributing exercises effectively split participants into treatment and control conditions of comparable size; the breakdown of the number of students in each condition by class is given in Table III. For comparison, the first Colorado implementation was done with 439 students and its second implementation included 363 students. In both of these cases, these participants were a subset of the course (73% and 60%, respectively) [27,101]. Our samples included only students who gave consent and completed both exercises, so consist of a subset as well; 76% of Physics 100 and 92% of Physics 212

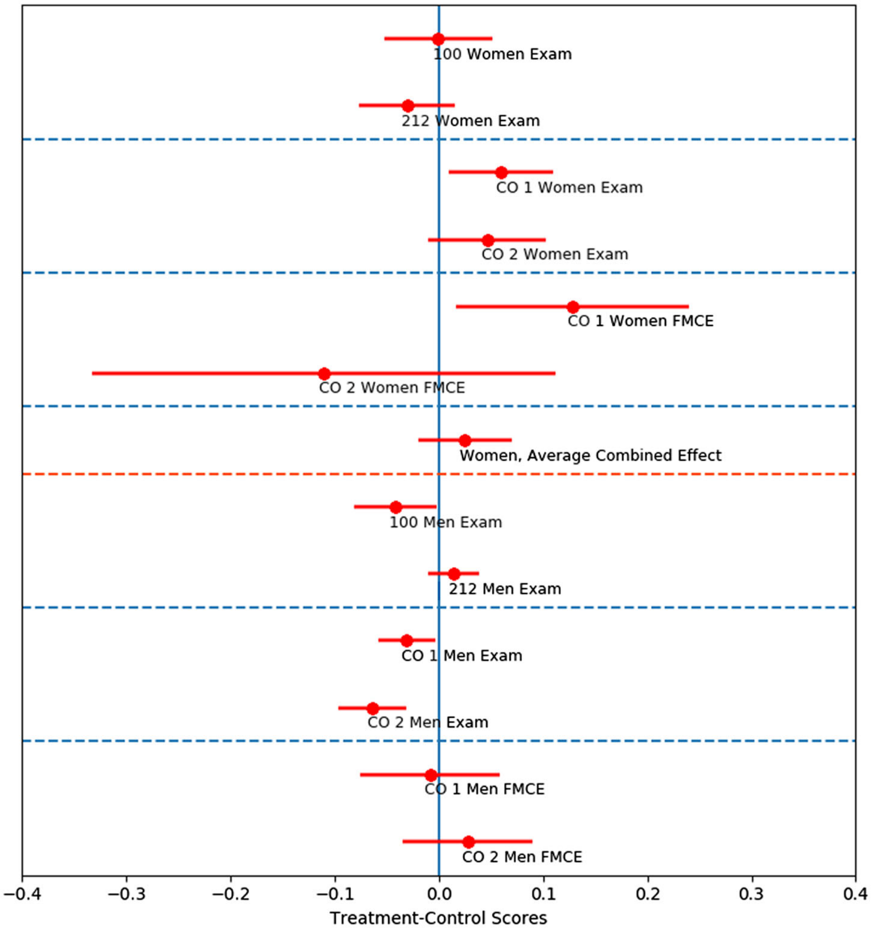


FIG. 5. Summary of effects within each gender and treatment. Points indicate mean value of treatment–control values and error bars are 95% confidence intervals. CO 1 refers to the first Colorado experiment [27] and CO 2 refers to the second implementation by Colorado [101]. The values for Colorado are adjusted scores, and their error bars are 95% confidence intervals, calculated from their reported standard errors. Note that within each experiment from Colorado, the exam and FMCE score effects are not independent; each pair of scores (exam and FMCE) is from the same group of students. The combined value for women's gains was calculated with a random effects model.

are included in analysis. Only a small fraction of consenting students were excluded from the data due to attrition within their courses (5% of Physics 100 and 2% of Physics 212).

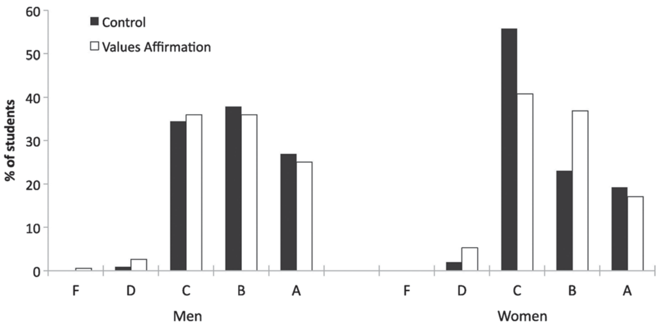


FIG. 6. Final grades of students by gender and treatment or control condition in the original Colorado implementation [27]. Reprinted with permission from AAAS.

The sensitivity listed in Table III was calculated using G*Power and represents the required effect size in order to have an 80% probability of making an observation with $p \leq 0.05$, using a two-tailed t test comparing the difference between the means of the control and treatment group. For comparison, the effect size of women's improved exam scores in the Miyake *et al.* study [27] was 0.45.

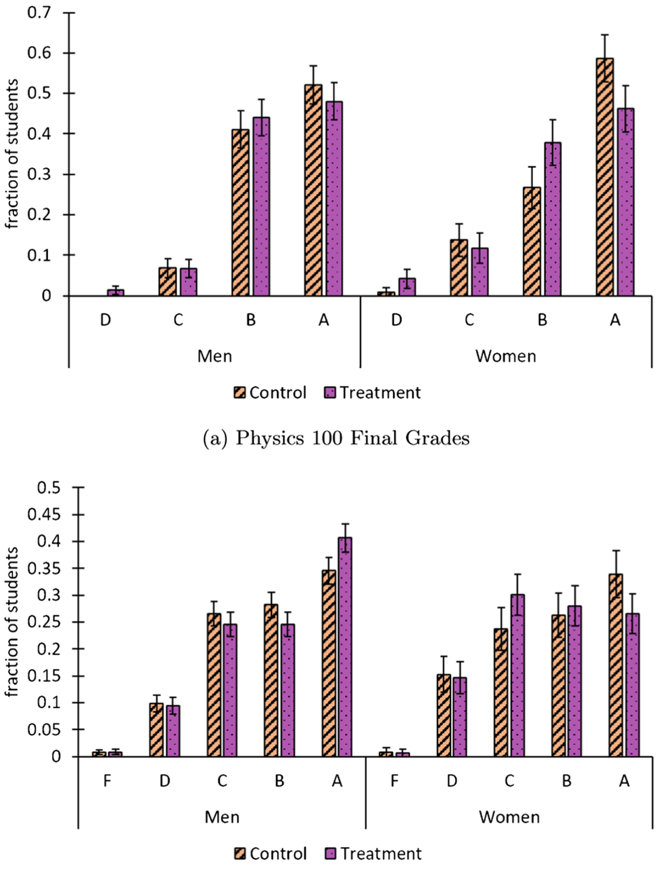
A. Exam results

1. Raw data

The first result presented by the Miyake *et al.* paper showed a significant reduction in gender differences for students in the treatment condition, which was presented as a bar chart and is shown for reference in Fig. 1 [27]. Their mean exam scores used were adjusted to account for baseline ACT/SAT math scores, which increased the difference between raw scores of women in the treatment and control groups. In contrast, we did not adjust the scores in our analysis, except where specifically noted, and focused on students' performance on their first exam immediately after the second treatment. Our unadjusted first exam scores for students in each group are shown in Fig. 2. We saw virtually no difference in women's scores in Physics 100, regardless of treatment, and lower performance by men in the treatment group. These changes slightly reduced the gender gap by about 4%, but did not elevate women's scores and the change was entirely due to the decline in men's scores with the treatment. In Physics 212, women who received the treatment did significantly worse and men with the treatment did slightly better, which increased the gap to about 7% for the treatment condition compared to 2% in control. Both of these cases illustrate the danger in considering gap sizes; the gap in Physics 100 is reduced despite no difference in the women's scores, and men's performance in Physics 212 inflates the gap, again, despite having no bearing on how the treatment affected women.

2. Bootstrap analysis

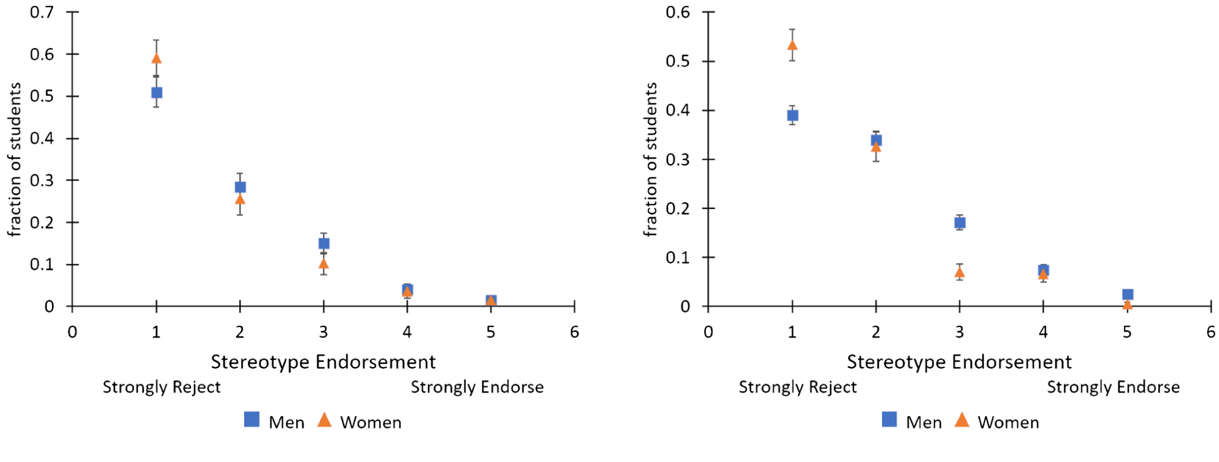
Students' scores were bootstrapped by randomly and proportionally selecting students from within each condition (gender × treatment) to simulate 10 000 trials with our data. Histograms of the distributions of bootstrapped average exam scores are shown in Fig. 3, with a focus on differences between conditions within groups. A quick look at these plots shows us that the only group that saw improved exam performance due to the treatment was men in Physics 212. Women in Physics 100 saw effectively no change from the affirmation exercise, while men in Physics 100 and women in Physics 212 saw a negative effect from the treatment.



(a) Physics 100 Final Grades

(b) Physics 212 Final Grades

FIG. 7. Distributions of students' final grades within each gender and treatment condition for (a) Physics 100 and (b) Physics 212. Each of these fractions is calculated out of the total number of students in each category (gender × condition).



(a) Physics 100 Stereotype Endorsement

(b) Physics 212 Stereotype Endorsement

FIG. 8. Fraction of men and women responding for each level of endorsement to the statement, "According to my own personal beliefs, I expect men to generally do better in physics than women." Students selecting 1 strongly disagree with the statement (a rejection of the stereotype), while students selecting 5 strongly agree with the statement (an endorsement of the stereotype).

According to theory, the affirmation activity should primarily help women, so a plot summarizing the bootstrapped effects for women in each course is in Fig. 4. This plot includes both raw scores and adjusted scores (for a subset of students in Physics 212).[3] We are primarily interested in the effects within gender, rather than the comparison of men to women; this plot shows contradictory effects in the two courses, whether the scores are adjusted or not. The mean value of the bootstrap distribution for Physics 100 shows a difference of $-0.05 \pm 2.7\%$, centered essentially at zero difference between the treatment and control groups. For Physics 212, the mean of the unadjusted distribution is $-3.0 \pm 2.4\%$, which becomes $-1.6 \pm 2.0\%$ for the adjusted distribution, both measurements showing women performing worse with the affirmation.

Within each gender group and course, a calculation of the average raw score gains for students with the treatment over students with the control condition are shown in Fig. 5 with a 95% confidence interval from the bootstrap as error. For completeness and comparison, this plot also includes the adjusted scores reported by Colorado in its two implementations on exams and FMCE gains, with 95% confidence intervals calculated from their cited standard error [27,101]. Note that the exam and FMCE score effects within each Colorado experiment are not independent; the results from "CO 1 Women Exam," for example, are from the same group of students as "CO 1 Women FMCE," although the two assessments are different.

B. Final grades

The researchers at Colorado also investigated the effect of the values affirmation exercises on students' final grades and saw an increase in the number of A's, B's, and reduction in number of C's for women in the treatment group compared to women in the control group [27]. For reference, their plot is shown in Fig. 6. The same analysis on our population showed the opposite effect; both courses saw a larger percentage of women in the control condition earning As than in the treatment, though these were the only differences in distribution that neared significance. These distributions are shown in Fig. 7; the values were not bootstrapped as the Colorado paper did not include error and could not be compared.

C. Stereotype endorsement effects

1. Raw data

A major result from the Colorado paper we attempted to replicate was the increased gains for women who endorse gender stereotypes in physics compared to those with a low endorsement [27]. For our analysis, we used the same Likert scale agreement survey statement: "According to my own personal beliefs, I expect men to generally do better in physics than women." In the following plots, students answering 1 (one) corresponds to strongly disagree, a strong rejection of the stereotype, through 5 (five) corresponding to strongly agree, a strong endorsement of the stereotype. Students also had the option to choose "no response," which was selected by about 1% of students in each course, and those students are not included in the following analysis. To get a snapshot of what students' level of endorsement is, a scatter plot of the fraction of students answering each choice from 1 (strong rejection) through 5 (strong endorsement) is given in Fig. 8.

TABLE IV. Number of male (M) and female (F) participants in each condition, by course and stereotype endorsement. Note that students who selected 2: disagree, no response, and those who did not fill out the survey are not included in this analysis.

		Physics 100		Physics 212	
		Treatment	Control	Treatment	Control
Low endorsing	M	43	59	127	119
	F	40	41	73	56
High endorsing	M	22	19	83	88
	F	9	12	18	16

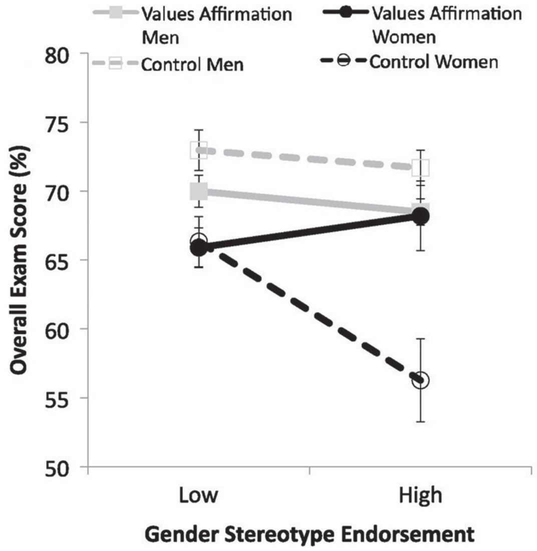
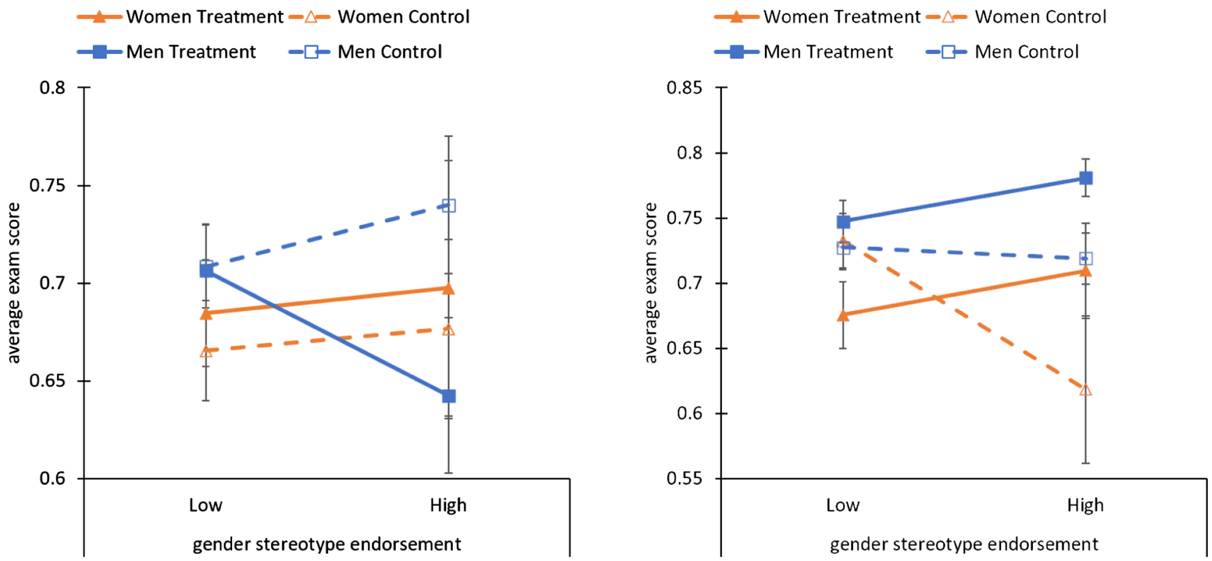


FIG. 9. Students' exam scores by treatment and gender group, also partitioned by high and low stereotype endorsement, for the first Colorado implementation [27]. Reprinted with permission from AAAS.

To classify our students as high and low endorsers of the stereotype, we used the same definitions as Miyake *et al.* [27] used in their plots: low endorsers were students who answered 0.75 times the standard deviation below the

[3] The subset of students in the adjusted set were Physics 212 students who completed the previous course in the physics sequence at the University of Illinois, which allowed us to subtract out their "baseline" exam score from the previous course.



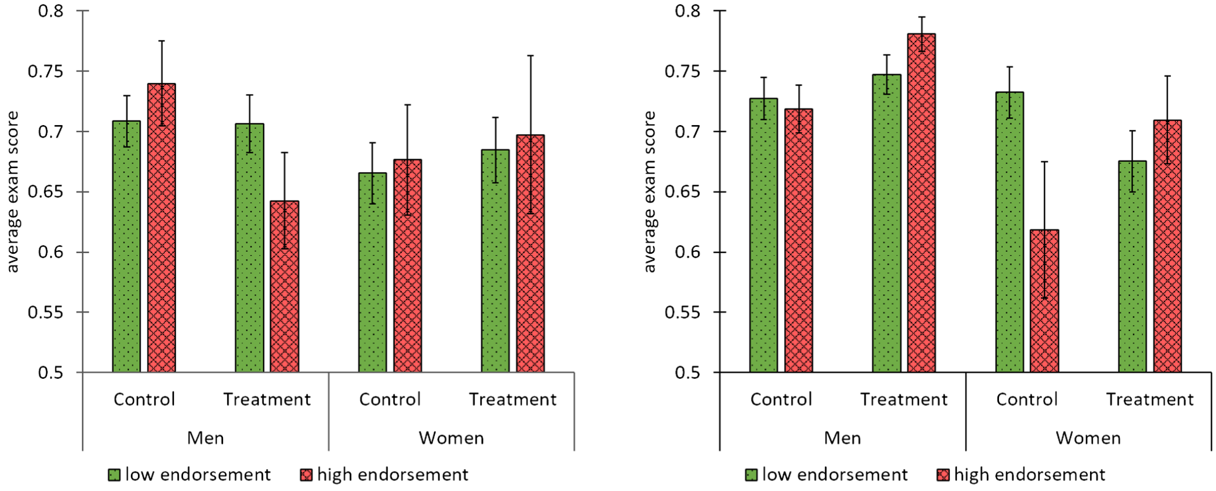
(a) Physics 100 Endorsement Effects

(b) Physics 212 Endorsement Effects

FIG. 10. Average exam grades for men and women in each treatment group, based on their classification of being a high or low endorser of gender stereotype in physics.

mean, and high endorsers were students who answered 0.75 times the standard deviations above the mean, treating students' responses as a continuous variable. In Physics 100, the mean value of students' responses was 1.71, with a standard deviation of 0.94. The mean value for Physics 212 responders was 1.92 with a standard deviation of 1.0. In both courses, the cutoff for "high" endorsers was any response of 3 (neutral) or higher, and only responses of 1 corresponded to "low" endorsers. Table IV gives the number of students in each condition using this categorization; these students are a subset of the original population as students who responded "2: Disagree" and no response are not included in either group, nor are the students who did not complete the survey. In both courses, most students responded to the survey (95% of Physics 100 students and 89% of Physics 212 students). We also include full regression analysis of all students (using their endorsement value as a continuous variable) in Appendix.

From their classification, Miyake *et al.* saw dramatic results, shown in Fig. 9. Our results, presented in the same way, are given in Fig. 10. In Physics 100, we see again that changes in the gap are predominantly due to men in the treatment group underperforming men in the control group. We do see higher scores for high endorsing women with the



(a) Physics 100 Exam Scores

(b) Physics 212 Exam Scores

FIG. 11. Average exam grades for men and women in each treatment group, based on their classification of being a high or low endorser of gender stereotype in physics.

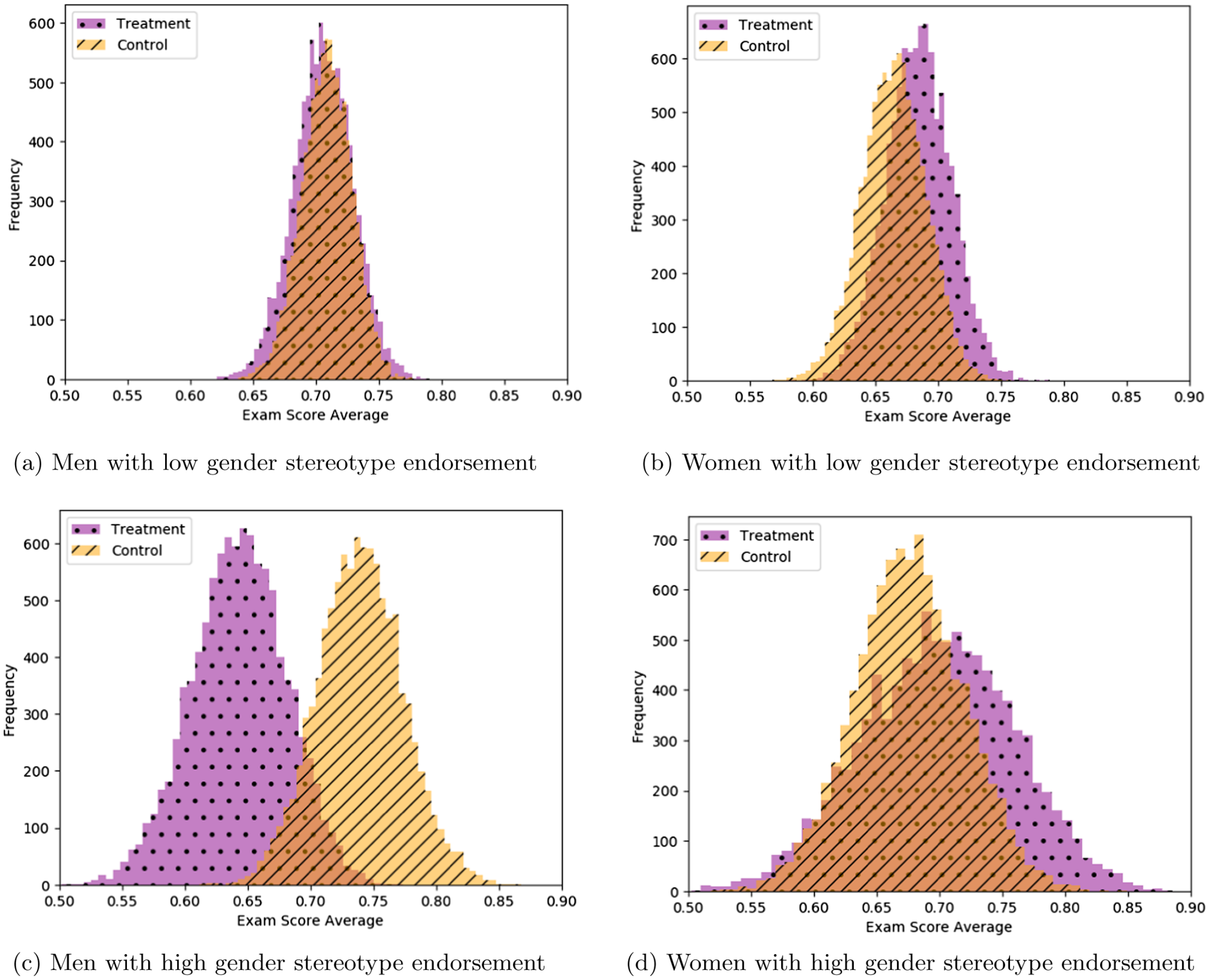
affirmation treatment in Physics 212 accompanied by lower scores for low endorsing women with treatment. However, these differences are within error bars and the representation of these groups as progressions can be misleading; each point is a separate population of students and the format was repeated for clean replication. Another comparison lies in looking at differences within like groups.

The same results are plotted as a bar chart in Fig. 11 which allows the reader to compare students with high or low endorsement within each group. The only statistically significant difference between high and low endorsing students is seen in women in the control group for Physics 212; women in the control group who endorse the stereotype did 10% worse on average on their exam than women who reject the stereotype. This difference between high and low endorsing women was not present in those who participated in the treatment, which could suggest that the treatment is protecting the high-endorsing women. However, the difference between high endorsing women in the control and treatment groups is still not statistically significant.

2. Bootstrap analysis

Students' scores were again bootstrapped by randomly selecting students from within each condition (gender× treatment × endorsement) to simulate 10 000 trials. The distributions of average exam scores for each subgroup, comparing average exam scores of students with treatment and control condition and separated by gender and stereotype endorsement are shown in Fig. 12 (Physics 100) and Fig. 13 (Physics 212).

Like in the previous bootstrap analysis, these groups can be subtracted from each other to see the effects of the affirmation treatment. Each of the plots in Fig. 14 show the control students' average scores subtracted from the treatment students' scores, essentially the gains by students who participated in the affirmation activity compared to their control counterpart group, and each subplot allows



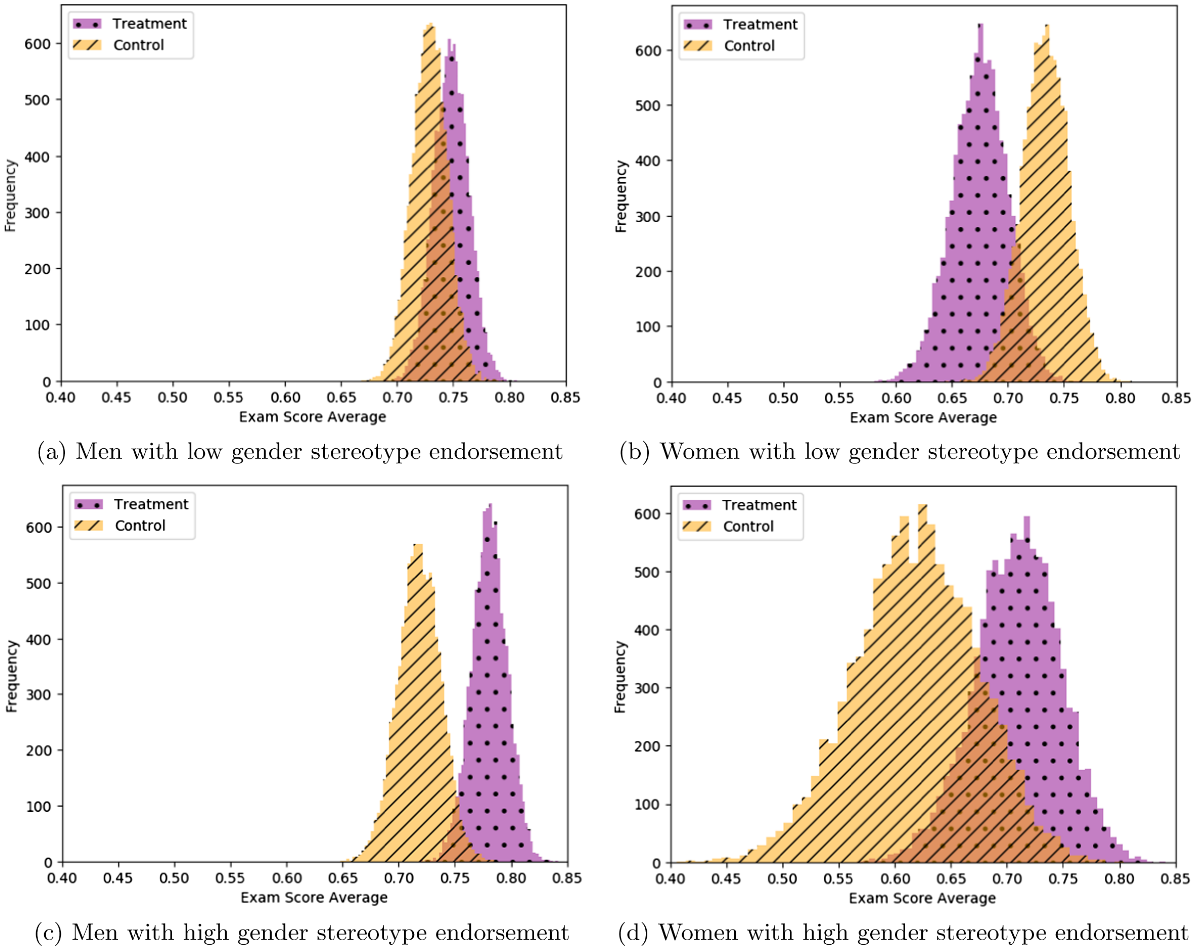
(a) Men with low gender stereotype endorsement

(b) Women with low gender stereotype endorsement

(c) Men with high gender stereotype endorsement

(d) Women with high gender stereotype endorsement

FIG. 12. Histogram of bootstrapped exam average ranges in Physics 100 for men and women in treatment (values affirmation) and control conditions with low and high endorsement of gender stereotype in physics. In all of these plots, the control condition is light orange with diagonal hatching up and to the right (//) and the treatment condition is magenta with spots; dark orange with both patterns shows their overlap.



(a) Men with low gender stereotype endorsement

(b) Women with low gender stereotype endorsement

(c) Men with high gender stereotype endorsement

(d) Women with high gender stereotype endorsement

FIG. 13. Histogram of bootstrapped exam average ranges in Physics 212 for men and women in treatment (values affirmation) and control conditions with low and high endorsement of gender stereotype in physics. In all of these plots, the control condition is light orange with diagonal hatching up and to the right (//) and the treatment condition is magenta with spots; dark orange with both patterns shows their overlap.

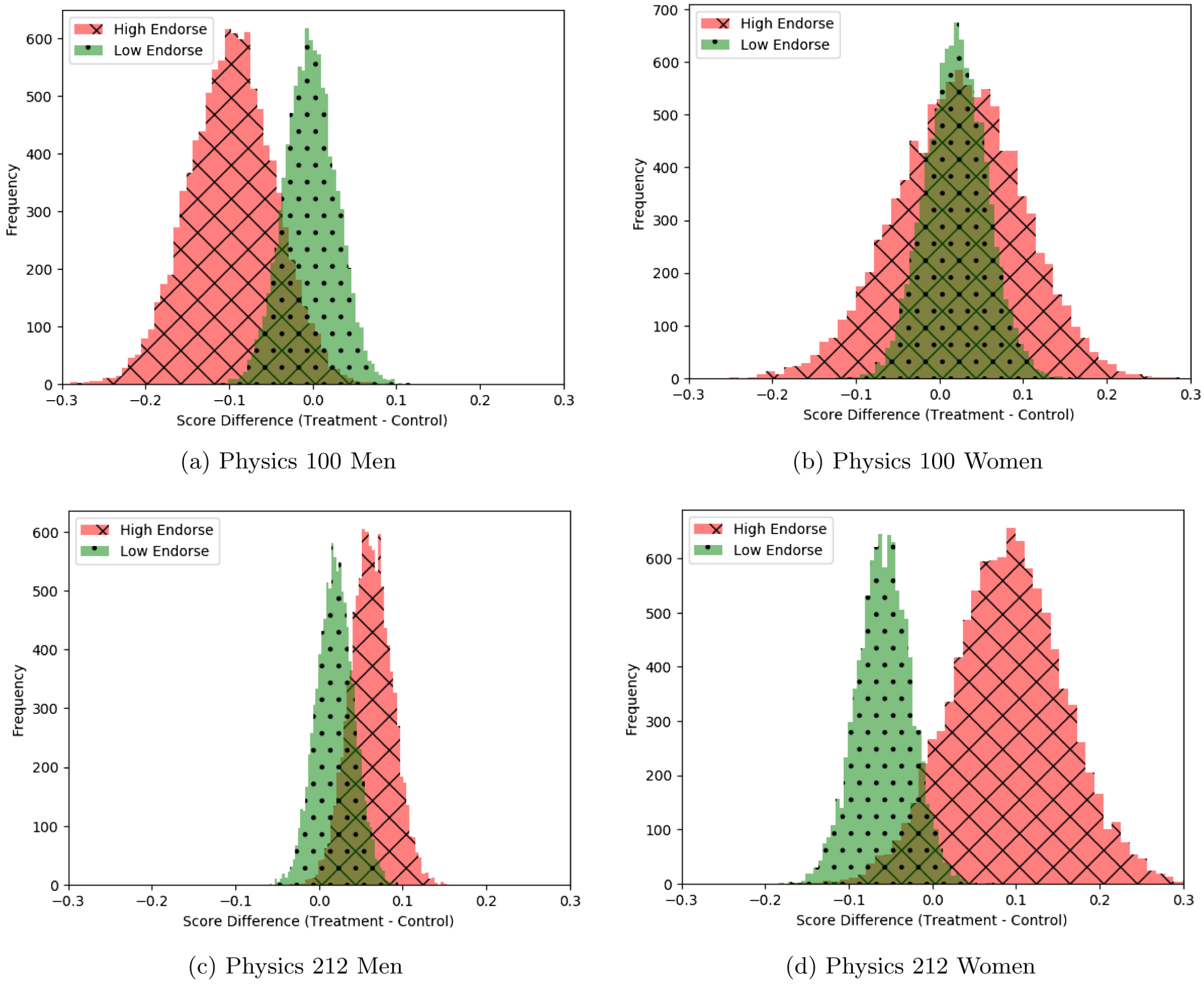
comparison of those gains between high and low endorsing students of each gender within each course.

More closely considering the group of students who should theoretically be most helped by this affirmation, the bootstrapped histogram of gains for high endorsing women in each course are shown in Fig. 15, with (a) unadjusted scores of all high endorsing women and (b) all Physics 100 high endorsing women and a subset of Physics 212 high endorsing women with adjusted scores. In this case, correcting for students' exam scores from their previous course more significantly affected the distribution than in the analysis represented by Fig. 4; without the adjustment, the gain from the affirmation in high endorsing women in Physics 212 was $9.1 \pm 6.7\%$, which was reduced to a gain of $4.0 \pm 4.3\%$ after adjusting for students' prior exam scores. This is significant for the unadjusted scores but loses significance when accounting for students' prior exam scores. The mean of the Physics 100 distribution was $2.1 \pm 6.7\%$, also implying a slight but statistically insignificant improvement for high endorsing women using the treatment. Physics 100 scores could not be adjusted with prior exam performance, as it is the first course in the engineering sequence.

A summary of the raw score difference distributions is shown in Fig. 16, with bootstrapped 95% confidence intervals given as error bars. As with the last summary, values from the first Colorado implementation [27] are included on the plot for comparison. The publication of the second Colorado implementation did not include values for students' stereotype endorsement effects that can be included in this summary. They do note that they saw high endorsing women with the treatment performing better on exams than those in the control group, but performing worse on the FMCE, and that both these effects were not statistically significant.

IV. DISCUSSION

In consolidating all our data and the data from Colorado, we primarily consider the effects of the affirmation on women; we argue that it is more important to lift women's



(a) Physics 100 Men

(b) Physics 100 Women

(c) Physics 212 Men

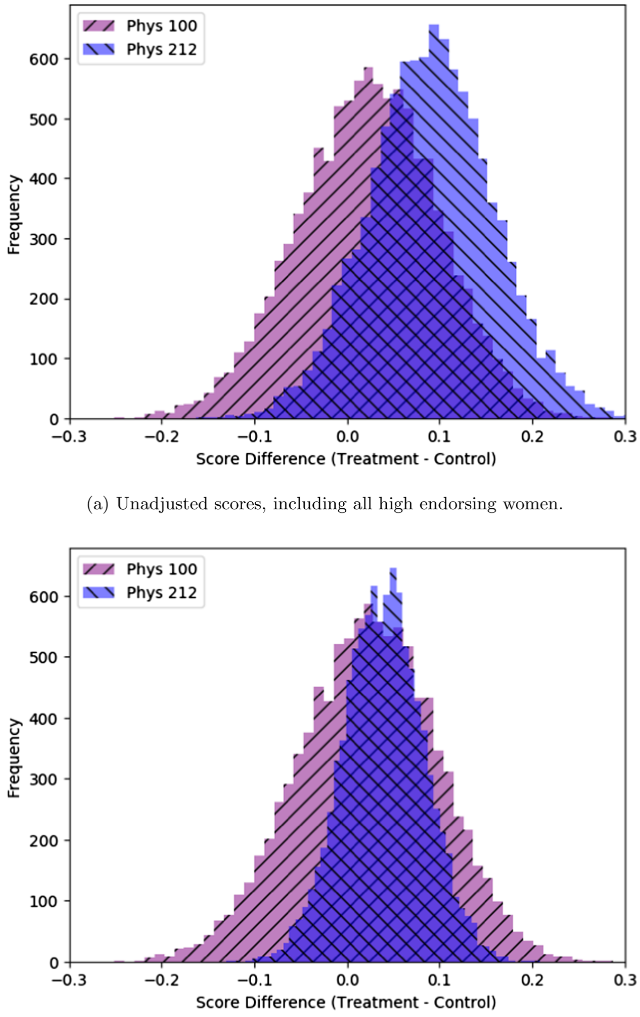
(d) Physics 212 Women

FIG. 14. Histogram of bootstrapped exam gains for treatment participants over control participants, comparing low and high stereotype endorsement. These gains were calculated by subtracting the average exam scores for those in the control condition from exam scores for those in the treatment condition, meaning positive gains imply higher scores from the affirmation exercise while negative gains show lower scores in those who completed the affirmation. In these plots, high endorsers are plotted in red with cross hatching and low endorsers are plotted in light green with spots; dark green with both patterns shows their overlap.

scores which may be affected by stereotype threat than to suppress men's scores to minimize the gap. In the original Colorado studies [27,101], some of the reduction in their gender gaps were due to lower men's scores in the treatment group, which we also saw in our results for Physics 100. The gap "decreased" due to lower men's performance in the treatment group although women's performance did not change from the treatment. Woolf *et al.* saw a similar "decrease" in gaps via white students performing worse and those from ethnic minority groups being unaffected [82]. These results exemplify some of the dangers of measuring gaps to identify success; setting the majority group as the "ideal" is not productive if the intervention is lowering their performance, and pitting the achievements of majority and minority groups against each other does not necessarily elevate minoritized students' successes.

When considering gains by women at the University of Illinois, our two results combine to give us a most likely value of the effect of the treatment for women to be $-1.8 \pm 3.5\%$ (where error is a 95% confidence interval). In comparison, the Colorado papers saw an average effect of $5.8 \pm 3.5\%$.[4] One can see from the summary in Fig. 5 that the spread of the results is large across different implementations, both at different universities and in different courses or iterations at the same university. Using a random effects model to combine all of the data yields $2.5 \pm 4.5\%$, with a $I^2 = 65\%$. These results do not support the use of this treatment to elevate women's performance in physics. Indeed, they suggest that most of the variation seen is due to effects that were not

[4] The weighted averages for Colorado's effects that are cited in this discussion section were calculated by treating each result as an independent experiment, but this is technically a simplification since students' FMCE and Exam scores come from the same groups of students.



(a) Unadjusted scores, including all high endorsing women.

(b) Unadjusted scores for Physics 100, and adjusted scores for Physics 212. The data set for Physics 212 includes the subset of students who took the previous physics course at the University of Illinois and were classified as high endorsing: 14/18 women in the treatment condition, and 11/16 women in the control condition.

FIG. 15. Histogram of bootstrapped difference in exam scores (treatment–control average) for high stereotype endorsing women due to the affirmation activity, from (a) unadjusted scores and (b) adjusted scores in Physics 212. The Physics 100 histogram is shown in reddish purple with diagonal hatching up and to the right (//) and the Physics 212 histogram is bluish purple with diagonal hatching down and to the right (\\); dark purple with cross hatching shows their overlap.

controlled between experiments. This is particularly concerning since both the University of Illinois and the University of Colorado Boulder are similar predominantly white R1 universities, and the protocol to implement the interventions were nearly identical. Indeed the combined statistics of our experiments suggests that our current understanding of affirmation activities is not sufficient to recommend their use without significant new insight into the necessary changes for a positive result.

There are many effects that contribute to differences in performance between men and women in physics classrooms, but if we isolate the effects of stereotype threat, the intervention should be most useful for women who endorse gender stereotypes in STEM. The threshold that we used to determine high and low endorsers is somewhat arbitrary but was borrowed from Colorado's implementation. Initially, including "neutral" responses (response 3) to the statement that men perform better than women as high endorsing felt strange, but further reflection has made us more comfortable with this categorization; if a person does not outright reject the notion that men perform better than women, this is a form of endorsement. The survey item is not suggesting that women must perform better than men for students to disagree with the statement, just to assert whether or not they believe women inherently perform worse.

When we restricted our analysis to only students who marked their survey responses with 4 or 5, corresponding to "agreeing" and "strongly agreeing" that men perform better than women, our sample sizes were very small and showed no statistically significant improvements. In particular, women with the treatment who endorsed by agreeing (a response of 4) saw a decline in performance from the treatment and both high endorsing men and women in Physics 212 saw the same improvement, which should not be true via the affirmation theory. Neither courses had women strongly agreeing (a response of 5) in both treatment and control groups to calculate the difference between the conditions. Although not elaborated on, students' responses to societal expectations showed much more variation, and believing that society expects you to fail whether you personally subscribe to the stereotype may also affect students' susceptibility to stereotype threat.

Using the more generous definitions of low and high endorsers for our analysis to be consistent with the Colorado implementation, we saw that the effects for high endorsing women averaged around $6 \pm 3\%$ at the University of Illinois, where again error cited is a 95% confidence interval. This value is buoyed by our largest positive effect (from the unadjusted data in Physics 212), but we note that the particular effect was diminished when we controlled for prior performance. At Colorado, their two reported effects can be combined to estimate an $15 \pm 8\%$ improvement, though this is driven up by their result for high endorsing women on the FMCE, and we could not quantify their additional contradictory results from the replication study to include in the overall value for their endorsement effect. The replication authors cite that they saw high-endorsing women with the treatment outperform the high-endorsing women in control on their exams, but perform worse on the FMCE, though both effects were not statistically significant [101]. Especially in contrast to the large positive signal of the FMCE in the original study, the negative result and the weaker positive result would bring down the overall endorsement effect value. Because of the small sample sizes of endorsing women in each of the studies, we did not

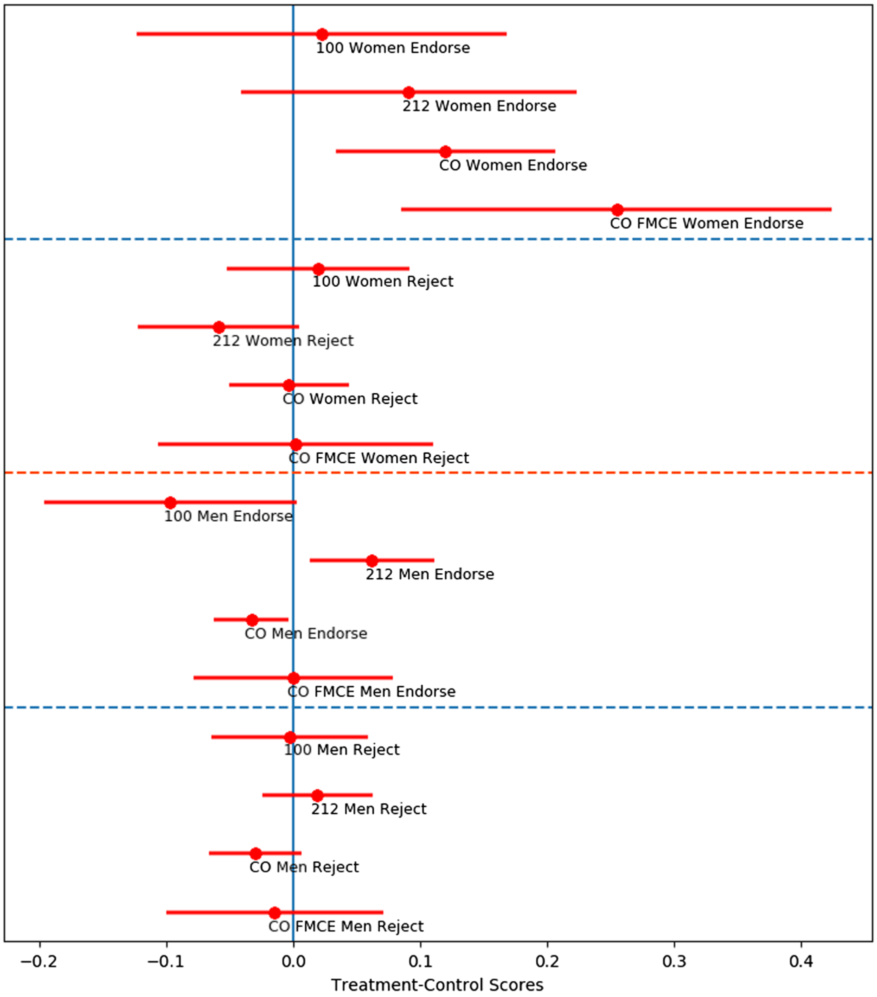


FIG. 16. Summary of effects within each gender and treatment, partitioned by stereotype endorsement. “Endorse” values refer to student averages of those who had high endorsement of the stereotype, while “Reject” values refer to low endorsement. Points indicate mean value of treatment–control average values and error bars are 95% confidence intervals. CO refers to the first Colorado experiment [27]. The values for Colorado are adjusted scores, and their error bars are 95% confidence intervals, calculated from their reported standard errors. Note that within Colorado’s experiment, the exam and FMCE score effects are not independent; the pair of scores (exam and FMCE) is from the same group of students.

quantitatively combine all of the endorsing women’s results across institutions.

The inconsistencies between our results at the University of Illinois and the published results of the values affirmation studies at the University of Colorado Boulder could be explained by statistical fluctuations in either our studies or theirs, or by our experiments being different in ways we did not anticipate or understand. When results are presented from individual studies, it is more likely that a statistically significant fluctuation may be interpreted as a positive result, and combining multiple studies together creates space to think critically about the underlying variables which make these interventions successful or not. Our summary plots showed a lot of variation in results that were not statistically consistent, except at the edges of the confidence intervals which were consistent with zero effect. Especially in the context of a very popular intervention, a takeaway from our disparate results is that the variables affecting student performance are not as simple as doing the activity versus not doing the activity. It is likely that additional factors contribute to the effect which are not explicitly defined or described; following the protocol outlined by the Colorado paper was not sufficient to replicate success. Values affirmation exercises may be a space for further research if one has new insights into the contributing variables, but we advise against repeating them as a broad intervention.

ACKNOWLEDGMENTS

The authors would like to thank Nadya Mason for her support of this project, including its funding via the Heising-Simons Foundation Grant No. 2017-0463.

APPENDIX: LINEAR REGRESSION RESULTS

For completeness, we include results from a linear regression fit for predicting the midterm exam score including factors of gender, treatment and endorsement for Physics 100 and Physics 212. The results are presented in Tables V–IX.

TABLE V. Parameters for models predicting Physics 100 midterm exam performance based on math ACT, gender, treatment, and endorsement. Math ACT scores where normalized to the class average and standard deviation (e.g., a score of 1 means 1 standard deviation above the class average). Numbers in parentheses represent the standard error. The endorsement field is the z score based on the integer value of the student selection 1 = low endorsement 5 = high endorsement. Model including endorsement excludes the 22 students who chose 6 (No Response) on the survey.

	ACT + female	ACT + female ∗ treatment	ACT + female ∗ treatment ∗ endorsement
Intercept	0.6943 (0.01)	0.7060 (0.014)	0.7004 (0.015)
ACT	0.0653 (0.008)	0.0645 (0.008)	0.0644 (0.008)
Female	−0.0265 (0.016)	−0.0352 (0.022)	−0.0253 (0.023)
Treatment		−0.0236 (0.019)	−0.0175 (0.021)
Female: Treatment		0.0178 (0.032)	−0.0095 (0.033)
Endorsement			−0.0147 (0.013)
Female: Endorsement			−0.0106 (0.021)
Treatment: Endorsement			0.0163 (0.020)
Female:Treatment: Endorsement			0.0037 (0.03)

TABLE VI. Parameters for models predicting Physics 212 midterm exam 1 performance based on math ACT, gender, treatment, and endorsement. Math ACT scores where normalized to the class average and standard deviation (e.g., a score of 1 means 1 standard deviation above the class average). Numbers in parentheses represent the standard error. First two models include 894 students who participated in the study and for whom we had ACT or SAT scores. The endorsement field is the z score based on the integer value of the student selection 1 = low endorsement 5 = high endorsement. Model including endorsement excludes the 30 students who chose 6 (No Response) on the survey.

	ACT + female	ACT + female ∗ treatment	ACT + female ∗ treatment ∗ endorsement
Intercept	0.739 (0.006)	0.7252 (0.009)	0.7306 (0.02)
ACT	0.0664 (0.005)	0.06664 (0.005)	0.0655 (0.006)
Female	−0.0319 (0.012)	0.0057 (0.018)	−0.0034 (0.018)
Treatment		0.0279 (0.013)	0.0215 (0.013)
Female: Treatment		−0.072 (0.024)	−0.06 (0.025)
Endorsement			−0.0085 (0.009)
Female: Endorsement			−0.0262 (0.02)
Treatment: Endorsement			0.0173 (0.013)
Female:Treatment: Endorsement			0.0193 (0.027)

TABLE VII. Parameters for models predicting Physics 212 average exam performance (3 midterms + final) based on math ACT, gender, treatment, and endorsement. Math ACT scores were normalized to the class average and standard deviation (e.g., a score of 1 means 1 standard deviation above the class average). Numbers in parentheses represent the standard error. First two models include 894 students who participated in the study and for whom we had ACT or SAT scores. The endorsement field is the z score based on the integer value of the student selection 1 = low endorsement 5 = high endorsement. Model including endorsement excludes the 30 students who chose 6 (No Response) on the survey.

	ACT + female	ACT + female ∗ treatment	ACT + female ∗ treatment ∗ endorsement
Intercept	0.7632 (0.005)	0.753 (0.007)	0.7546 (0.005)
ACT	0.0592 (0.005)	0.0591 (0.005)	0.0587 (0.008)
Female	−0.0319 (0.01)	−0.0143 (0.015)	−0.0203 (0.015)
Treatment		0.0206 (0.011)	0.0197 (0.011)
Female: Treatment		−0.0345 (0.02)	−0.0267 (0.021)
Endorsement			0.0015 (0.007)
Female: Endorsement			−0.038 (0.017)
Treatment: Endorsement			−0.0015 (0.011)
Female: Treatment: Endorsement			0.0399 (0.023)

TABLE VIII. Parameters for models predicting Physics 100 midterm exam performance for women based on math ACT, treatment, and endorsement. Math ACT scores where normalized to the class average and standard deviation (e.g., a score of 1 means 1 standard deviation above the class average). Numbers in parentheses represent the standard error. First two models include 137 female students who participated in the study and for whom we had ACT or SAT scores. The endorsement field is the integer value of the student selection 1 = low endorsement 5 = high endorsement. Model including endorsement excludes the 3 students who chose 6 (No Response) on the survey.

	ACT	ACT + treatment	ACT + treatment ∗ endorsement
Intercept	0.667 (0.012)	0.6707 (0.018)	0.7142 (0.043)
ACT	0.0814 (0.012)	0.0816 (0.013)	0.0815 (0.013)
Treatment		−0.0071 (0.025)	−0.032 (0.061)
Endorsement			−0.0159 (0.013)
Treatment: Endorsement			0.01 (0.018)

TABLE IX. Parameters for models predicting Physics 212 midterm exam 1 performance for women based on math ACT, treatment, and endorsement. Math ACT scores were normalized to the class average and standard deviation (e.g., a score of 1 means 1 standard deviation above the class average). Numbers in parentheses represent the standard error. First two models include 243 female students who participated in the study and for whom we had ACT or SAT scores. The endorsement field is the z score based on the integer value of the student selection 1 = low endorsement 5 = high endorsement. Model including endorsement excludes the 5 students who chose 6 (No Response) on the survey.

	ACT	ACT + treatment	ACT + treatment ∗ endorsement
Intercept	0.7081 (0.011)	0.7321 (0.016)	0.7282 (0.017)
ACT	0.0709 (0.011)	0.0716 (0.011)	0.0695 (0.011)
Treatment		−0.0447 (0.022)	−0.0390 (0.023)
Endorsement			−0.0346 (0.019)
Treatment: Endorsement			0.0364 (0.025)

[1] J. M. Nissen and J. T. Shemwell, Gender, experience, and self-efficacy in introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 020105 (2016).

[2] M. Ong, C. Wright, L. L. Espinosa, and G. Orfield, Inside the double bind: A synthesis of empirical research on undergraduate and graduate women of color in science, technology, engineering,and mathematics, Harv. Educ. Rev. **81**, 172 (2011).

[3] V. Sawtelle, E. Brewe, and L. H. Kramer, Exploring the relationship between self-efficacy and retention in introductory physics, J. Res. Sci. Teach. **49**, 1096 (2012).

[4] A. Traxler and E. Brewe, Equity investigation of attitudinal shifts in introductory physics, Phys. Rev. ST Phys. Educ. Res. **11**, 020132 (2015).

[5] E. M. Marshman, Z. Yasemin Kalender, T. Nokes-Malach, C. Schunn, and C. Singh, Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm?, Phys. Rev. Phys. Educ. Res. **14**, 020123 (2018).

[6] K. L. Lewis, J. G. Stout, S. J. Pollock, N. D. Finkelstein, and T. A. Ito, Fitting in or opting out: A review of key social-psychological factors influencing a sense of belonging for women in physics, Phys. Rev. Phys. Educ. Res. **12**, 020110 (2016).

[7] S. L. Eddy and S. E. Brownell, Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines, Phys. Rev. Phys. Educ. Res. **12**, 020106 (2016).

[8] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and a complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).

[9] M. Bruun, S. Willoughby, and J. L. Smith, Identifying the stereotypical who, what, and why of physics and biology, Phys. Rev. Phys. Educ. Res. **14**, 020125 (2018).

[10] A. Maries, N. I. Karim, and C. Singh, Is agreeing with a gender stereotype correlated with the performance of female students in introductory physics?, Phys. Rev. Phys. Educ. Res. **14**, 020119 (2018).

[11] Z. Hazari, P. M. Sadler, and G. Sonnert, The science identity of college students: Exploring the intersection of gender, race, and ethnicity, J. Coll. Sci. Teach. **42**, 82 (2013), https://www.jstor.org/stable/43631586.

[12] L. E. Kost-Smith, S. J. Pollock, and N. D. Finkelstein, Gender disparities in second-semester college physics: The incremental effects of a “smog of bias”, Phys. Rev. ST Phys. Educ. Res. **6**, 020112 (2010).

[13] V. Seyranian, A. Madva, N. Duong, N. Abramzon, Y. Tibbetts, and J. M. Harackiewicz, The longitudinal effects of STEM identity and gender on flourishing and achievement in college physics, Int. J. STEM Educ. **5**, 40 (2018).

[14] K. Rosa and F. Moore Mensah, Educational pathways of Black women physicists: Stories of experiencing and overcoming obstacles in life, Phys. Rev. Phys. Educ. Res. **12**, 020113 (2016).

[15] X. R. Quichocho, J. Conn, E. M. Schipull, and E. W. Close, Who does physics? Understanding the composition of physicists through the lens of women of color and LGBTQ + women physicists, *Proceedings of the Physics Education Research Conference 2019, Provo, UT*, https://doi.org/10.1119/perc.2019.pr.Quichocho.

[16] L. Bian, A. Cimpian, and S.-J. Leslie, Gender stereotypes about intellectual ability emerge early and influence children’s interests, Science **355**, 389 (2017).

[17] A. M. Kelly, Physics teachers’ perspectives on factors that affect urban physics participation and accessibility, Phys. Rev. ST Phys. Educ. Res. **9**, 010122 (2013).

[18] N. Reid and E. A. Skryabina, Gender and physics, Int. J. Sci. Educ. **25**, 509 (2003).

[19] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).

[20] I. Rodriguez, G. Potvin, and L. H. Kramer, How gender and reformed introductory physics impacts student success in advanced physics courses and continuation in the physics major, Phys. Rev. Phys. Educ. Res. **12**, 020118 (2016).

[21] P. W. Irving and E. C. Sayre, Becoming a physicist: The roles of research, mindsets, and milestones in upper-division student perceptions, Phys. Rev. ST Phys. Educ. Res. **11**, 020120 (2015).

[22] V. Sawtelle, E. Brewe, R. Michelle Goertzen, and L. H. Kramer, Identifying events that impact self-efficacy in physics learning, Phys. Rev. ST Phys. Educ. Res. **8**, 020111 (2012).

[23] E. W. Close, J. Conn, and H. G. Close, Becoming physics people: Development of integrated physics identity through the Learning Assistant experience, Phys. Rev. Phys. Educ. Res. **12**, 010109 (2016).

[24] R. M. Lock, Z. Hazari, and G. Potvin, Physics career intentions: The effect of physics identity, math identity, and gender, AIP Conf. Proc. **1513**, 262 (2013).

[25] Z. Hazari, G. Sonnert, P. M. Sadler, and M.-C. Shanahan, Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study, J. Res. Sci. Teach. **47**, 978 (2010).

[26] G. L. Cohen, J. Garcia, N. Apfel, and A. Master, Reducing the racial achievement gap: A social-psychological intervention, Science **313**, 1307 (2006).

[27] A. Miyake, L. E. Kost-Smith, N. D. Finkelstein, S. J. Pollock, G. L. Cohen, and T. A. Ito, Reducing the gender achievement gap in college science: A classroom study of values affirmation, Science **330**, 1234 (2010).

[28] S. Andersson and A. Johansson, Gender gap or program gap? Students’ negotiations of study practice in a course in electromagnetism, Phys. Rev. Phys. Educ. Res. **12**, 020112 (2016).

[29] R. Gutiérrez, A “gap-gazing” fetish in mathematics education? Problematizing research on the achievement gap, J. Res. Math. Educ. **39**, 357 (2008), https://www.jstor.org/stable/40539302.

[30] A. Danielsson, Gender in physics education research: A review and a look forward, in *Never Mind the gap! Gendering Science in Transgressive Encounters* (2010), http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-254006.

[31] G. Ladson-Billings, From the achievement gap to the education debt: Understanding achievement in U.S. schools, Educ. Res. **35**, 3 (2006).

[32] D. Bauder, AP says it will capitalize Black but not white, 2020, https://www.ap.org/ap-in-the-news/2020/ap-says-it-will-capitalize-black-but-not-white.

[33] K. A. Appiah, The case for capitalizing the B in Black, The Atlantic **18** (2020), https://www.theatlantic.com/ideas/archive/2020/06/time-to-capitalize-blackand-white/613159/.

[34] M. Laws, Why we capitalize ‘Black’ (and not ‘white’), Columbia Journalism Rev. **16** (2020), https://www.cjr.org/analysis/capital-b-black-styleguide.php.

[35] E. Zorn, Column: Should ‘white’ be capitalized? It feels wrong, but it’s the way to go, Chicago Tribune (2020), https://www.chicagotribune.com/columns/eric-zorn/ct-column-capitalize-white-black-language-race-zorn-20200709-e42fag6ivbazdblizpopsp4p2a-story.html.

[36] N. Irvin Painter, Opinion: Why ‘White’ should be capitalized, too, The Washington Post (July 22, 2020), https://www.washingtonpost.com/opinions/2020/07/22/why-white-should-be-capitalized/.

[37] G. L. Cochran, A. Gupta, S. Hyater-Adams, A. V. Knaub, and B. Zamarripa Roman, Emerging reflections from the people of color (POC) at PERC discussion space, 2019, arXiv:1907.01655.

[38] K. Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*, University of Chicago Legal Forum Vol. 1989, p. 8, https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8.

[39] S. Kanim and X. C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).

[40] S. Theule Lubienski, On “gap gazing” in mathematics education: The need for gaps analyses, J. Res. Math. Educ. **39**, 350 (2008), https://www.jstor.org/stable/40539301.

[41] C. M. Steele, S. J. Spencer, and J. Aronson, Contending with group image: The psychology of stereotype and social identity threat, Adv. Exp. Soc. Psychol. **34**, 379 (2002).

[42] G. M. Walton and S. J. Spencer, Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students’, Psychol. Sci. **20**, 1132 (2009).

[43] G. L. Cohen and D. K. Sherman, The psychology of change: Self-affirmation and social psychological intervention, Annu. Rev. Psychol. **65**, 333 (2014).

[44] A. Madsen, S. B. McKagan, and E. C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).

[45] D. S. Yeager and G. M. Walton, Social-psychological interventions in education: They're not magic, Rev. Educ. Res. **81**, 267 (2011).

[46] N. Shnabel, V. Purdie-Vaughns, J. E. Cook, J. Garcia, and G. L. Cohen, Demystifying values-affirmation interventions: Writing about social belonging is a key to buffering against identity threat, Pers. Soc. Psychol. Bull. **39**, 663 (2013).

[47] S. Stoessner and C. Good, Stereotype threat: An overview, 2011, https://diversity.arizona.edu/sites/default/files/stereotype_threat_overview.pdf.

[48] C. M. Steele and J. Aronson, Stereotype threat and the intellectual test performance of African Americans, J. Personality Soc. Psychol. **69**, 797 (1995).

[49] S. J. Spencer, C. M. Steele, and D. M. Quinn, Stereotype threat and women's math performance, J. Exp. Soc. Psychol. **35**, 4 (1999).

[50] M. Walsh, C. Hickey, and J. Duffy, Influence of item content and stereotype situation on gender differences in mathematical problem solving, Sex Roles **41**, 219 (1999).

[51] C. Good, J. Aronson, and J. Ann Harder, Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses, J. Appl. Dev. Psychol. **29**, 17 (2008).

[52] M. Inzlicht and T. Ben-Zeev, A threatening intellectual environment: Why females are susceptible to experiencing problem-solving deficits in the presence of males, Psychol. Sci. **11**, 365 (2000).

[53] N. Ambady, M. Shih, A. Kim, and T. L. Pittinsky, Stereotype susceptibility in children: Effects of identity activation on quantitative performance, Psychol. Sci. **12**, 385 (2001).

[54] J.-C. Croizet and T. Claire, Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds, Pers. Soc. Psychol. Bull. **24**, 588 (1998).

[55] L. A. Harrison, C. M. Stevens, A. N. Monty, and C. A. Coakley, The consequences of stereotype threat on the academic performance of White and non-White lower income college students, Social Psychol. Educ. **9**, 341 (2006).

[56] D. K. Sherman, K. A. Hartson, K. R. Binning, V. Purdie-Vaughns, J. Garcia, S. Taborsky-Barba, S. Tomassetti, A. David Nussbaum, and G. L. Cohen, Deflecting the trajectory and changing the narrative: How self-affirmation affects academic performance and motivation under identity threat, J. Personality Soc. Psychol. **104**, 591 (2013).

[57] P. M. Gonzales, H. Blanton, and K. J. Williams, The effects of stereotype threat and double-minority status on the test performance of Latino women, Personality Soc. Psychol. Bull. **28**, 659 (2002).

[58] T. Schmader and M. Johns, Converging evidence that stereotype threat reduces working memory capacity, J. Personality Soc. Psychol. **85**, 440 (2003).

[59] B. Levy, Improving memory in old age through implicit self-stereotyping, J. Personality Soc. Psychol. **71**, 1092 (1996).

[60] J. K. Bosson, E. L. Haymovitz, and E. C. Pinel, When saying and doing diverge: The effects of stereotype threat on self-reported versus non-verbal anxiety, J. Exp. Soc. Psychol. **40**, 247 (2004).

[61] J. Stone, M. Sjomeling, C. I. Lynch, and J. M. Darley, Stereotype threat effects on black and white athletic performance, J. Personality Soc. Psychol. **77**, 1213 (1999).

[62] L. J. Kray, A. D. Galinsky, and L. Thompson, Reversing the gender gap in negotiations: An exploration of stereotype regeneration, Organ. Behav. Human Decis. Processes **87**, 386 (2002).

[63] N. C. J. Yeung and C. von Hippel, Stereotype threat increases the likelihood that female drivers in a simulator run over jaywalkers, Accid. Anal. Prev. **40**, 667 (2008).

[64] C. M. Steele, S. J. Spencer, and J. Aronson, Contending with group image: The psychology of stereotype and social identity threat, in *Advances in Experimental Social Psychology* (Academic Press, New York, 2002), Vol. 34, pp. 379–440.

[65] C. M. Steele, A threat in the air: How stereotypes shape intellectual identity and performance, Am. Psychol. **52**, 613 (1997).

[66] A. C. Krendl, J. A. Richeson, W. M. Kelley, and T. F. Heatherton, The negative consequences of threat: A functional magnetic resonance imaging investigation of the neural mechanisms underlying women's underperformance in math, Psychol. Sci. **19**, 168 (2008).

[67] J. C. Croizet, G. Despres, M. E. Gauzins, P. Huguet, J. P. Leyens, and A. Meot, Stereotype threat undermines intellectual performance by triggering a disruptive mental load, Pers. Soc. Psychol. Bull. **30**, 721 (2004).

[68] J. Keller and D. Dauenheimer, Stereotype threat in the classroom: Dejection mediates the disrupting threat effect on women's math performance, Pers. Soc. Psychol. Bull. **29**, 371 (2003).

[69] C. Stangor, C. Carr, and L. Kiang, Activating stereotypes undermines task performance expectations, J. Personality Soc. Psychol. **75**, 1191 (1998).

[70] J. Thomas Kellow and B. D. Jones, The effects of stereotypes on the achievement gap: Reexamining the academic performance of African American high school students', J. Black Psychol. **34**, 94 (2008).

[71] C. Graham, R. W. Baker, and S. Wapner, Prior interracial experience and Black student transition into predominantly White colleges, J. Personality Soc. Psychol. **47**, 1146 (1984).

[72] J. Aronson, C. B. Fried, and C. Good, Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence, J. Exp. Soc. Psychol. **38**, 113 (2002).

[73] C. S. Dweck, C.-y. Chiu, and Y.-y. Hong, Implicit theories and their role in judgments and reactions: A word from two perspectives, Psychol. Inquiry **6**, 267 (1995).

[74] C. S. Dweck, *Mindset: The New Psychology of Success* (Random House Digital, Inc., New York, 2008), ISBN: 9780345472328, 0345472322.

[75] C. M. Steele, The psychology of self-affirmation: Sustaining the integrity of the self, Adv. Exp. Soc. Psychol. **21**, 261 (1988).

[76] C. M. Steele, Name-calling and compliance, J. Personality Soc. Psychol. **31**, 361 (1975).

[77] A. McQueen and W. M. P. Klein, Experimental manipulations of self-affirmation: A systematic review, Self & Identity **5**, 289 (2006).

[78] D. K. Sherman and G. L. Cohen, The psychology of self-defense: Self-affirmation theory, Adv. Exp. Soc. Psychol. **38**, 183 (2006).

[79] G. L. Cohen, J. Garcia, V. Purdie-Vaughns, N. Apfel, and P. Brzustoski, Recursive processes in self-affirmation: Intervening to close the minority achievement gap, Science **324**, 400 (2009).

[80] D. K. Sherman, G. L. Cohen, L. D. Nelson, A. David Nussbaum, D. P. Bunyan, and J. Garcia, Affirmed yet unaware: Exploring the role of awareness in the process of self-affirmation, J. Personality Soc. Psychol. **97**, 745 (2009).

[81] N. K. Bowen, K. M. Wegmann, and K. C. Webber, Enhancing a brief writing intervention to combat stereotype threat among middle-school students, J. Educ. Psychol. **105**, 427 (2013).

[82] K. Woolf, I. Chris McManus, D. Gill, and J. Dacre, The effect of a brief social intervention on the examination results of UK medical students: a cluster randomised controlled trial, BMC Med. Educ. **9**, 35 (2009).

[83] J. M. Harackiewicz, E. A. Canning, Y. Tibbetts, C. J. Giffen, S. S. Blair, D. I. Rouse, and J. S. Hyde, Closing the social class achievement gap for first-generation students in undergraduate biology, J. Educ. Psychol. **106**, 375 (2014).

[84] E. D. B. Riggle, K. A. Gonzalez, S. S. Rostosky, and W. W. Black, Cultivating positive LGBTQA identities: An intervention study with college students, J. LGBT Issues Counseling **8**, 264 (2014).

[85] G. M. Walton, C. Logel, J. M. Peach, S. J. Spencer, and 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 (2015).

[86] D. N. Bradley, E. P. Crawford, and S. E. Dahill-Brown, Defining and assessing FoI in a large-scale randomized trial: Core components of values affirmation, Studies Educ. Eval. **49**, 51 (2016).

[87] G. Borman, *Examination of a Self-Affirmation Intervention in St. Paul Public Schools*, The Senior Urban Education Research Fellowship Series (2012), Vol. IX, https://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/85/St_Paul_Self-Aff irmation.pdf.

[88] T. S. Dee, Social identity and achievement gaps: Evidence from an affirmation intervention, J. Res. Educ. Effectiveness **8**, 149 (2015).

[89] G. D. Borman, J. Grigg, and P. Hanselman, An effort to close achievement gaps at scale through self-affirmation, Educ. Eval. Policy Anal. **38**, 21 (2016).

[90] G. D. Borman, J. Grigg, C. S. Rozek, P. Hanselman, and 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 (2018).

[91] H. Jordt, S. L. Eddy, R. Brazil, I. Lau, C. Mann, S. E. Brownell, K. King, and S. Freeman, Values affirmation intervention reduces achievement gap between underrepresented minority and white students in introductory biology classes, CBE Life Sci. Educ. **16** (2017).

[92] S. Lauer, J. Momsen, E. Offerdahl, M. Kryjevskaia, W. Christensen, and L. Montplaisir, Stereotyped: Investigating gender in introductory science courses, CBE Life Sci. Educ. **12**, 30 (2013).

[93] A. Franco, N. Malhotra, and G. Simonovits, Publication bias in the social sciences: Unlocking the file drawer, Science **345**, 1502 (2014).

[94] J. Mervis, Research transparency: Why null results rarely see the light of day, Science **345**, 992 (2014).

[95] D. Fanelli, Negative results are disappearing from most disciplines and countries, Scientometrics **90**, 891 (2012).

[96] D. Chavalarias, J. D. Wallach, A. H. T. Li, and J. P. A. Ioannidis, Evolution of reporting P values in the biomedical literature, 1990-2015: Evolution of reporting P values in the biomedical literature, 1990–2015, J. Am. Med. Assoc. **315**, 1141 (2016).

[97] R. Rosenthal, The file drawer problem and tolerance for null results, Psychol. Bull. **86**, 638 (1979).

[98] L. D. Conlin, E. Kuo, and N. R. Hallinen, Nothing's plenty: The significance of null results in physics education research, arXiv:1810.10071.

[99] T. D. Sterling, W. L. Rosenbaum, and J. J. Weinkam, Publication Decisions Revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa, Am. Statistician **49**, 108 (1995).

[100] Z. Wu, T. F. Spreckelsen, and G. L. Cohen, A meta-analysis of the effect of values affirmation on academic achievement, J. Soc. Issues 1 (2021).

[101] L. E. Kost-Smith, S. J. Pollock, N. D. Finkelstein, G. L. Cohen, T. A. Ito, and A. Miyake, Replicating a self-affirmation intervention to address gender differences: Successes and challenges, AIP Conf. Proc. **1413**, 231 (2012).

[102] M. Rifkin, Addressing underrepresentation: Physics teaching for all, Phys. Teach. **54**, 72 (2016).

[103] Z. Hazari, G. Potvin, R. M. Lock, F. Lung, G. Sonnert, and P. M. Sadler, Factors that affect the physical science career interest of female students: Testing five common hypotheses, Phys. Rev. ST Phys. Educ. Res. **9**, 020115 (2013).

[104] E. A. Eschenbach, M. Virnoche, E. M. Cashman, S. M. Lord, and M. M. Camacho, Proven practices that can reduce stereotype threat in engineering education: A literature review, in *Proceedings of the 2014 IEEE Frontiers in Education Conference (FIE)* (2014), pp. 1–9, https://doi.org/10.1109/FIE.2014.7044011.

[105] L. Aguilar, G. Walton, and C. Wieman, Psychological insights for improved physics teaching, Phys. Today **67**, No. 5, 43 (2014).

[106] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).

[107] B. W. Fitzgerald, Using superheroes such as Hawkeye, Wonder Woman and the Invisible Woman in the physics classroom, Phys. Educ. **53**, 035032 (2018).

[108] N. Abramzon, P. Benson, E. Bertschinger, S. Blessing, G. L. Cochran, A. Cox, B. Cunningham, J. Galbraith-Frew, J. Johnson, L. Kerby, E. Lalanne, C. O'Donnell, S. Petty, S. Sampath, S. Seestrom, C. Singh, C. Spencer, K. SparksWoodle, and S. Yennello, Women in physics in the United States: Recruitment and retention, AIP Conf. Proc. **1697**, 060045 (2015).

[109] T. Scott, A. Gray, and P. Yates, A controlled comparison of teaching methods in first-year university physics, J. R. Soc. New Zealand **43**, 88 (2013).

[110] J. M. Valla and W. M. Williams, Increasing achievement and higher-education representation of under-represented groups in science, technology, engineering, and mathematics fields: A review of current L-12 intervention programs, J. Women Minorities Sci. Eng. **18**, 21 (2012).

[111] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler, Exploring the gender gap in the Conceptual Survey of Electricity and Magnetism, Phys. Rev. Phys. Educ. Res. **13**, 020114 (2017).

[112] K. Kreutzer and A. Boudreaux, Preliminary investigation of instructor effects on gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **8**, 010120 (2012).

[113] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[114] B. Efron, Bootstrap methods: Another look at the jackknife, Ann. Stat. **7**, 1 (1979).