



## Supporting Online Material for

### **Reducing the Gender Achievement Gap in College Science: A Classroom Study of Values Affirmation**

Akira Miyake,\* Lauren E. Kost-Smith, Noah D. Finkelstein, Steven J. Pollock, Geoffrey L. Cohen, Tiffany A. Ito

\*To whom correspondence should be addressed. E-mail: [akira.miyake@colorado.edu](mailto:akira.miyake@colorado.edu)

Published 26 November 2010, *Science* **330**, 1234 (2010)  
DOI: 10.1126/science.1195996

#### **This PDF file includes:**

Materials and Methods  
SOM Text  
Table S1  
References and Notes

## MATERIALS AND METHODS

### Subjects

Subjects were recruited from an introductory calculus-based physics course at the University of Colorado. Approximately  $\frac{3}{4}$  of the enrolled students were men (74%), which is typical for this course.

The study sample consisted of 439 students (311 men and 128 women) who (a) were over 18 years of age, (b) completed both writing exercises in weeks 1 and 4, and (c) took the final exam at the end of the semester. This sample represented 72.9% of the 602 students over 18 years of age who received a grade in the course. Of those 439 students in the sample, 7 students were removed from the analysis for failure to follow the writing instructions for one or both of the writing exercises (e.g., by discussing the personal importance of a value when in the control condition). Twenty-three students did not complete the stereotype endorsement survey question, and 8 students did not have prior mathematics performance data (either SAT or ACT Math scores). An additional 2 students were removed based on two multivariate outlier analyses, one involving the four exam scores and one involving the Force and Motion Conceptual Evaluation (FMCE) (*SI*) scores obtained at the beginning (week 1) and end (week 15) of the semester. These students had Mahalanobis distances of 22.6 for the exam scores (critical value=18.5) and 14.0 for the FMCE scores (critical value=13.8), respectively.

This left a final sample of 399 students (283 men and 116 women) on which the reported analyses were based. An analysis of the attrition from the original study sample (439 students) to the final sample (399 students) showed no difference in attrition by condition for all students [ $\chi^2(1, N=439)=1.35, P=0.24$ ], or for women specifically [ $\chi^2(1, N=128)=0.01, P=0.94$ ].

As noted later in more detail, we randomly assigned approximately 60% of students to

the values affirmation condition and 40% to the control condition. This assignment plan was largely successful; the final sample of 399 students consisted of 178 men and 69 women in the affirmation condition (62.9% of men and 59.5% of women in the final sample) and 105 men and 47 women in the control condition (37.1% and 40.5%, respectively). Because 308 (77.2%) of the 399 students (212 men and 96 women) in the final sample took the FMCE in both weeks 1 and 15, the sample used for the analysis of the FMCE scores consisted of 137 men and 55 women in the affirmation condition (64.6% of men and 57.3% of women in this sample) and 75 men and 41 women in the control condition (35.4% and 42.7%, respectively).

### **Course Description**

We conducted this study in the first-semester, calculus-based introductory mechanics course at the University of Colorado. This is the first course in a three-semester introductory physics sequence for science and engineering majors. This 15-week course covers traditional mechanics content including Newton's laws, work, energy, momentum, and waves. Students met three times per week (Monday, Wednesday, and Friday) for 50-minute lectures. There were two lecture sections of the course offered, one at 9:00 AM and the other at 11:00 AM. The two lecture sections were taught by the same instructor using identical materials, and the outcome measures used in this study were identical in the two sections. The same exam questions were given to students in both sections. All exams were administered at the same time, but students were separated into four rooms for each of the three midterm exams. Students were assigned to one of the four exam rooms based on their assigned teaching assistant (TA). The two sections were graded on the same scale and so were combined for analysis.

As in prior offerings of the course, the instructor implemented pedagogical features previously shown to minimize gender differences in performance (*S2*). The 50-minute lecture

period was interspersed with conceptual questions that students would discuss in small groups (*Peer Instruction (S3)*) and then answer using personal response (clicker) systems. The instructor usually asked between 4 and 6 conceptual questions during each lecture (*S4*). The instructor was a highly experienced teacher, well versed in the interactive engagement techniques that he used during lecture sections.

In addition to lectures, students also attended a 50-minute recitation section once each week (Thursday). There were 24 recitation sections (12 recitation sections for each lecture section) with about 30 students in each section. The recitation sections were led by graduate TAs, assisted by undergraduate Learning Assistants (LAs) (*S5*). There were 6 TAs for the course (5 men and 1 woman) and 7 LAs (4 men and 3 women). During the recitation section, students worked through a conceptual workbook called *Tutorials in Introductory Physics (S6)* in small groups of about four students. While students were working on the workbook activities, one TA and one LA would circle around the room answering student questions and engaging in Socratic dialogue with the students. These recitation sections focused students less on generating the correct answers to questions in the workbook and more on discussing the key scientific concepts for the week and engaging with the relevant ideas (*S7, S8*).

Data gathered for over five years in this course indicate that, relative to men, women (a) are less prepared for this course, having lower SAT or ACT Math scores and being less likely to have taken a high school physics class; (b) score lower on the in-class exams; and (c) score worse on the standardized test of conceptual physics understanding (FMCE) both at the beginning and end of the semester in this course (*S9*).

### **Values Affirmation and Control Writing Exercises**

Students completed two brief writing exercises in a randomized, double-blind design.

Each student completed either a values-affirmation writing intervention or a control writing exercise of similar format and length. The writing assignments were delivered early in class (week 1) to maximize the effect of the intervention across the semester, with a second administration shortly before the first exam (week 4) to ensure the potency of the intervention.

The first writing exercise was presented during the first recitation section (week 1). During each of the two lectures that students attended prior to the first recitation section, the professor provided pedagogical context for the writing exercise that they would complete. Specifically, he told the students that effective communication was an important skill for success in physics-related careers and that, to practice communication, they would complete a 10–15 minute writing exercise in recitation. The professor told them that they would not be writing about physics, but about something that they already knew about.

Recitation sections were led by graduate TAs who were naïve to the purpose of the study. Student attendance was mandatory in these sections, and all recitation sections met on the same day. TAs were given a scripted introduction explaining to the students that they would be completing a writing exercise in which they would think about values that are important to people. To further standardize administration, TAs were given scripted answers to possible questions from the students. They distributed to each student a manila envelope containing the informed consent form and writing assignment that was prepared in advance by study personnel. Although there were two versions for the writing assignment (values affirmation and control), the envelopes and formatting of the two exercises looked similar. Additionally, the TA established the expectation of silent concentration. These steps served to minimize the possibility of students' becoming aware of differences in the exercises. The consent form described the study as examining the relation between critical writing and students' experiences in college,

including physics classes. All students over the age of 18 were invited to participate. Students who were younger than 18 and hence could not legally grant consent were instructed to complete an alternative writing assignment, of comparable length, asking about past physics experiences.

The values affirmation and control exercises closely followed procedures developed and validated in prior research (*S10, S11*). Students in each writing condition received a three-page packet. The first page listed 12 values: *being good at art; creativity; relationships with family and friends; government or politics; independence; learning and gaining knowledge; athletic ability; belonging to a social group (such as your community, racial group, or school club); music; career; spiritual or religious values; and sense of humor*. The values were similar to those used in past research (*S10, S11*), though modified somewhat for the present sample, and were selected to represent a range of values that students may or may not endorse. We avoided values that explicitly dealt with science and math. Students in the affirmation condition were instructed to circle the two or three values *most* important to them, whereas students in the control condition were instructed to circle the two or three *least* important values.

Through a series of structured prompts, the second page of the packet instructed students to describe in a few sentences either why the selected values were important to *them* (affirmation condition) or why they might be important to *someone else* (control condition). To decrease evaluation apprehension, students were told to focus on their thoughts and feelings, without worrying about spelling and grammar or how well written their answer was. Lines were provided on two thirds of the page for students to provide their answer.

The final page reinforced the manipulation by asking students to again look at the values they had selected earlier. They were then asked to list either the top two reasons why these values were important to them (affirmation condition) or the top two reasons why these values

might be picked as important by someone else, such as another student at their school or a person they have heard about (control condition). To further encourage reflection about the values, the third page ended by asking students to indicate their agreement with several items using numerical scales (e.g. *In general, I try to live up to these values* in the affirmation condition vs. *In general, some people try to live up to these values* in the control condition).

In each recitation section, 60% of the packets given to each TA contained the values-affirmation writing assignment. We overrepresented the affirmation condition so that any possible benefits could be conferred to the greatest number of students without undermining the rigor of the study. The affirmation and control packets were intermixed randomly. Random assignment to condition occurred when students received a packet. The two writing exercises were formatted identically and were of nearly identical length. Students put all materials (the writing exercise and the consent form) back in the manila envelope when they were done. TAs collected the envelopes after 15 minutes and were instructed to not open them at any time.

Study personnel discreetly monitored the administration of the writing exercise in each recitation from the back of the classroom. They verified that all TAs properly instructed and administered the exercise. Study personnel discreetly collected the completed materials at the end of each recitation after students had departed.

A second administration of the writing exercise was delivered shortly before the first midterm exam (week 4) via a regular, weekly online homework assignment. Students received the assignment on a Friday, 11 days before the first midterm exam, and could complete it any time within the next 8 days, with a Saturday 8 AM deadline. Consequently, students could have completed the second writing exercise between 3 and 11 days before the first midterm. The majority of students (76%) completed the exercise toward the end of the homework period (the

following Thursday, Friday, or Saturday morning). The homework assignment was delivered and completed online. The final question on the homework assignment asked students to follow a link to the university-wide online course portal. Doing so allowed us to customize the content delivered, ensuring that each subject received the writing exercise for the same condition that he or she had completed in recitation in week 1. The writing exercise was similar to the one completed in recitation, but the second exercise was presented online, with students typing their answers and submitting them via the online portal. Instructions suggested that students spend about 15 minutes on the exercise.

Several steps were taken to ensure that all instructional personnel associated with the course were unaware of students' condition assignment. All but the course instructor were blind to the study's purpose and hypotheses. Both writing exercises occurred without the course instructor present (in TA-led recitations and online). All data were handled only by study personnel. The course instructor had no access to information about students' assignment to experimental condition (values affirmation or control) and was not told of any results until after the semester was over. Further precluding any threats to the study's validity, the instructor did not grade exams, and all exams were multiple-choice, with objectively correct vs. incorrect answers, graded by machine. The TAs who distributed the writing exercises in recitation were told only that they would be administering a writing exercise and were not told of the hypotheses or even of the presence of two different writing exercises. The affirmation and control exercises given in recitation were distributed in closed envelopes, with study personnel present in every recitation to verify that TAs never viewed the students' responses. The writing exercises distributed in recitation were formatted identically, preventing TAs from noticing condition assignment. TAs were also instructed to remain at the front of the room while students completed

the exercise in recitation. Completed writing exercises were collected by study personnel immediately at the end of each recitation.

Because the writing exercise portion of the homework assignment (week 4) was not required and students were allowed to opt out, there was some attrition across the two administrations of the intervention. Of the total students completing the first writing exercise in week 1, approximately 74% (439 students) completed the second writing exercise in week 4 and took the final exam (thus comprising the students in the study sample). There was a tendency that, among students who completed the first administration of the intervention, more students completed the second administration in the affirmation condition than in the control condition [ $\chi^2(1, N=591)=3.30, P=0.07$ ], a trend that was almost significant for women [ $\chi^2(1, N=158)=3.70, P=0.054$ ] but not for men [ $\chi^2(1, N=433)=1.31, P=0.25$ ].

### **Stereotype Endorsement Measure**

In the second week of the course, students were asked to complete a survey about their attitudes toward science. This survey is a typical part of this introductory physics course. Embedded in the larger survey was an item asking them to report their expectation that men do better in physics than women. The survey was included as a link on their weekly online homework assignment. Students were asked to follow the link to complete the survey and were told that they would receive extra credit (equivalent to one homework problem) if they included their name on the survey (with no requirement that they answer any questions). Clicking on the link brought students to an online survey containing various attitude measures (e.g., the Colorado Learning Attitudes about Science Survey (*SL2*)). Stereotype endorsement was measured by assessing students' agreement with the statement, *According to my own personal beliefs, I expect men to generally do better in physics than women*, answered on a 5-point scale ranging from

*strongly disagree to strongly agree.*

Because the measure of stereotype endorsement was collected in week 2, after the first administration of the writing exercise, we evaluated whether the writing exercise affected stereotype perceptions. Neither the condition main effect nor the gender  $\times$  condition interaction was significant [ $F_{1,395}=0.05$ ,  $P=0.82$  and  $F_{1,395}=0.31$ ,  $P=0.58$ , respectively], suggesting that there was no effect of the affirmation versus control writing exercises in week 1 on students' reported endorsement of the stereotype in week 2.

## **Outcome Measures**

### **Exam Scores**

The physics course was 15 weeks long (excluding the fall break week), during which the students took three midterm exams and a final exam. The midterm exams were given in weeks 5, 9, and 14 of the course. The final exam was given the morning (Saturday) following the last day of lecture for the course. Each of the three midterm exams was given in the evening, and students had 1.5 hours to complete each exam. The midterm exams each contained 25 multiple-choice questions with 4 or 5 answer options. Students from the two lecture sections took the midterm exams at the same time, but were assigned to four different locations depending on who their TA was. The final exam was cumulative and contained 40 multiple-choice questions, again with 4 or 5 answer options for each question. Students had 2.5 hours to complete the final exam. For the final exam, all students were in the same location. For all exams, two versions of the exam were distributed to students. The same questions were included in both versions of the exam, but the order of the questions and the answer options were shuffled between the two versions. There were no significant differences in students' average scores between each of the exam versions for each of the four exams.

The exams contained mostly conceptual questions, but there were also more procedural, mathematical, or computational questions. The conceptual questions tested students' understanding of the basic ideas of the course, without necessarily requiring the application of extended mathematical algorithms. For example, conceptual questions might ask students to pick the direction of a resultant force, compare the magnitudes of two accelerations, or describe the motion of an object. Conceptual questions often did not give students any numerical quantities, but instead gave relative magnitudes of quantities. The procedural, mathematical, or computational questions usually involved specific quantities and required students to calculate analytic or numerical answers. For example, students might have to calculate the final momentum of a space station after an astronaut has pushed off from it, given the astronaut's initial and final momentum.

### **Final Course Grade**

Students' overall course scores were composed of their exam, homework, and participation scores. Each of the midterm exams accounted for 14% of the final score (a total of 42%), the final exam for 33%, and the homework and participation for the remaining 25%. Students' final course scores were first computed as a percentage. Based on the distribution of these course score percentages, the professor assigned course grades (A, B, C, etc.).

### **Standardized Test of Conceptual Mastery of Physics**

The Force and Motion Concept Evaluation (FMCE) (*SI*) was administered to students both at the beginning of the semester (week 1), immediately after the first intervention, and at the end of the semester (week 15). The FMCE is a research-based, validated survey instrument used to measure student understanding of introductory mechanics concepts and their conceptual mastery of introductory physics. The survey does not require any mathematics except graphical

interpretation. The precourse administration of the FMCE took place during the first week of the course in recitation. After students completed the first writing exercise for the values affirmation or control condition, they completed the FMCE for the remainder of the recitation section (approximately 35 minutes). The postcourse administration of the FMCE took place during the last week of the semester (week 15), again in recitation. Students were given the entire 50 minutes to complete the test. In both administrations, students were explicitly told that the FMCE scores would not affect their course grades in any way. As is standard procedure whenever the FMCE is analyzed, we discarded scores from students who had left two or more of the questions blank to remove students who did not take the test seriously (*S9, S13*).

## ADDITIONAL ANALYSES, RESULTS, AND DISCUSSION

### Effects of Affirmation on Outcome Measures

#### Analytical Strategy

The effectiveness of values affirmation in reducing the gender gap was assessed on three outcome measures that have previously demonstrated significant gender differences (*S9*): exam scores, final course grade, and end-of-the-semester performance on the standardized test of conceptual physics knowledge (FMCE). Separate multiple regression analyses were conducted on each outcome measure. Of particular theoretical interest in each analysis was the interaction between subject gender and condition, which tested whether the performance of women in the affirmation condition was improved relative to women in the control condition and whether the gender gap was reduced in the affirmation condition. We were also interested in the three-way interaction between gender, condition, and stereotype endorsement, which tested whether the gender  $\times$  condition interaction varied depending on students' level of agreement with the stereotype that men do better in physics than women.

To test for these critical effects, we conducted a series of regression analyses and included the following predictors in the models: subject gender (1=women, -1=men), condition (1=values affirmation, -1=control), and stereotype endorsement (mean centered for all students, following *S14*), plus all two-way interactions (gender  $\times$  condition, gender  $\times$  stereotype endorsement, and condition  $\times$  stereotype endorsement) and the three-way gender  $\times$  condition  $\times$  stereotype endorsement interaction.

When assessing the effects of identity threat on performance, it is critical to evaluate these theoretically predicted effects while controlling for prior relevant performance (*S15*, *S16*). Because previous research has shown that background preparation in math predicts physics

grades (*S17*) and accounts for a substantial amount of variance in gender differences in performance in physics (*S9*), a mean-centered measure of prior math background was included as a covariate in the analyses of exam scores and course grade. This measure was calculated first by standardizing students' SAT and ACT Math scores (provided from university records) and then using whichever of the two scores was available or, in cases where scores for both tests were available from student university records, using the average of the two. The analysis of the end-of-semester FMCE scores included the same predictors, except that the beginning-of-semester FMCE score (mean centered) was used as a covariate instead of prior math background.

Although the inclusion of background covariate variables in regression models is common, such analysis often neglects to include in the same models terms representing the interaction of those background variables with the main experimental or individual differences variables. Such neglect, however, has been shown to potentially produce serious biases in the estimation of the regression coefficients for the main experimental or individual differences variables of interest (*S18*). This is particularly the case when the individual difference variable in question, such as gender, correlates with the covariate, such as prior performance, and one wants a pure read on the interaction of gender with condition above and beyond gender's covariance with performance. Thus, in addition to the key covariate variable above (SAT/ACT Math or beginning-of-semester FMCE scores), we also included terms representing the interaction of these background variables with gender, condition, and stereotype endorsement. Specifically, the regression models predicting exam scores and final course grades contained the following 11 predictors: gender; affirmation condition; stereotype endorsement; gender  $\times$  condition; gender  $\times$  stereotype endorsement; condition  $\times$  stereotype endorsement; gender  $\times$  condition  $\times$  stereotype endorsement; SAT/ACT math; SAT/ACT math  $\times$  gender; SAT/ACT math  $\times$  condition; and

SAT/ACT math  $\times$  stereotype endorsement. The analysis of end-of-semester FMCE score was similar, but used beginning-of-semester FMCE score as the covariate.

Prior performance, as measured either by the SAT/ACT Math or by the beginning-of-semester FMCE, correlated highly with the outcomes (*S19*, *S20*, *S21*). There was some evidence for the course grade measures (but not the exam and end-of-semester FMCE scores) that it correlated differentially with outcome as a function of experimental condition, as suggested by its interaction with condition (*S19*). This differential predictiveness, however, did not manifest for any of the outcome measures when analyzing the focal group, women [all  $F_s < 2.50$ ,  $P_s > 0.11$ ]. Our analytic models take into account this heterogeneity in regression slopes by including the relevant condition  $\times$  prior performance interaction term.

Because these interaction terms involving the covariates were included to guarantee that the tests of our predicted effects were unbiased, and because those interaction terms themselves are not of direct theoretical relevance to the understanding of gender identity threat effects, results involving the covariates are presented in the notes here (*S19*, *S20*, *S21*). We also evaluated more complex models that included all possible 3-way and 4-way interactions with the background covariate variables (e.g., the SAT/ACT math  $\times$  gender  $\times$  condition interaction in the analyses of exam scores), but in no case were any 3-way or 4-way interactions significant. Thus, for simplicity, we report here the analyses that included only the 11 predictors described above.

All  $\beta$  weights reported for the regression analyses in the main article and in this supporting online document are standardized weights. All reported  $P$  levels are two-tailed.

### **Exam Scores**

An overall composite exam score, created by averaging the percent correct on each of the three midterms and the final exam, was used as the primary dependent measure for exam scores.

Fig. 1A in the main article presents mean exam scores, adjusted for prior math background (the raw, covariate unadjusted means are shown in Table S1) (S19).

Replicating past research on gender gaps in physics (S9), student gender was a significant predictor of exam scores [ $\beta=-0.23$ ,  $t(387)=-4.62$ ,  $P<0.01$ ], with men ( $M=70.7\%$ ) scoring higher than women ( $M=64.2\%$ ) across the four exams. This gender difference was present even when controlling for differences in prior math performance. As shown in Fig. 1A, the predicted gender  $\times$  condition interaction was significant [ $\beta=0.16$ ,  $t(387)=3.08$ ,  $P<0.01$ ]. Examining the simple effects, we found that for women, affirmation resulted in higher exam scores compared to those in the control condition [ $F_{1,387}=7.61$ ,  $P<0.01$ ]. There was also an unexpected yet significant tendency for men to have lower exam scores in the affirmation than control condition [ $F_{1,387}=5.37$ ,  $P=0.02$ ] (S22). This finding will be discussed in more detail in a later section entitled *Effects of Affirmation on the Performance on Men*. When the same data were examined in terms of the gender gap within each condition, the gender difference in the control condition was significant [ $F_{1,387}=36.71$ ,  $P<0.01$ ], while the gender gap was eliminated in the affirmation condition [ $F_{1,387}=2.35$ ,  $P=0.13$ ].

The benefit of values affirmation in improving the performance of women was also moderated by students' endorsement of the gender stereotype. Stereotype endorsement [ $\beta=-0.13$ ,  $t(387)=-2.31$ ,  $P=0.02$ ] and the condition  $\times$  stereotype endorsement interaction [ $\beta=0.15$ ,  $t(387)=2.64$ ,  $P<0.01$ ] were significant predictors of exam scores, but both were qualified by the predicted three-way gender  $\times$  condition  $\times$  stereotype endorsement interaction [ $\beta=0.16$ ,  $t(387)=2.74$ ,  $P<0.01$ ].

We decomposed the three-way interaction by examining the simple effect of stereotype endorsement within each level of gender and condition. As can be seen in Fig. 3A in the main

article, there was a significant negative relation between stereotype endorsement and exam scores for women in the control condition [ $\beta=-0.50$ ,  $t(387)=-3.29$ ,  $P<0.01$ ]. This result shows that among women who were not affirmed, stronger endorsement of the stereotype that men are better at physics than women was associated with poorer exam scores. This effect was eliminated by the values affirmation, as reflected in the nonsignificant relation between stereotype endorsement and exam score among women in the affirmation condition [ $\beta=0.12$ ,  $t(387)=0.94$ ,  $P=0.35$ ]. The benefit of values affirmation for women was also assessed by comparing exam scores for women at higher levels of stereotype endorsement, defined as 0.75 standard deviations (*SDs*) above the stereotype endorsement mean for all students. When stereotype endorsement was relatively high for women, values affirmation significantly improved exam scores relative to the control condition [ $t(115)=3.04$ ,  $P<0.01$ ].

For men, there was no relation between stereotype endorsement and exam scores in the control condition [ $\beta=-0.07$ ,  $t(387)=-0.92$ ,  $P=0.36$ ]. There was a marginally significant negative relation between stereotype endorsement and exam scores for men in the affirmation condition, indicating decreases in exam score as stereotype endorsement increased [ $\beta=-0.08$ ,  $t(387)=-1.70$ ,  $P=0.09$ ].

### **Final Course Grade**

The dependent variable for the analysis of final course grades was the percentage of the total points earned for the course (the raw, covariate unadjusted means and the covariate adjusted means are shown in Table S1) (*S20*).

The course grades showed a pattern of effects highly similar to the exams. This is not surprising, given that 75% of the final course grade was based on the four exams. The average final course score was higher for men ( $M=74.7\%$ ) than women ( $M=70.0\%$ ) [ $\beta=-0.18$ ,  $t(387)=-$

3.70,  $P < 0.01$ ], even after controlling for differences in prior math background. The predicted gender  $\times$  condition interaction was also significant [ $\beta = 0.14$ ,  $t(387) = 2.50$ ,  $P = 0.01$ ]. For women, affirmation improved course grade relative to women in the control condition [ $F_{1,387} = 5.26$ ,  $P = 0.02$ ], whereas the affirmation versus control difference was marginally significant for men [ $F_{1,387} = 3.23$ ,  $P = 0.07$ ]. This resulted in a significant gender gap in the control condition [ $F_{1,387} = 23.82$ ,  $P < 0.01$ ], which was eliminated in the affirmation condition [ $F_{1,387} = 1.43$ ,  $P = 0.23$ ].

As with exam scores, the benefit of values affirmation on course grades was moderated as predicted by stereotype endorsement. The two-way condition  $\times$  stereotype endorsement was significant [ $\beta = 0.15$ ,  $t(387) = 2.51$ ,  $P = 0.01$ ] as well as the predicted higher-order gender  $\times$  condition  $\times$  stereotype interaction [ $\beta = 0.16$ ,  $t(387) = 2.69$ ,  $P < 0.01$ ]. Decomposing the three-way interaction, we found that there was no relationship between stereotype endorsement and course grade for men in either the affirmation or control condition [ $\beta = -0.06$ ,  $t(387) = -1.46$ ,  $P = 0.14$ , and  $\beta = -0.04$ ,  $t(387) = -0.56$ ,  $P = 0.58$ , respectively]. For women, however, there was a significant negative relationship between stereotype endorsement and course grade in the control condition [ $\beta = -0.46$ ,  $t(387) = -2.96$ ,  $P < 0.01$ ], but not in the affirmation condition [ $\beta = 0.15$ ,  $t(387) = 1.23$ ,  $P = 0.22$ ]. Among the women relatively high in stereotype endorsement (0.75 *SDs* above the mean), the course grade was significantly higher in the affirmation condition than in the control condition [ $t(115) = 2.74$ ,  $P < 0.01$ ].

### **End-of-Semester FMCE Scores**

**Analyses of end-of-semester FMCE scores.** FMCE scores at the end of the semester provide the opportunity to directly assess learning of conceptual knowledge across the semester on identical questions administered in the 1st and 15th weeks of class. Fig. 1B in the main article presents end-of-semester FMCE scores, adjusted for beginning-of-semester FMCE scores (the

raw, covariate unadjusted means for the end-of-semester as well as beginning-of-semester FMCE scores are shown in Table S1) (S21).

Values affirmation was successful in reducing the gender gap on this measure. There was an overall gender gap across the entire sample, with men ( $M=73.4\%$ ) demonstrating better conceptual mastery at the end of the semester than women ( $M=60.4\%$ ) [ $\beta=-0.06$ ,  $t(296)=-1.05$ ,  $P=0.30$ ]. This gender effect, however, was moderated by the affirmation condition, as reflected in the predicted gender  $\times$  condition interaction [ $\beta=0.12$ ,  $t(296)=2.13$ ,  $P=0.03$ ]. As shown in Fig. 1B, for the FMCE, the reduction of the gender gap in the affirmation condition was due almost entirely to women's increased score in the affirmation condition. Specifically, women in the affirmation condition had significantly higher FMCE scores than women in the control condition [ $F_{1,296}=7.71$ ,  $P<0.01$ ], whereas the scores of men in the two conditions did not differ [ $F_{1,296}=0.08$ ,  $P=0.78$ ]. Thus, even though there was some unexpected tendency for men in the affirmation condition to perform worse than men in the control condition for the two highly correlated measures of the composite exam score and the final course score, there was no such effect for the end-of-semester FMCE data. The gender gap in the control condition was significant [ $F_{1,296}=6.23$ ,  $P=0.01$ ], whereas the gender gap in the affirmation condition was not [ $F_{1,296}=0.96$ ,  $P=0.33$ ].

Finally, the effects of values affirmation on the performance of women were once again moderated by the levels of gender stereotype endorsement, as seen in Fig. 3B in the main article. The two-way condition  $\times$  stereotype endorsement interaction [ $\beta=0.16$ ,  $t(296)=2.69$ ,  $P<0.01$ ], as well as the predicted gender  $\times$  condition  $\times$  stereotype endorsement interaction [ $\beta=0.15$ ,  $t(296)=2.45$ ,  $P=0.02$ ], was significant. Decomposing the three-way interaction, we found no significant relationship between stereotype endorsement and FMCE scores for men in either the

affirmation or control condition [ $\beta=-0.10$ ,  $t(296)=-1.41$ ,  $P=0.16$ , and  $\beta=-0.13$ ,  $t(296)=-1.33$ ,  $P=0.18$ , respectively]. For women, however, there was a significant negative relationship between stereotype endorsement and FMCE scores only in the control condition [ $\beta=-0.39$ ,  $t(296)=-2.55$ ,  $P=0.01$ ]. By contrast, for women in the affirmation condition, the relationship was not significant [ $\beta=0.22$ ,  $t(296)=1.54$ ,  $P=0.13$ ]. Among the women with higher stereotype endorsement (0.75 *SDs* above the mean), the end-of-semester FMCE scores were significantly higher in the affirmation condition than in the control condition [ $t(115)=3.01$ ,  $P<0.01$ ].

**FMCE data as an index of learning.** Hypothesized decreases in identity threat that occur following affirmation could benefit women by improving either test performance, actual learning of new concepts, or both. We cannot differentiate between the contribution of solely performance-related and solely learning-related benefits on exam scores and final course grades. However, because evaluation apprehension was low for the FMCE (students took this test in recitation both times and were told that their performance on the FMCE would not affect their grade) and because performance was assessed on identical standardized items across the semester, FMCE performance provides a better indication of actual learning effects. The improvement among women following the affirmation on the end-of-semester FMCE (with the beginning-of-semester FMCE controlled for) is, therefore, promising evidence that affirmation can produce beneficial effects through the facilitation of better learning.

To further support this idea, we assessed the effects of the first writing exercise on performance on the beginning-of-semester FMCE, which was completed immediately after the writing exercise in the first recitation. Any effects of values affirmation observed here would reflect primarily performance benefits. Neither the condition main effect [ $F_{1,304}=0.01$ ,  $P=0.96$ ] nor the gender  $\times$  condition interaction [ $F_{1,304}=0.53$ ,  $P=0.47$ ] was significant. This absence of any

affirmation effect on the beginning-of-semester FMCE scores suggests that end-of-semester FMCE effects were not solely due to better performance, but, rather, that values affirmation likely had benefits on the actual acquisition of new concepts over the semester.

### **Effects of Affirmation on Course Letter Grade**

The values affirmation was particularly beneficial in elevating women's course grades from average to above average. As can be seen in Fig. 2, a large majority of women in the control condition (55.8%) earned a grade in the C range (including C-, C, and C+), with only 23.1% earning a grade in the B range (including B-, B, and B+). The percentage of Cs was reduced to 40.8% among women in the affirmation condition, and Bs increased to 36.8%. This difference in the percentage of women getting Bs and Cs across the two conditions was statistically significant [ $\chi^2(1, N=91)=4.07, P=0.04$ ]. There was no difference in the distribution of Bs and Cs for men as a function of affirmation condition [ $\chi^2(1, N=202)=0.02, P=0.88$ ].

### **Effects of Affirmation on the Performance of Men**

There was no significant difference in the performance of men in the control and affirmation conditions on the end-of-semester FMCE score. However, we obtained an unexpected effect of affirmation on the performance of men on exam scores and course grades. The pattern was such that, although affirmation improved performance for women relative to the control condition, it decreased performance for men (this negative effect for men was significant for exam scores and marginal for course grades).

This pattern was not predicted and was not obtained in the original affirmation field experiments showing that a similar intervention closed the achievement gap in course grades between minority and nonminority middle school students (*SI0*). Because this pattern was not consistently observed either in this past research or across all measures in the present study (e.g.,

end-of-semester FMCE, letter grade distribution), it should be regarded tentatively. At the same time, though not predicted here, negative effects of affirmation have sometimes been observed in the affirmation literature (S23). More research needs to be conducted to specify when and why such negative effects may occur. In the present case, there are possible reasons why affirmation could be counterproductive for a person belonging to a group free of pervasive threats tied to their social identity. For instance, affirmation might divert their attention away from the performance domain, remind them of alternative domains where they could invest their effort, or make their sense of self-integrity less dependent on performing well. These are admittedly speculative explanations for a finding that should, as noted, be regarded tentatively. Researchers and educators concerned with this pattern might consider providing only the control exercise to men. As in the present study, it is feasible to do so without calling students' attention to the differences in exercises.

Despite the lower exam scores for men in the affirmation condition, it is important to emphasize that such effects were not observed on the end-of-semester FMCE score and, perhaps more important, that the reduction in the gender gap associated with affirmation does not simply reflect a negative effect of affirmation on men. As noted in the above presentation of the results, the affirmation significantly improved women's performance relative to those in the control group across all outcome measures. Such consistent results found for women indicate that, on the whole, affirmation closed the gender gap by directly benefitting women.

**Table S1.** Mean exam scores (mean of three midterms and final exam), final course grades, and FMCE scores from the beginning and end of the semester as a function of gender and affirmation condition. Values are shown as raw (covariate unadjusted) or covariate adjusted values. Standard deviations are in parentheses.

	Men	Women
Raw (Covariate Unadjusted) Means		
Mean Exam Score (%)		
Values Affirmation	69.4 (13.2)	65.2 (13.8)
Control	72.7 (12.5)	62.7 (11.9)
Final Exam Score (%)		
Values Affirmation	70.4 (14.2)	66.7 (15.6)
Control	73.3 (12.8)	61.3 (13.6)
Final Course Grade (%)		
Values Affirmation	73.9 (10.8)	70.5 (12.1)
Control	76.0 (10.5)	69.3 (9.9)
End-of-semester FMCE (%)		
Values Affirmation	72.7 (26.3)	63.6 (30.6)
Control	74.7 (27.3)	56.2 (25.1)
Beginning-of-Semester FMCE (%)		
Values Affirmation	39.1 (27.9)	25.9 (18.0)
Control	41.5 (29.7)	23.7 (13.4)

---

Covariate Adjusted Means		
Final Exam Score (%)		
Values Affirmation	70.4 (12.9)	68.5 (14.9)
Control	73.2 (12.8)	60.2 (14.7)
Final Course Grade (%)		
Values Affirmation	73.7 (9.7)	72.3 (11.2)
Control	75.7 (9.7)	68.2 (11.1)

---

### Supporting References and Notes

- 
- S1. R.K. Thornton, D.R. Sokoloff, *Am. J. Phys.* **66**, 338 (2006).
- S2. M. Lorenzo, C. Crouch, E. Mazur, *Am. J. Phys.* **74**, 118 (2006).
- S3. E. Mazur, *Peer Instruction: A User's Manual* (Prentice-Hall, Upper Saddle River, NJ, 1997).
- S4. C. Turpen, N.D. Finkelstein, *Phys. Rev. ST. Phys. Ed. Res.* **5**, 020101 (2009).
- S5. V. Otero, N.D. Finkelstein, R. McCray, S. Pollock, *Science* **313**, 445 (2006).
- S6. L.C. McDermott, P.S. Schaffer, *Tutorials in Introductory Physics* (Prentice-Hall, Upper Saddle River, NJ, 2002).
- S7. N.D. Finkelstein, S.J. Pollock, *Phys. Rev. ST. Phys. Ed. Res.* **1**, 010101 (2005).
- S8. S.J. Pollock, N.D. Finkelstein *Phys. Rev. ST. Phys. Ed. Res.* **4**, 010110 (2008).
- S9. L.E. Kost, S.J. Pollock, N.D. Finkelstein, *Phys. Rev. ST Phys. Ed. Res.* **5**, 010101 (2009).
- S10. G. L. Cohen, J. Garcia, N. Apfel, A. *Science* **313**, 1307 (2006).
- S11. G.L. Cohen, J. Aronson, C.M. Steele, *Per. Soc. Psychol. Bull.* **26**, 1151 (2000).
- S12. W.K. Adams, K.K. Perkins, N. Podolefsky, M. Dubson, N.D. Finkelstein, C.E. Wieman, *Phys. Rev. ST. Phys. Ed. Res.* **2**, 010101 (2006).
- S13. S.J. Pollock, N.D. Finkelstein, L.E. Kost, *Phys. Rev. ST Phys. Ed. Res.* **3**, 010107 (2007).
- S14. T. Schmader, M. Johns, M. Barquissau, *Sex Roles*, **50**, 835 (2004).
- S15. C. M. Steele, J. Aaronson, *J. Per. Soc. Psychol.* **69**, 797. (1995).
- S16. T. Schmader, M. Johns, *J. Per. Soc. Psychol.* **85**, 440. (2003).
- S17. Z. Hazari, R.H. Tai, P. M. Sadler, *Sci. Ed.* **91**, 847 (2007).
- S18. V.Y. Yzerbyt, D. Muller, C.M. Judd. *J. Exp. Soc. Psych.* **40**, 424 (2004).
- S19. There were two significant effects on mean exam scores involving the SAT/ACT Math covariate: SAT/ACT math score [ $\beta=0.43$ ,  $t(387)=9.00$ ,  $P<0.01$ ] and SAT/ACT Math score  $\times$

---

stereotype endorsement [ $\beta=-0.10$ ,  $t(387)=-2.08$ ,  $P=0.04$ ]. In all cases, higher SAT/ACT Math scores predicted higher exam scores. The interaction indicated that this relationship increased as stereotype endorsement decreased.

S20. There were three significant effects on final course grades involving the SAT/ACT Math covariate: SAT/ACT Math score [ $\beta=0.41$ ,  $t(387)=8.47$ ,  $P<0.01$ ], SAT/ACT Math score  $\times$  condition [ $\beta=0.10$ ,  $t(387)=2.24$ ,  $P=0.03$ ], and SAT/ACT Math score  $\times$  stereotype endorsement [ $\beta=-0.10$ ,  $t(387)=-2.19$ ,  $P=0.03$ ]. In all cases, higher SAT/ACT Math scores predicted higher course grades. The interactions indicated that this relationship was stronger for students in the values affirmation condition and for those who reported lower stereotype endorsement.

S21. There were two significant effects on end-of-semester FMCE involving the beginning-of-semester covariate: beginning-of-semester FMCE [ $\beta=0.66$ ,  $t(296)=8.80$ ,  $P<0.01$ ] and beginning-of-semester FMCE  $\times$  gender [ $\beta=0.17$ ,  $t(296)=2.40$ ,  $P=0.02$ ]. In all cases, higher initial FMCE scores predicted higher end-of-semester FMCE scores. The interaction indicated that this relationship was stronger for women.

S22. This unexpected effect of affirmation on men was not significant when the beginning-of-semester FMCE score, rather than the SAT/ACT Math score, was used as the covariate in the regression model to control for prior background differences [ $F_{1,296}=0.79$ ,  $P=0.37$ ]. In this model, the effect of affirmation on women was still highly significant [ $F_{1,296}=7.97$ ,  $P<0.01$ ]. All the other statistical conclusions remained the same in this analysis, including the gender  $\times$  condition interaction [ $\beta=0.15$ ,  $t(296)=2.69$ ,  $P<0.01$ ] and the gender  $\times$  condition  $\times$  stereotype endorsement interaction [ $\beta=0.22$ ,  $t(296)=3.88$ ,  $P<0.01$ ]. The beginning-of-the-semester FMCE score was a strong predictor of the exam scores [ $\beta=0.61$ ,  $t(296)=8.48$ ,

---

$P < 0.01$ ], although none of the interaction terms involving the covariate was significant in this model [all  $P$ s  $> 0.30$ ].

S23. D. Sherman, G.L. Cohen, in *Advances in Experimental Social Psychology*, M.P. Zanna, Ed. (Academic Press, San Diego, 2006), pp. 183–242.