Replicating a Self-Affirmation Intervention to Address Gender Differences: Successes and Challenges

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

*Department of Physics, University of Colorado at Boulder, Boulder, Colorado 80309 USA
†School of Education and Department of Psychology, Stanford University, Stanford, California 94305 USA
‡Department of Psychology and Neuroscience, University of Colorado at Boulder, Boulder, Colorado 80309 USA

Abstract. We previously reported on the success of a psychological intervention implemented to reduce gender differences in achievement in an introductory college physics course. In this prior study, we found that the gender gap on exams and the FMCE among students who completed two 15-minute self-affirmation writing exercises was significantly reduced compared to the gender gap among students who completed neutral writing exercises. In a follow-up study we replicated the self-affirmation intervention in a later semester of the same course, with the same instructor. In this paper, we report the details and preliminary results of the replication study, where we find similar patterns along exams and course grades, but do not observe these patterns along the FMCE. We begin to investigate the critical features of replicating educational interventions, finding that replicating educational interventions is challenging, complex, and involves potentially subtle factors, some of which we explore and others that require further research.

Keywords: gender, stereotype threat, self-affirmation, conceptual learning, introductory physics
PACS: 01.40.Fk, 01.40.G-, 01.40.gb, 01.75.+m

INTRODUCTION

For the past several years, we have been investigating gender differences in the first-semester, calculus-based mechanics course (Physics 1) at the University of Colorado. Collecting data from twelve semesters of Physics 1, we find consistent gender gaps (differences in male and female performance) on in-class exams and a conceptual survey of mechanics [1]. On average, males outscore females on exams by about 5%, and on a conceptual mechanics survey by 10% at the beginning of the semester and by 12% at the end of the semester. These gender gaps exist despite the use of interactive engagement methods [2,3], and in courses where all students show significant learning gains over the semester.

Based on our prior work on self-efficacy, showing that females are more worried and nervous about taking exams [4], we hypothesized that identity threat (described below) was inhibiting females' performance in the course. Researchers have demonstrated that identity threat can be alleviated through self-affirmation [5,6]. The results of our initial study to test the impact of a self-affirmation exercise on the gender gap in Physics 1 were reported earlier [7,8]. We found that the gender gap among students who affirmed their personal values was significantly less than the gender gap among students who did not affirm their own values.

Given the success of our initial intervention, we conducted a straightforward replication in a following semester. After the study, we identified several problems with this initial replication including a small sample size, inconsistent implementation of the intervention, and a different instructor and thus, different course practices, culture, and assessments, all of which are difficult to characterize. The results from this first replication neither supported nor discredited the original results, but were inconclusive.

These inconclusive results lead us to replicate a second time paying careful attention to issues of fidelity. We replicated the self-affirmation study in a later semester of the same course with the same instructor as in the original study. In this paper, we report the details and preliminary results of this second replication study. Our initial findings are that there are similar results in terms of reducing the gender gap on exams, but we do not observe these same results on a conceptual mechanics survey.

IDENTITY THREAT AND SELF-AFFIRMATION

Identity threat [9] occurs when a person realizes that he/she can be negatively judged based on his/her social identity (sex, ethnicity, social class, etc.). This awareness and threat of being negatively judged has been shown to undermine the performance of

threatened students [10,11]. Several researchers have found that identity threat can be alleviated through self-affirmation [5,6]. Self-affirmation is a process through which a person affirms his or her overall sense of self-worth and integrity [12]. Work on self-affirmation assumes that people are motivated to maintain a positive sense of overall integrity, identity and worth. When our integrity or identity is threatened, we seek ways to resolve the threat. Because it is often difficult (or even impossible) to resolve the specific identity threat (e.g., that "girls can't do science"), individuals can affirm a more general sense of worth and integrity (e.g., "I'm a good person") or a specific, but unrelated, aspect of their identity (e.g., "I'm good at music"), which will help to protect them from the threat. In a laboratory experiment with college students, Martens, *et al*. [6] found that females who were given the opportunity to write about a characteristic of themselves that they valued performed better on a subsequent difficult math test than females who wrote about a characteristic that they did not value. Similar results were found by Cohen, *et al.* [5] comparing the school achievement of middle-school African-Americans who did and did not affirm their personal values.

STUDY DESIGN

In both the original study (Study 1) and the replication study reported here (Study 2) we conducted a randomized experiment in order to test the impact of self-affirmation on the performance of males and females in Physics 1. Physics 1 is the first semester of the three-semester, calculus-based introductory physics sequence for science majors and engineers. In each semester, there were about 600 students in the class. Peer Instruction and ConcepTests [2] were employed during lecture and students worked through *Tutorials in Introductory Physics* [3] during recitation. In terms of the curriculum, these courses were nearly identical to each other and to previous semesters of Physics 1. The lead instructor, who taught during both studies, was experienced in interactive engagement methods.

The experimental design of the two studies was the same. Students were randomly assigned to complete either a self-affirmation exercise where they wrote about values that were important to them, or a control exercise, where they wrote about values that were important to others. These writing exercises took place in the first and fourth weeks of the course and took about 15 minutes to complete. Students took the Force and Motion Concept Evaluation (FMCE) [13] at both the beginning and end of the semester, as well as three in-class, midterm exams and a final. The exams in the two studies were similar in character, but not identical. Additional details of the writing exercises and study design can be found in Ref. 7. In both studies, while the instructor was blind to individual students' condition, he was aware of the purpose of the studies.

In addition to the writing exercises, students were also asked to fill out an online, optional survey early in the semester that assessed self-efficacy and identity in physics [4]. While the exact content of the survey in the two studies differed, both surveys contained the following gender stereotype item of interest to this intervention: *According to my own personal beliefs, I expect men to generally do better in physics than women*. Students responded to this statement on a 5-point, Likert-like scale from strongly disagree to strongly agree.

In both Study 1 and 2 we were interested in the impact of the self-affirmation intervention on student performance on exams and the FMCE. We hypothesized that females who completed the self-affirmation exercise would perform better on exams and the post-FMCE than females who completed the control exercise. Also, because we expected that there would be no significant impact of the self-affirmation exercise on male performance, we expected that the gender gap among students who completed the self-affirmation exercise would be smaller than the gender gap among students who completed the control exercise. These results were found in Study 1 [7].

RESULTS

Just as in Study 1, students were only included in the analyses if they satisfied the following conditions: 1) they completed both writing exercises, 2) they took the final exam, 3) they responded to the gender stereotype item, and 4) we had data for either the SAT- or ACT-Math test. This left us with a total of 363 students (60% of the class) [14]. The self-affirmation group included 168 males and 74 females, while the control group was comprised of 86 males and 35 females. Of the 363 students in the final sample, only 283 took both the pre- and post-FMCE, and are thus included in that analysis (affirmation: 129 males, 60 females; control: 65 males, 29 females). By design, there are more students in the self-affirmation than the control condition (60% versus 40%). This design ensured that more students would receive the potentially beneficial self-affirmation exercise. There were no significant differences on any prior factors [15] between the self-affirmation and control groups, by gender.

We used a multiple regression analysis to test the effect of the self-affirmation exercise on students' exam and post-FMCE scores [16]. This standard

statistical approach allowed us to create a model of scores that accounts for other factors in addition to gender and experimental condition. In these models, we were specifically interested in the interaction between gender and condition. The significance of this interaction indicates that the gender gap in the self-affirmation condition is not equal to the gender gap in the control condition. The final models that we present include the variables: *gender*, *condition*, *stereotype endorsement*, and a measure of student background (either a prior *math score* [17] used in the analysis of exams or the *FMCE pre-test* used in the analysis of the FMCE post-test), as well as all two-, three-, and four-way interactions between these variables. For all analyses we report two-tailed significance tests.

Exams

Students took three midterm exams and a final exam over the course of Physics 1. Students' scores on all four exams were averaged to get an average exam score for each student. Students' average exam scores were regressed on *gender*, *condition*, *stereotype (ST) endorsement*, prior *math score*, and all interactions. The variables included in the final model account for a significant fraction (39%) of the variance in exam scores (p<0.01). Just as in Study 1, females in the affirmation condition outscored females in the control condition (p=0.03). Additionally, the gender gap (controlling for prior *math score* and *ST endorsement*) in the self-affirmation group is 3% ± 2%, and is smaller than in the gender gap in the control group, which is 9% ± 3% (uncertainties are standard errors). The *gender*×*condition* interaction, however, is not significant (p=0.10), meaning that the gender gaps in the affirmation and control groups are not significantly different. While the interaction is not statistically significant, the patterns in this replication study are similar to the results in the original study. Further, when data from both studies are combined, the *gender*×*condition* is significant (p<0.01), and the *gender*×*condition*×*semester* interaction is not significant (p=0.46), indicating that the *gender*×*condition* interaction does not vary significantly by semester. Results from both studies are shown in Figure 1, which plots the average exam score (adjusted by prior *math score* and *ST endorsement*) by gender and condition.

In Study 1, we found a significant three-way interaction between *gender*, *condition*, and *ST endorsement*, which indicated that the *gender*×*condition* interaction varied depending on how much students endorsed the stereotype. That is, the more females endorsed the gender stereotype the worse they performed on exams, unless they completed the self-affirmation. Female students who completed the self-affirmation exercises were buffered from the negative impact of stereotype endorsement [7]. In Study 2, though we observed the same general patterns, i.e., endorsing the gender stereotype was negatively related to exam performance for females unless they had self-affirmed, the three-way interaction was not statistically significant (p=0.27). When combining the data from both studies, the *gender*×*condition*×*ST endorsement* interaction is significant (p<0.01).

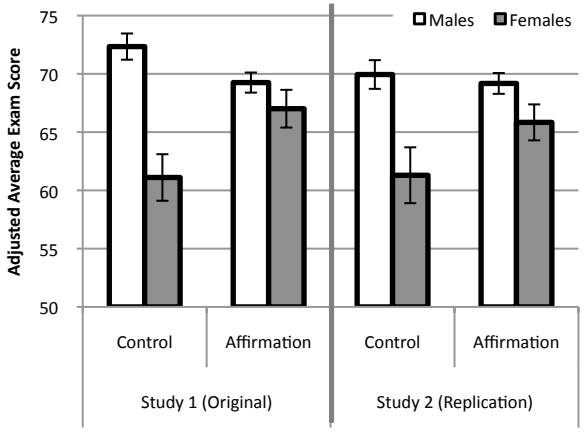

FIGURE 1. Adjusted average exam scores for males and females by condition in Study 1 and 2.

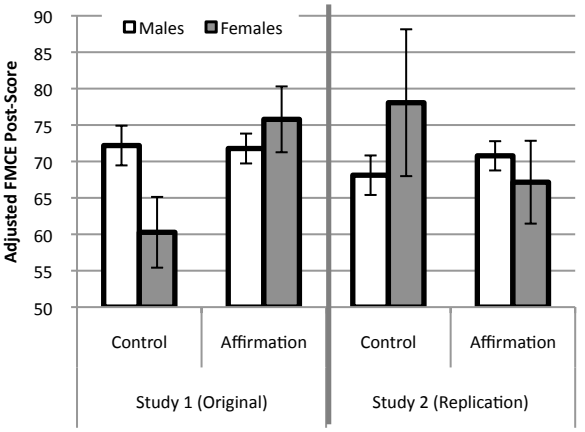

FIGURE 2. Adjusted FMCE post-scores for males and females by condition in Study 1 and 2.

FMCE Post-test

Students' FMCE post-test scores were regressed on *gender*, *condition*, *ST endorsement*, *FMCE pre-test*, and all interactions. The variables included in the final model account for a significant fraction (42%) of the variance in FMCE post-test scores (p<0.01). In Study 2, the *gender*×*condition* interaction is not significant (p=0.26), meaning that the post-FMCE gender gaps in the affirmation and control groups are not significantly

different. Unlike with the exam results (where we see similar patterns in Study 2 as in Study 1), the patterns from Study 1 are not replicated in Study 2 for FMCE post-test scores. Once *FMCE pre-test* scores and *ST endorsement* are controlled for, females in the control condition outscore males in the control condition (p=0.02) and females in the affirmation condition (p=0.02), as seen in Figure 2. Additionally, the three-way interaction, *gender*×*condition*×*ST endorsement*, is not significant (p=0.54), and the pattern of relationships between FMCE post-test score and *ST endorsement* are dissimilar to those in Study 1.

DISCUSSION

Our preliminary analysis of the outcomes in this replication study indicates that, just as in the original study, females who completed the self-affirmation exercises outperformed control females on exams. Additionally, the pattern of exam results in terms of the gender gap match those of the original study, i.e. the gender gap in the affirmation condition was reduced compared to the gender gap in the control condition. Further, when data from the original and replication studies are combined, the effect of the intervention on exam performance and the moderation of the effect by stereotype endorsement are both statistically significant, suggesting that small sample size may be an issue in the replication study. Though not discussed here, we also note that females in the affirmation condition had average course grades that were about 1/3 of a letter grade higher (p=0.03) than the course grades of females in the control condition.

The replication results on the FMCE are surprising and puzzling. We may have expected a higher chance of replicating the results of the original study on the FMCE since it is a nationally-normed, standardized instrument, and it is the assessment on which we observe the largest gender differences. We are still working towards understanding these results.

This replication study suggests that replicating educational interventions is difficult and requires more research into the factors and conditions that support replication. We have begun to explore the impact of some of these factors. For instance, we have looked for differences between the populations of students in the two studies. Females in Study 2 have lower math scores than females in Study 1, and we have evidence from Study 2 that math score moderates the effect of the affirmation for females, such that the affirmation effect is stronger for females with higher math scores. These data suggest that the affirmation may be most beneficial for better-prepared females, who were a smaller fraction of the females in Study 2 compared to Study 1. There are also other factors that require further research: the impact of different pre-course surveys, differences in instructor practices and classroom culture, differences in how the intervention is implemented, and differences between graded exams versus voluntary, un-graded conceptual surveys. We argue that the self-affirmation intervention shows promise for reducing gender differences in introductory physics performance, but further research is required to understand the conditions and factors that support its success.

ACKNOWLEDGMENTS

This work is supported by iSTEM (#0833364) and DRL (#0910373). We thank the PER at Colorado group, the faculty that were involved, and Lindsey Newnes and Natalie Golaszewski for their assistance with data collection.

REFERENCES

1. L. E. Kost, et al., *Phys. Rev. ST Phys. Ed. Res.* **5**, 010101 (2009).
2. E. Mazur, *Peer Instruction: A User's Manual*, Upper Saddle River, NJ: Prentice-Hall, 1997.
3. L. C. McDermott & P. S. Schaffer, *Tutorials in Introductory Physics*, Upper Saddle River, NJ: Prentice-Hall, 2002.
4. L. E. Kost, et al. in *2009 PERC Proceedings*, ed. by M. Sabella, et al., AIP Conference Proceedings 1179, AIP, Melville, NY, 2009, pp. 177.
5. G. L. Cohen, et al., *Science*, **313**, 1307 (2006).
6. A. Martens, *J. Exp. Soc. Psych*. **42**, 236 (2006).
7. A. Miyake, et al., *Science*, **330**, 1234 (2010).
8. L. E. Kost-Smith, et al., in *2010 PERC Proceedings*, ed. by M. Sabella, et al., AIP Conference Proceedings 1289, AIP, Melville, NY, 2010, pp. 197.
9. C. M. Steele, et al., in *Advances in Experimental Social Psychology*, ed. by M. P. Zanna, Academic Press, NY, 2002, pp. 379.
10. S. J. Spencer, et al., *J. Exp. Soc. Psych.* **35**, 4 (1999).
11. C. M. Steele & J. Aronson, *J. Pers. Soc. Psych.* **69**, 797 (1995).
12. C. M. Steele, in *Advances in Experimental Social Psychology*, ed. by L. Berkowitz, Academic Press, NY, 1988, pp. 261.
13. R. K. Thornton & D. R. Sokoloff, *Am. J. Phys*. **74**, 118 (2006).
14. As is usually the case, students who participated in the study had higher course grades than those who did not.
15. Prior factors included SAT-Verbal, SAT-Math, SAT total score, ACT-Reading, ACT-English, ACT-Math, ACT-Scientific Reasoning, ACT total score, years of high school physics and calculus, and high school GPA.
16. L. S. Aiken & S. G. West, *Multiple Regression: Testing and Interpreting Interactions*, Thousand Oaks, CA: Sage, 1991.
17. The math score is an average of normalized SAT- and ACT-Math scores.