Can Short Psychological Interventions Affect Educational Performance? Revisiting the Effect of Self-Affirmation Interventions

Authors: M. Serra-Garcia¹*, K. Hansen¹, U. Gneezy¹,²

Accepted Version, Published in Psychological Science 31 (7), 865-872 (2020)

Affiliations:
¹ UC San Diego, Rady School of Management
² CREED, University of Amsterdam
*Correspondence to: mserragarcia@ucsd.edu

Abstract: Large amounts of resources are spent annually to improve educational achievement and to close the gender gap in sciences with typically very modest effects. Miyake et al., (Science, 2010) introduced a 15-minute “self-affirmation” intervention and reported a dramatic reduction in the gender gap. This paper re-analyzes the original data and finds several critical problems. First, the self-affirmation hypothesis states that women’s performance will improve. However, the data showed no improvement for women. There was an interaction effect between self-affirmation and gender, caused by a negative effect on men’s performance. Second, the findings were based on covariate-adjusted interaction effects, which imply that self-affirmation only reduced the gender gap for the small sample of men and women who did not differ in the covariates. Third, specification-curve analyses with over 1,500 possible specifications showed that fewer than a quarter yielded significant interaction effects and fewer than three percent showed significant improvements among women.
Keywords: self-affirmation, gender gap, education.
Education in the United States is far from reaching desired outcomes (Swanson, 2009). The problem is not in the number of policy ideas that are tried; it is in the success of these policies. For example, in a large-scale field experiment to study whether monetary incentives can improve educational outcomes, some students were paid to read books, others for performance on interim assessments (Fryer, 2011). Despite spending $9.4 million, the impacts were statistically indistinguishable from zero (see also, Springer et al., 2009, Gneezy et al., 2011, Fryer et al., 2012).

A particular focus of many programs is to reduce the gender gap in science, technology, engineering, and mathematics (STEM). Given the limited effectiveness and high cost of most interventions, the success of a very simple psychological manipulation tested by Miyake et al. (2010) is impressive. In the intervention, called values affirmation, 399 students in a college-level introductory physics class either wrote about their most important values or their least important values (in the control condition), twice for 15 minutes within the first four weeks of the 15-week course.

Miyake et al. (2010) reported that the intervention reduced the gender gap in physics exams at the 1% significance level. While studies using monetary incentives to improve performance find average effects ranging from –0.031 to 0.079 standard deviations (SD) (Fryer, 2011), the effects in Miyake et al. (2010) ranged from 0.12 to 0.19 SD. Similar strong results of the self-affirmation intervention have been found in schools (Cohen et al., 2006), even in two-year and 7-to-9-year follow up studies (Cohen et al., 2009; Goyer et al., 2017).

Miyake et al. (2010) is a widely cited study that has had a substantial impact on the field of STEM education (Nisbett, 2010; Hanselman et al., 2017) and the broader use of
self-affirmation interventions. However, other studies (Borman, 2012; Lauer et al., 2013; Dee, 2015; De Jong et al., 2016; Bratter, Rowley and Chukhray, 2016; Hanselman et al., 2017; Hoffman and Kurtz-Cortes 2019) failed to replicate the findings.

We re-analyzed the Miyake et al. (2010) data, and examined it at a conceptual and a statistical level. At the conceptual level, value affirmation theory starts by observing the common stereotype that men are better at math and science than women. This may generate increased academic pressure for women who subscribe to this stereotype and possibly negatively affect their academic performance. To offset this, the theory proposes that a values affirmation task, in which people reflect on self-defining values, can buffer people against such psychological threat. The resulting hypothesis is that (some) women will improve as a result of the intervention. This theory has no prediction for male students. However, the original data in Miyake et al. (2010) showed that the observed reduction in the gender gap was a result of the interaction effect of the intervention on the performance of men and women. We examined the results for male and female students separately.

From a statistical perspective, the analysis in Miyake et al. (2010) was based on covariate-adjusted means. The interaction effect of values affirmation must be thus interpreted as conditional on a given level of prior performance and stereotype endorsement (Cochran, 1957; Miller and Chapman, 2001). However, since prior performance and stereotype endorsement on average differed substantially for men and women, the reported effect was restricted to a small portion (28%) of the sample: those females with the same performance and stereotype endorsement as males.

Miyake et al. (2010) reported the results of one empirical model specification. Using specification-curve analyses (Simonsohn et al., 2015; see also, Steegen et al., 2016),
we examined the robustness of the effect of values affirmation on the gender gap and on females’ performance.

Finally, Miyake et al. (2010) argued that results were consistent with theoretical accounts of stereotype threat because the intervention improved the performance of women displaying high stereotype endorsement. To show this, they compared women whose stereotype endorsement is either 0.75 SD higher than the mean (high endorsement) or 0.75 SD lower than the mean (low endorsement). We re-examined these heterogeneous effects considering the responses of all women along the 5-point scale measuring stereotype endorsement.

We also ran a small replication study of Miyake et al. (2010) in a physics class at UC San Diego, because replication is an important part of social science (Simmons et al., 2011; Open Science Collaboration, 2012, 2015; Simonsohn, 2016; Camerer et al., 2016), and report the results of our suggestive study in the SOM.

Method

Revisiting Miyake et al. (2010)

The self-affirmation intervention worked as follows. Students in the treated group wrote about personally important values (such as friends and family). Students in the control group selected their least important values from the same list and wrote why these values might be important to other people. Thus, all students wrote about values and their importance, but the exercise was self-relevant only for the affirmation group (the students in the treated group). In Miyake et al. (2010) this 15-minute writing exercise was conducted during class once in week one and once in an online homework assignment during week
four of the semester of an introductory physics class in college. This brief exercise was not related to the subject matter of the course.

The theory is that “Values affirmation, in which people reflect on self-defining values, can buffer people against such psychological threat. When they affirm their core values in a threatening environment, people reestablish a perception of personal integrity and worth, which in turn can provide them with the internal resources needed for coping effectively…”, according to Miyake et al. (2010). The theory does not predict a negative effect on men but only a positive one for women.

We solicited the raw data from the original study by Miyake et al. (2010) to better understand their results. These data included three continuous measures of performance: mean exam scores, scores on a standardized physics test (the Force and Motion Concept Evaluation or FMCE), and the final course score, based substantially (75%) on the exam scores.

**Results**

We found that, using the original data by Miyake et al. (2010), values affirmation had no significant effect on average female performance (covariate-unadjusted). If at all, it had a negative effect on male performance in some dimensions. Covariate-unadjusted means are reported in Table S1 in the supplementary materials of Miyake et al. (2010).

Values affirmation had no significant effect on the mean exam score of females (Cohen’s $d=0.19$) ($t(114)=1.00, p=0.318$). However, the performance of male students was significantly lower in the values affirmation condition relative to the control condition
(Cohen’s $d=-0.25$) ($t(281)=-2.06, p=0.040$). The interaction effect was significant ($\beta_t=0.43, t(395)=1.97, p=0.050$), but it was driven by a drop in male performance.

Values affirmation did not have a significant effect on male or female scores on a standardized physics test (FMCE score) at the end of the semester. The scores of female students did not differ significantly across the values affirmation and the control condition (Cohen’s $d=0.31$) ($t(94)=1.28, p=0.21$). Male students did not obtain significantly different scores across conditions either (Cohen’s $d=-0.09$) ($t(210)=-0.54, p=0.587$). The interaction effect between values affirmation and gender was not significant ($\beta_t=0.34, t(304)=1.40, p=0.164$).

Fig. 1 presents the results obtained by Miyake et al. (2010) for final course scores. The final course scores obtained by females and males were not affected by the values affirmation intervention, and the interaction effect was not significant either. All details of the regression analysis are reported in the SOM.
Fig. 1: Mean final course scores by condition and gender in Miyake et al (2010). Confidence interval bars are indicated.

The interpretation of covariate-adjusted effects

These results contrast with the covariate-adjusted results reported by Miyake et al. (2010). Miyake et al. (2010) reported a significant positive effect of the values affirmation intervention on the performance of females using covariate-adjusted means. The adjustment was based on two covariates: (1) prior performance, measured by the SAT score for the mean exam score and the course score, and the beginning-of-semester FMCE score for the FMCE score at the end of the semester; (2) and stereotype endorsement, measured by students’ agreement with the statement “According to my own personal beliefs, I expect men to generally do better in physics than women” on a 5-point scale ranging from
“strongly disagree” to “strongly agree.” The authors argued that it is critical to assess the effects of the self-affirmation intervention “controlling for prior relevant performance” (Supplementary Materials, p. 13, Miyake et al. 2010) and that background variables, including stereotype endorsement, should be included and interacted with the experimental difference variables. They pooled male and female students together and examined the effect of the self-affirmation intervention, including variables for gender, stereotype endorsement and prior performance, as well as interaction terms between these.

The interpretation of covariate-adjusted effects is that there was a reduction in the gender gap for a population of men and women who have the same SAT score (and level of stereotype endorsement). We demonstrate the interpretation of covariate-adjusted effects formally in the SOM. Is the covariate-adjusted effect the effect we are interested in? It could be if the distributions of SAT scores and stereotype endorsement for men and women had similar means. However, there was a significant difference in prior performance of men and women (for SAT scores, t(397)=2.62, p=0.01; for beginning-of-semester FMCE scores, t(306)=4.80, p<0.01) and in stereotype endorsement ($\chi^2(4)=41.64$ p<0.01). Hence, the effect that was estimated with covariate adjustment was only relevant for the subset of women who had the same prior performance and stereotype endorsement as men. In Miyake et al. (2010), this was 56% of the sample, considering only SAT or ACT scores. Including stereotype endorsement, only 28% of the sample featured male and female students with the same SAT scores and stereotype endorsement.

The relevant question is whether values affirmation had a significantly positive effect on average female performance. We examined this effect as part of the specification-curve analysis described next.
Evaluating Analytical Approaches: Specification-Curve Analysis

As we have discussed above, adding covariates such as SAT scores, and their interactions, changes the interpretation of the coefficients, and also invalidates standard linear model analysis. The researcher may be interested in understanding how robust the effect of self-affirmation interventions is to the inclusion of different covariates in the regression model, as well as other decisions regarding the data analysis (e.g., Steegen et al., 2016; Simonsohn et al., 2015). To investigate this, we conducted a specification curve analysis (Simonsohn et al., 2015).

The regression model used in Miyake et al. (2010) included 11 independent variables. In addition to gender ($F_i$), affirmation condition ($Z_i$) and prior performance ($S_i$), it included stereotype threat, which we denote as $T_i$. The specification included 2-way and 3-way interaction effects of these variables, and was as follows:

\[
Y_i = \beta + \beta_Z \times Z_i + \beta_F \times F_i + \beta_{FZ} \times F_i \times Z_i + \beta_S \times S_i + \beta_{SZ} \times S_i \times Z_i + \beta_{SF} \times S_i \times F_i \\
+ \beta_T \times T_i + \beta_{TZ} \times T_i \times Z_i + \beta_{TF} \times T_i \times F_i + \beta_{TS} \times T_i \times S_i + \beta_{TZF} \times T_i \times Z_i \times F_i + \epsilon_i
\]

Miyake et al. (2010) focused on the average exam score as their main outcome of interest, but also included final course score and score in the FMCE at the end of the semester in some of the analyses. We investigated how the gender gap in academic performance (values affirmation $X$ gender interaction) changed under the following alternatives: (1) using the three different dependent variables mentioned above, all of which measure academic performance; (2) including stereotype threat as a covariate, interacted or not with other covariates; (3) using different definitions of stereotype threat (as a continuous variable, or a dummy for those above median); (4) including prior performance, interacted
or not with other covariates; (5) using different variables and definitions of prior performance (SAT/ACT score or FMCE score at the beginning at the semester, continuous or as a dummy for those performing above median); (6) allowing for different exclusions of missing observations; (7) and estimating robust standard errors. These decisions yielded a total of 1566 unique estimated interaction effects of values affirmation and gender. Further details of the analysis and results are provided in the SOM. Since the relevant question is whether the values affirmation has a significantly positive effect on average female performance, we also conducted the specification curve analysis for the effect of values affirmation on female students.

Panel A of Figure 2 plots the coefficient and confidence interval of the interaction of values affirmation and gender at the sample average. The dependent variable was standardized, such that the coefficient can be interpreted in standard deviations. Panel B of Figure 2 plots the coefficient and confidence interval of the effect of values affirmation on female students. Each figure also shows the distribution of effects for each of the researcher’s decisions, described on the left-hand side.

We replicated the $t$-statistic for the interaction effect in the original work of Miyake et al. (2010), 3.08, and indicate where it lies on the specification curve. It was the 15$^{th}$ highest out of 1566 specifications. The interaction effect was not statistically significant ($p>0.05$) in 1205 specifications. That is, in 77% of specifications the effect of values affirmation was not different between female and male students. As shown in Figure 2, a key decision in obtaining a significant interaction effect was to include the 3-way interaction between values affirmation, gender and stereotype threat endorsement. If it was not included, 91% of specifications (1000 out of 1098) were not significant. By contrast,
of the 361 significant interaction effects, 73% include the 3-way interaction. This indicates that, not only was covariate-adjustment needed, but that a 3-way interaction term was needed to find a significant reduction of the gender gap in academic performance.

Within the group of female students, the treatment was assigned randomly and females in the treatment and control conditions do not vary in their prior performance (for SAT scores by condition, $t(114)=-0.79, p=0.4329$; for beginning-of-semester FMCE score by condition, $t(94)=-0.65, p=0.5187$) or stereotype endorsement ($\chi^2(4)=2.46, p=0.653$). Hence, covariate adjustment could be used to reduce variance (Miller and Chapman, 2001) without reducing the effect of values affirmation to a specific group of students. In the data shared with us by Miyake et al. (2010), SAT/ACT grades were standardized for men and women and, with the data available, we could not recode them and use them within the sample of women. Stereotype endorsement and FMCE scores at the beginning of the semester were sample centered. They could be recoded and we used these as covariates within the sample of women. The specification-curve is shown in Panel B of Figure 2. The effect of values affirmation on female students was not significant in 97.0% of specifications.
Panel A. Effect of Values Affirmation on the Gender Gap - Specification Curve

Panel B. Effect of Values Affirmation on Female Students – Specification Curve

NOTES: (a) Std Coefficient is the average interaction effect of values affirmation and gender (Panel A) or values affirmation on female students’ performance (Panel B), where performance is always measured on the standardized dependent variable (DV). (b) To conduct the analysis in Panel B, SAT/ACT grade cannot be used because it is standardized for male and female students in the original data and its raw value cannot be recovered.

Figure 2. Specification Curve Analysis
Miyake et al. (2010) argued that “values affirmation was particularly beneficial for women who tended to endorse the gender stereotype” (p. 1236). They split the sample between high and low stereotype endorsement based on whether the student was +/- 0.75 SD from the mean. We examined this finding in further detail to understand the role of this particular sample split. Stereotype endorsement was based on agreement with the statement “According to my own personal beliefs, I expect men to generally do better in physics than women,” with responses ranging from strongly disagree to strongly agree on a 5-point scale. Out of 96 female students without missing information, 7 (4 treated, 3 control) agreed with the stereotype, 10 (4 treated, 6 control) neither agreed nor disagreed, 24 disagreed (14 treated, 10 control) and 55 (33 treated, 22 control) strongly disagreed. Values affirmation had a positive and significant effect on the exam scores of women who agreed with the stereotype (4 women) and for women who neither agreed nor disagreed (a total of 8 women). Surprisingly, for women who strongly disagreed with the statement we observed negative effects of values affirmation on their exam score (33 women).

Using a regression analysis of performance on the female sample, controlling for stereotype endorsement and prior performance, the interaction between Values Affirmation and Stereotype Endorsement was significantly positive. Yet, this only indicates that the effect of values affirmation was more positive on the students with higher stereotype endorsement. It should not be interpreted as being a positive effect for everyone, and more strongly positive for women. As described above, the effect was negative for a large group of students. Detailed results are presented in the SOM.
Discussion

Understanding how to improve academic performance, in particular for struggling students, is an important challenge for social scientists and policy makers. Miyake et al. (2010) offered a remarkably strong and cost-effective way of doing that. Our investigation was motivated by a desire to better understand the psychological instrument in Miyake et al. (2010), and hopefully use it in large-scale interventions. A closer look at the data of Miyake et al. (2010) revealed two critical problems. The self-affirmation hypothesis only predicted an effect on women. However, their conclusions were based on the significant interaction effect between the self-affirmation intervention and gender. This effect was driven by a reduction in the performance of men, and not by an improvement in the performance of women. Further, the conclusions from the original study were highly sensitive to the empirical specification. Miyake et al. (2010) wrote that “values affirmation is a promising intervention that can help reduce the gender achievement gap in physics” (p. 1237). The statistical analysis presented here suggests that this conclusion was not supported by the data. Taking the evidence together, the results were supportive of a null effect. It is important to carefully understand the effects of such interventions because promoting ineffective interventions is costly in terms of resource allocation and negatively affects the success of policy makers’ attempts to reduce the gender achievement gap.

Author Contributions

U. Gneezy and M. Serra-Garcia developed the study concept. K. Hansen analyzed the regression models in Miyake et al. (2010) and provided the interpretation of covariate-
adjusted effects. M. Serra-Garcia conducted the reanalysis of the data in Miyake et al. (2010). All authors drafted and revised the manuscript.

Acknowledgments

We thank Tiffany Ito for sharing the materials and data, and Anna Dreber, Leif Nelson, and Johannes Haushofer for valuable comments.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Supplemental Material

This project was conducted under IRB #130701XX. Additional supporting information is provided in the SOM.

Open Practices

STATA codes used for different analyses are available on OSF: https://osf.io/gzq9k/. Data was obtained from the authors of the original study (Miyake et al., 2010).

References:


Supplementary Information for

Can Short Psychological Interventions Affect Educational Performance? Revisiting the Effect of Self-Affirmation Interventions

Marta Serra-Garcia, Karsten Hansen and Uri Gneezy

Correspondence to: mserragarcia@ucsd.edu

This PDF file includes:

Re-Analysis of Miyake et al. (2010)
Tables S1 – S3
Female Performance, Stereotype Threat and Values Affirmation
Figure S1
Interpreting Covariate-adjusted Effects: Details
Specification Curve Analysis
Figure S2 and Table S4
Suggestive Replication Study
Figure S3 and Tables S5 – S6
Re-Analysis of Miyake et al. (2010)

Below we first show the complete regression models underlying the effects summarized on page 6 of the paper. The data were obtained directly from Tiffany Ito and are the same as reported in Miyake et al. (2010). Table S1 shows the results of linear regression models of the effect of values affirmation on the FMCE score (at the end of the course), the mean exam score and the course score. For each performance measure, Table S1 shows the effect of values affirmation on males and females separately (columns (1)-(6)), and the interaction between values affirmation and gender, when all subjects are pooled together (columns (7)-(9)).

Table S2 takes an alternative approach to the use of covariates, which is to examine the interaction effect between values affirmation and gender by quartiles of the ability distribution. The results below show that in 11 out of the 12 specifications, the interaction between gender and treatment is not significant.

Table S3 focuses on the sample of female students, and reports the coefficient estimates of the effect of the values affirmation condition on female performance, controlling for stereotype endorsement and prior math performance, mean exam score, course score and end-of-semester FMCE score.

Throughout, performance measures (exam scores and test scores) are standardized. Therefore, the coefficient of values affirmation can be interpreted as a “standardized coefficient” with respect to the dependent variable, and in standard deviations for this variable.
Table S1. Effect of values affirmation on student performance, without covariates (students in Miyake et al. (2010)).

<table>
<thead>
<tr>
<th></th>
<th>Exam Score</th>
<th>Course Score</th>
<th>FMCE Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male Values</td>
<td>0.037</td>
<td>-0.246**</td>
<td>0.191</td>
</tr>
<tr>
<td>Female Values</td>
<td>0.046</td>
<td>-0.186</td>
<td>0.075</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male Values</td>
<td>0.029</td>
<td>-0.186</td>
<td>0.075</td>
</tr>
<tr>
<td>Female Values</td>
<td>0.120</td>
<td>-0.751***</td>
<td>0.751***</td>
</tr>
</tbody>
</table>

Note: Values Affirmation is a dummy variable that takes value 1 if the subject completed the values affirmation exercise and 0 if the subject completed the control exercise. Female is a dummy variable that takes value 1 if the subject is female, 0 if male. Standard errors in brackets.

- ** p<0.01, * p<0.05, * * p<0.01, * * * p<0.001.
Table S2. Effect of values affirmation on student performance, by quartile of the distribution of ability (students in Miyake et al. (2010))

<table>
<thead>
<tr>
<th>Distribution of Ability (Beginning-of-Semester FMCE Score)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st quartile</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Values Affirmation</td>
<td>-0.514**</td>
<td>-0.065</td>
<td>0.250</td>
<td>-0.153</td>
</tr>
<tr>
<td></td>
<td>[0.248]</td>
<td>[0.268]</td>
<td>[0.239]</td>
<td>[0.179]</td>
</tr>
<tr>
<td>Female</td>
<td>-0.401</td>
<td>-0.582*</td>
<td>-0.544</td>
<td>-1.072***</td>
</tr>
<tr>
<td></td>
<td>[0.348]</td>
<td>[0.328]</td>
<td>[0.358]</td>
<td>[0.243]</td>
</tr>
<tr>
<td>Female X Values Affirmation</td>
<td><strong>0.651</strong></td>
<td><strong>-0.215</strong></td>
<td><strong>0.001</strong></td>
<td><strong>0.457</strong></td>
</tr>
<tr>
<td></td>
<td>[0.450]</td>
<td>[0.439]</td>
<td>[0.436]</td>
<td>[0.320]</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.403**</td>
<td>0.073</td>
<td>0.274</td>
<td>1.168***</td>
</tr>
<tr>
<td></td>
<td>[0.198]</td>
<td>[0.228]</td>
<td>[0.189]</td>
<td>[0.138]</td>
</tr>
<tr>
<td>2nd quartile</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Values Affirmation</td>
<td>-0.528**</td>
<td>-0.028</td>
<td>0.302</td>
<td>-0.168</td>
</tr>
<tr>
<td></td>
<td>[0.245]</td>
<td>[0.278]</td>
<td>[0.238]</td>
<td>[0.185]</td>
</tr>
<tr>
<td>Female</td>
<td>-0.420</td>
<td>-0.578*</td>
<td>-0.431</td>
<td>-0.912***</td>
</tr>
<tr>
<td></td>
<td>[0.345]</td>
<td>[0.341]</td>
<td>[0.355]</td>
<td>[0.252]</td>
</tr>
<tr>
<td>Female X Values Affirmation</td>
<td><strong>0.611</strong></td>
<td><strong>-0.230</strong></td>
<td><strong>-0.093</strong></td>
<td><strong>0.371</strong></td>
</tr>
<tr>
<td></td>
<td>[0.446]</td>
<td>[0.455]</td>
<td>[0.432]</td>
<td>[0.332]</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.284</td>
<td>0.123</td>
<td>0.245</td>
<td>1.149***</td>
</tr>
<tr>
<td></td>
<td>[0.196]</td>
<td>[0.237]</td>
<td>[0.188]</td>
<td>[0.143]</td>
</tr>
<tr>
<td>3rd quartile</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Values Affirmation</td>
<td>-0.143</td>
<td>-0.282</td>
<td>0.271</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>[0.272]</td>
<td>[0.255]</td>
<td>[0.245]</td>
<td>[0.135]</td>
</tr>
<tr>
<td>Female</td>
<td>-0.465</td>
<td>-0.804**</td>
<td>-0.582</td>
<td>-0.812***</td>
</tr>
<tr>
<td></td>
<td>[0.382]</td>
<td>[0.312]</td>
<td>[0.366]</td>
<td>[0.183]</td>
</tr>
<tr>
<td>Female X Values Affirmation</td>
<td><strong>0.224</strong></td>
<td><strong>0.115</strong></td>
<td><strong>0.138</strong></td>
<td><strong>0.658</strong>*</td>
</tr>
<tr>
<td></td>
<td>[0.493]</td>
<td>[0.417]</td>
<td>[0.446]</td>
<td>[0.242]</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.569**</td>
<td>0.302</td>
<td>0.206</td>
<td>0.868***</td>
</tr>
<tr>
<td></td>
<td>[0.217]</td>
<td>[0.217]</td>
<td>[0.194]</td>
<td>[0.104]</td>
</tr>
<tr>
<td>4th quartile</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Values Affirmation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female X Values Affirmation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel B. Course Score

| Values Affirmation                                         | -0.528** | -0.028 | 0.302 | -0.168 |
|                                                           | [0.245]   | [0.278] | [0.238] | [0.185] |
| Female                                                     | -0.420  | -0.578* | -0.431 | -0.912*** |
|                                                           | [0.345]   | [0.341] | [0.355] | [0.252] |
| Female X Values Affirmation                                | **0.611** | **-0.230** | **-0.093** | **0.371** |
|                                                           | [0.446]   | [0.455] | [0.432] | [0.332] |
| Constant                                                   | -0.284  | 0.123  | 0.245  | 1.149*** |
|                                                           | [0.196]   | [0.237] | [0.188] | [0.143] |

Panel C. End-of-Semester FMCE Score

| Values Affirmation                                         | -0.143  | -0.282 | 0.271  | 0.019  |
|                                                           | [0.272]   | [0.255] | [0.245] | [0.135] |
| Female                                                     | -0.465  | -0.804** | -0.582 | -0.812*** |
|                                                           | [0.382]   | [0.312] | [0.366] | [0.183] |
| Female X Values Affirmation                                | **0.224** | **0.115** | **0.138** | **0.658*** |
|                                                           | [0.493]   | [0.417] | [0.446] | [0.242] |
| Constant                                                   | -0.569** | 0.302  | 0.206  | 0.868*** |
|                                                           | [0.217]   | [0.217] | [0.194] | [0.104] |

Observations: 82, 80, 71, 75

Note: Values Affirmation is a dummy variable that takes value 1 if the subject completed the values affirmation exercise, and 0 if the subject completed the control exercise. Female is a dummy that takes value 1 if the subject is a female, 0 if male. Column (1) restricts the sample to students in the 1st quartile of the distribution of beginning-of-semester FMCE scores, column (2) restricts the sample to students in the 2nd quartile, column (3) to those in the 3rd quartile, and (4) to those in the 4th quartile. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Standard errors in brackets.
Table S3. Effect of values affirmation on female performance, including covariates (students in Miyake et al. (2010))

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Exam Score</td>
<td>Course Score</td>
<td>FMCE Score</td>
</tr>
<tr>
<td>Values Affirmation</td>
<td>0.053</td>
<td>0.010</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td>[0.163]</td>
<td>[0.187]</td>
<td>[0.183]</td>
</tr>
<tr>
<td>Stereotype endorsement</td>
<td>-0.128</td>
<td>-0.107</td>
<td>-0.090</td>
</tr>
<tr>
<td></td>
<td>[0.081]</td>
<td>[0.093]</td>
<td>[0.091]</td>
</tr>
<tr>
<td>Values Affirmation X Stereotype Endorsement</td>
<td>0.324***</td>
<td>0.351***</td>
<td>0.245***</td>
</tr>
<tr>
<td></td>
<td>[0.081]</td>
<td>[0.093]</td>
<td>[0.091]</td>
</tr>
<tr>
<td>FMCE prior score</td>
<td>0.334**</td>
<td>0.419**</td>
<td>0.399**</td>
</tr>
<tr>
<td></td>
<td>[0.155]</td>
<td>[0.177]</td>
<td>[0.174]</td>
</tr>
<tr>
<td>FMCE prior score X Values Affirmation</td>
<td>0.054</td>
<td>-0.000</td>
<td>0.146</td>
</tr>
<tr>
<td></td>
<td>[0.191]</td>
<td>[0.218]</td>
<td>[0.214]</td>
</tr>
<tr>
<td>FMCE prior score X Stereotype Endorsement</td>
<td>-0.002</td>
<td>-0.068</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>[0.096]</td>
<td>[0.110]</td>
<td>[0.108]</td>
</tr>
<tr>
<td>Constant</td>
<td>-0.102</td>
<td>-0.227</td>
<td>-0.416***</td>
</tr>
<tr>
<td></td>
<td>[0.124]</td>
<td>[0.141]</td>
<td>[0.139]</td>
</tr>
<tr>
<td>Observations</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.297</td>
<td>0.282</td>
<td>0.308</td>
</tr>
</tbody>
</table>

Note: The sample includes female students only. The variable Stereotype Endorsement is centered for the female students in the sample with available FMCE scores, following Miyake et al. (2010). The FMCE prior score is standardized for the female students in the sample with available FMCE scores. The interaction effects of Values Affirmation X Stereotype Endorsement, FMCE prior score X Values Affirmation and FMCE prior score X Stereotype Endorsement are included following the same specification as Miyake et al. (2010), but dropping the gender main effect and interaction terms. Instead of adding SAT/ACT scores as a covariate in columns (1) and (2), we add FMCE prior score, because this measure can be standardized for the sample of only female students, based on the raw data provided by Miyake et al., and this cannot be done for the SAT scores. *** p<0.01, ** p<0.05, * p<0.1. Standard errors in brackets.
Female Performance, Stereotype Threat and Values Affirmation

Table S3 indicates that there is a positive interaction between stereotype threat endorsement and values affirmation. This means that the values affirmation intervention has a more positive effect on female students who show high stereotype endorsement. It is an interaction effect that needs to be carefully interpreted. Particularly, by only considering the interaction effect, one cannot know whether the positive sign stems from a negative effect on female students with a low stereotype endorsement, that disappears among students with a high stereotype endorsement. Or, whether it is positive throughout, and more strongly positive on female students.

Figure S1 shows the covariate-adjusted effects of values affirmation on female students, by stereotype endorsement, and their confidence intervals. Panel A shows the effects on mean exam score, Panel B those on final course score and Panel C those on end-of-semester FMCE score. Panel A reveals that values affirmation significantly decreases the exam scores of female students who strongly disagree with the stereotype (N=55). By contrast, values affirmation significantly increases the exam score of female students who either neither agree nor disagree, or agree, with the stereotype. This applies to a group of 17 students. The cell of female students who agree with the stereotype and are part of the values affirmation intervention is based on 4 students. Panel B and C show qualitatively similar results, but the negative effects are no longer significant and smaller.

Figure S1. Effects of Values Affirmation on Female Students, by Stereotype Threat
Panel B: Mean Course Score

Note: The 95% confidence interval is shown for each performance outcome.

Panel C: End-of-Semester FMCE Score

...I expect men to generally do better in physics than women
Interpreting Covariate-adjusted Effects: Details

In what follows we first discuss covariate-unadjusted effects, and then covariate-adjusted effects. We present the estimated regression models and discuss the interpretation of the parameters of interest.

Covariate-unadjusted effects. Let $Z_i$ be 1 if student $i$ was assigned to self-affirmation condition, 0 otherwise. Gender is denoted as $F_i = 1$ if student $i$ was female, 0 otherwise.

The regression model (without covariate adjustment) is:

$$Y_i = \beta + \beta_Z \times Z_i + \beta_F \times F_i + \beta_{FZ} \times F_i \times Z_i + \epsilon_i,$$

$i = \{1, \ldots, N\}$

The effect of the intervention on women is therefore:

$$\Delta_Z(\text{women}) \equiv E[Y \mid Z=1, F=1] - E[Y \mid Z=0, F=1] = \beta_Z + \beta_{FZ},$$

while the corresponding effect for men is:

$$\Delta_Z(\text{men}) \equiv E[Y \mid Z=1, F=0] - E[Y \mid Z=0, F=0] = \beta_Z$$

The effect of self-affirmation on the gender gap (defined as outcome for women minus men) is then:

$$\Delta_Z(\text{women}) - \Delta_Z(\text{men}) = \beta_{FZ}$$

If $\beta_{FZ}>0$, then the size of the gender gap is reduced as a result of the self-affirmation intervention. The results of our analysis revealed that we cannot reject the hypothesis that $\beta_{FZ}=0$. Hence, there was no significant effect of self-affirmation.

Covariate-adjusted effects. Here we specify the regression model used in Miyake et al (2010) and compare the interpretation of this model with that in the model without covariates. The regression model used by Miyake et al. (2010) to estimate the effect of the self-affirmation intervention on the gender achievement gap includes covariates. One of the two covariates included, and potentially the most important one is prior performance $S_i$, which corresponds to student $i$’s SAT score. The regression model including this covariate is:

$$Y_i = \beta + \beta_Z \times Z_i + \beta_F \times F_i + \beta_{FZ} \times F_i \times Z_i + \beta_S \times S_i + \beta_{SZ} \times S_i \times Z_i + \beta_{SF} \times S_i \times F_i + \epsilon_i,$$

$i = \{1, \ldots, N\}$

For simplicity, we omit triple interaction effects, and the other covariate used in the estimation of covariate-adjusted effects, gender stereotype endorsement. The same results hold when we include these.

There are two problems associated with this specification:
1. **Parameter of Interest**: The problem with conditioning on SAT Score is that it changes the parameter that is estimated from an unconditional effect to a conditional effect. The effect of the self-affirmation for women conditional on $S=s$ (i.e., the effect for women with SAT score equal to $s$) is:

$$
\Delta_Z(\text{women, } S=s) = E[Y | Z=1, F=1, S=s] - E[Y | Z=0, F=1, S=s] = \beta_Z + \beta_{FZ} + \beta_{SZ} \times s
$$

The corresponding effect for men is:

$$
\Delta_Z(\text{men, } S=s) = E[Y | Z=1, F=0, S=s] - E[Y | Z=0, F=0, S=s] = \beta_Z + \beta_{SZ} \times s
$$

Hence, the effect of the self-affirmation on the gender gap is found by subtracting the two effects, which gives:

$$
\Delta_Z(\text{women, } S=s) - \Delta_Z(\text{men, } S=s) = \beta_{FZ}
$$

This looks just like the unadjusted effect and it is indeed tempting to think of it as identical (just more precisely estimated due to the covariates decreasing the standard error). However, it is not. The interpretation of this effect is the reduction in the gender gap for a *population of men and women who have the same SAT score*.

Is this the effect we are interested in? It could be if the distributions of SAT scores for men and women had similar means. However, the prior performance and stereotype endorsement of males and females differ (for SAT scores, $t(397)=2.62$, $p=0.01$; for beginning-of-semester FMCE scores, $t(306)=4.80$, $p<0.01$; for stereotype endorsement, $\chi^2(4)=41.64$, $p<0.01$). Hence, the effect that is estimated with covariate adjustment is only relevant for the small subset of women who have the same SAT scores as men. In Miyake et al. (2010), this is 56% of the sample, considering only SAT or ACT scores. Including stereotype endorsement, only 28% of the sample features male and female students with the same SAT scores and stereotype endorsement.

What we want is to compare the average effect for women

$$
\Delta_Z(\text{women}) = \beta_Z + \beta_{FZ} + \beta_{SZ} \times E[S|\text{women}]
$$

to the average effect for men

$$
\Delta_Z(\text{men}) = \beta_Z + \beta_{SZ} \times E[S|\text{men}],
$$
or – more generally – compare the effects at different quantiles of the S distribution for men and women. We showed these two results above. First, since SAT scores and stereotype endorsement are standardized and sample-centered in the data, the average effect for women and men is simply the covariate-unadjusted effect, which we showed is not statistically significant (Table S1). Second, the effects per quartile of the S distribution are generally null,
in 11 out of 12 comparisons (Table S2).

2. **Endogeneity of $S$**: A more worrying problem in including $S$ is the potential endogeneity. Note that $S$ is a prior test score and $Y$ is a current test or exam score. A reasonable assumption is that both of these are correlated with some common underlying unobserved ability $\theta$ (we can call it something like “science aptitude”). Since $\theta$ is unobserved it is part of the error term, $\varepsilon$, in the regression model (because it is unobserved) so we may write the model as:

$$Y_i = \beta + \beta Z \times Z_i + \beta F \times F_i + \beta FZ \times F_i \times Z_i + \beta S \times S_i + \beta SZ \times S_i \times Z_i + \beta SF \times F_i \times Z_i + \alpha \theta_i + \phi_i \quad i={1, \ldots, N}$$

where $\alpha$ is expected to be positive. Similarly, we would expect $X$ to be generated as something like

$$X_i = \mu_x + \theta_i + \xi_i.$$ 

where $\xi_i$ is an error term, capturing the notion that test scores like SAT do not measure ability perfectly (measurement error). Unless the variance of $\xi_i$ is zero, and thus the SAT score is a perfect proxy for ability, then $S$ will be correlated with the joint error term in the regression, leading to biased estimates.
Specification Curve Analysis

In what follows, we detail the steps taken in the specification-curve analyses (Simonsohn et al., 2015).

1. Identification of the set of potential specifications
1.A. Specification curve for the interaction effect between Values Affirmation and Gender

The specification by Miyake et al. (2010) includes 11 independent variables, several of which are interaction effects. Their main outcome of interest is students’ average performance in all exams, though other outcomes such as scores in the FMCE are also considered in their analysis.

Based on the dependent and independent variables available in the dataset shared with the authors, several potential specifications could be chosen. In what follows we describe several choices, indicating the choice by Miyake et al. (2010) and other reasonable alternatives that could also have been chosen:

1) Dependent Variable (DV): The data by Miyake et al. (2010) included 3 measures of performance, all of which can be dependent variables. They are the average score in all exams (“Average Exam Score”), the average grade in the class (“Average Course Score”) or the score obtained in the End-of-Semester FMCE (“FMCE End of Sem”). All variables were considered by Miyake et al. (2010) at different points of their paper. A further dependent variable reported in the paper was the letter grade of each student (A through F), but this variable was not made available for the present re-analysis.

2) Controlling for Stereotype Threat: The first option is not to include stereotype threat as a covariate (“No”), to measure the covariate-unadjusted effect of values affirmation. The second option is to only include it as a control, without interactions (“Yes”). The third option is to include it as control and interacted with treatment assignment and gender (“2-way interaction with fem & treat”). The fourth option, chosen by Miyake et al. (2010) is to, in addition, include a 3-way interaction between Stereotype threat, values affirmation treatment assignment and gender (“3-way interaction”).

3) Stereotype Threat: The first option is to include stereotype threat as a continuous variable as in Miyake et al. (2010) (“Stereotype threat, continuous”). This variable takes values 1 to 5, from “strongly disagree” to “strongly agree”, and is sample-centered. Another approach is to split the sample by the median, and control for whether the student’s stereotype threat is above median or not (“Stereotype threat, median split”).

4) Controlling for Math Ability: Again, the first option is not to include math ability as a covariate, to measure the covariate-unadjusted effect of values affirmation (“No”). Second, regressions could only include math ability as a covariate, without interactions (“Yes”). Third, regressions could include math ability as a control and interacted with treatment assignment and gender (“2-way interaction with fem & treat”). The fourth option, chosen by Miyake et al. (2010) is to also include a 2-way interaction between math ability and stereotype threat (“2-way interactions with fem & treat & stereotype”).

5) Math Ability: There are different measures of math ability in the data: the Beginning-of-Semester FMCE score, and the SAT/ACT score. We allow for each measure to be used as a control, specified as a continuous variable (“SAT/ACT, continuous”, “FMCE Begin of Sem, continuous”) and also as a dummy indicating whether the student was above or below median in each variable (“SAT/ACT, median split”, “FMCE Begin of Sem,
median split”). In Miyake et al. (2010), SAT/ACT score was included as a control for specifications in which the dependent variable was the average exam score, while the FMCE Beginning of Semester Score was included as a control for specifications in which the dependent variable was the FMCE End of Semester Score. We allow all possible combinations to better understand the robustness of the estimated interaction effects.

(6) Sample Restriction: There are three potential samples: all students (N=668), students who have no missing information about their ACT/SAT scores (N=399) and students who have no missing information about their FMCE scores at the beginning of the semester (N=308).

(7) Robust Standard Errors: includes robust standard errors, estimated using the Huber-White or sandwich estimator (Yes) or does not (No). Note that this does not affect the estimated coefficients, but rather the standard errors.

Recall that performance measures (exam scores and test scores) have been standardized throughout. Therefore, the coefficient of the interaction effect of values affirmation and gender can be interpreted as a “standardized coefficient” with respect to the dependent variable, and in standard deviations for this variable. To replicate the results in Miyake et al. (2010) within this analysis, all variable definitions were kept as in the original paper. That is, treatment assignment is a dummy variable that takes value -1 in the control, and 1 in the treatment. Gender (female) is a dummy that takes value -1 if the student is male, and 1 if it is a female. Stereotype threat and beginning of semester FMCE scores were sample-centered, and SAT/ACT scores were standardized.

1.B. Specification curve for the effect of Values Affirmation on Female Students
Miyake et al. (2010) focused on the coefficient of the interaction term between values affirmation and gender. This interaction term indicated whether the effect of values affirmation on academic performance was different for female students, compared to male students. Yet, the main hypothesis of the intervention was that it reduced stereotype threat and thereby improved performance of the stereotyped group, women.

Next, we focus on the effect of values affirmation on female students. Based on the dependent and independent variables available, several potential specifications could be also chosen. The specifications considered in the analysis were the following:

(1) Dependent Variable (DV): This may be the average score in all exams (Average Exam Score), the average grade in the class (Average Course Score) or the score obtained in the End-of-Semester FMCE (FMCE End of Sem).

(2) Controlling for Stereotype Threat: Regressions may not include this measure as a covariate (No), may include it only as a control, without interactions (Yes), may include it as control and interacted with treatment assignment (2-way interaction with treatment).

(3) Stereotype Threat: The first option is to include stereotype threat as a continuous variable as in Miyake et al. (2010) (“Stereotype threat, continuous”). This variable takes values 1 to 5, from “strongly disagree” to “strongly agree”, and is sample-centered. Another approach use the median stereotype in the class, and control for whether the female student’s stereotype threat is above median or not (“Stereotype threat, median split”).
Controlling for Math Ability: Regressions may not include math ability as a covariate (No), may include it as a control only, without interactions (Yes), may include it as a control and interacted with treatment assignment (2-way interaction with treat), as well as interacted with treatment assignment and stereotype threat (2-way interactions with treat & stereotype).

Math Ability: This allows for different specifications of math ability. First, the Beginning-of-Semester FMCE, centered for the sample of female students only. Second, a dummy indicating whether the Beginning-of-Semester FMCE score is above median in the sample of female students or not. Third, a dummy indicating whether the SAT/ACT score is above median in the sample of female students or not. Unfortunately, the SAT/ACT scores provided by Miyake et al. (2010) were standardized considering the sample of male and female students and could not be standardized for the sample of female students without further information. Hence, we did not include a continuous measure of SAT/ACT grades for female students as a measure of math ability.

Sample Restriction: There are three potential samples: all female students (N=181), female students who have no missing information about their ACT/SAT scores (N=116) and female students who have no missing information about their FMCE scores at the beginning of the semester (N=96).

Robust Standard Errors: includes robust standard errors, estimated using the Huber-White or sandwich estimator (Yes) or does not (No). Note that this does not affect the estimated coefficients, but rather the standard errors.

2. Results and Inference
2.A. Results for the interaction effect between Values Affirmation and Gender
Considering potential and reasonable regression models that could have been run by the original authors with the available data, we obtained 1566 unique interaction effects of values affirmation and gender.¹

Figure 2 shows the coefficient for the average effect of the interaction effect of the treatment (values affirmation) and female students, and its 95% confidence interval, for each specification. As mentioned above, the dependent variable was standardized such that the coefficient can be interpreted in terms of standard deviations of the dependent variable. If the regression did not include a 3-way interaction effect this is simply the coefficient of the interaction term. If the interaction effect of values affirmation and gender was interacted with stereotype threat (3-way interaction), as was done in Miyake et al. (2010), we were interested in the average interaction effect. Using the same notation as in the text, if the regression included the treatment assignment (Z) and gender (F) jointly interacted with stereotype threat (T), we calculated the coefficient of the interaction effect as:

\[ \hat{\beta} = \beta_{ZF} + \beta_{ZFT} \times E(T) \]

¹ Note that a total of 1728 regression specifications are possible out of the 7 potential choices listed. However, these are not all unique. For example, if the regression specification does not include math ability as a measure of prior performance, the definition of math ability is irrelevant.
Figure 2 (main text) plots $\hat{\beta}$ and its standard error.

Out of 1566 specifications, 1,205 (76.95%) yielded an interaction effect that was not statistically significant. The average $t$-statistic across all specifications was 1.56 ($sd=0.55$). Considering average course score as the dependent variable to capture a student’s performance in the physics class, the specification reported in Miyake et al. (2010) had a $t$-statistic of 3.08. The resulting interaction effect was the 15th highest out of 1566 specifications. It was in the 99th percentile of the distribution resulting from the specification curve analysis.

As argued in the text, the inclusion of covariates changes the interpretation of the estimated coefficients. There are three cases one could compare:

1. No covariates: The specification curve included 18 regression models without covariates. 11.1% of these yielded a significant interaction effect ($p$-value<0.05). The average estimated interaction effect was 0.076.

2. Covariates, excluding a three-way interaction: What was the average estimated interaction effect and how often was it significant if we included both covariates (stereotype endorsement and prior ability), but did not include a 3-way interaction effect? The average interaction effect was 0.072. There were 864 possible regressions of this sort and 9.4% of them yielded a significant coefficient ($p$-value<0.05).

3. Regressions including a three-way interaction effect: If a 3-way interaction effect was included, the average estimated interaction effect was 0.11, and 56% of 468 specifications were statistically significant ($p$-value<0.05).

Lastly, of the 361 specifications that yielded a significant interaction effect, 72.9% included a 3-way interaction effect.

We did not include an inference analysis as described in Simonsohn et al. (2015) for the interaction effect because gender is not randomly assigned. We did this for the effect of values affirmation on female students (next subsection).

To conclude, the results revealed that the interaction effect between gender and treatment was generally not significant, using the original data. The analysis showed that a key driver of significant effects was the inclusion of 3-way interaction effects.

2.B. Results for the Effect of Values Affirmation on Female Students
Considering all potential unique combinations of regressions measuring the effect of values affirmation on female students, we obtained 726 plausible specifications that could have been run with the available data.² For each specification, Panel B of Figure 2 in the body of the text plots the average effect of the treatment (values affirmation) on female students, and its confidence interval.

² Note that a total of 1296 regression specifications are possible out of the 7 potential choices listed. However, these are not all unique. For example, if the regression specification does not include math ability as a measure of prior performance, the definition of math ability is irrelevant.
If the treatment variable, values affirmation, was not interacted with any other covariate in the regression specification, Panel B of Figure 2 (in the main text) shows the coefficient of the treatment variable. If the treatment variable was interacted with math ability and/or stereotype threat, we calculated the coefficient for the average effect of the treatment in the following way. For example, if the regression included the treatment assignment \((Z_i)\) and the treatment interacted with math ability \((S_i)\), using the same notation as in the text, we calculated the coefficient of the treatment effect as:

\[
\hat{\beta} = \beta_Z + \beta_{ZS} \times E(S_i)
\]

We plot \(\hat{\beta}\) and its confidence interval in Panel B of Figure 2. Out of 726 specifications, 704 (96.97%) yielded an interaction effect that was not statistically significant. The average \(t\)-statistic across all specifications was 0.82 (sd=0.61).

Next, we proceeded to explore statistical inference for the specification curve by asking, “considering the full set of reasonable specifications jointly, how inconsistent are the results with the null hypothesis of no effect?” (Simonsohn et al., 2015). This involved conducting a permutation test, using 500 shuffled samples. The results are shown in Figure S2. The observed data always lies within the 95% confidence interval. We present three test statistics that compare the observed data with the permuted datasets in Table S4. The median effect of values affirmation in the observed data was 0.08. Such an effect or higher was obtained in 32% of the shuffled samples. Ninety-three percent of specifications yielded a positive sign for the effect of values affirmation on female students, a share that was not significantly different from the share under the null. Finally, 3% of specifications featured a significant effect of values affirmation on female students, a share that was not significantly higher than that under the null.

Overall, the results revealed that the effect of values affirmation on female students was generally not significant. The inference analysis using the specification curve yielded the same result.
Figure S2. Observed and Expected Under-The-Null Specification Curves for the Effect of Values Affirmation on Female Students

Notes: The expected under-the-null specification curves are based on 500 shuffled samples, in which values affirmation (treatment) is shuffled. All specifications are estimated in each shuffled sample. The resulting coefficient estimates for the observed data (blue dots), as well as the median and 2.5th and 97.5th percentiles are shown.

<table>
<thead>
<tr>
<th></th>
<th>Observed Result</th>
<th>p-value (% of shuffled samples with as or more extreme results)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Values Affirmation Effect on Female Students</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Median coefficient</td>
<td>0.08</td>
<td>0.320</td>
</tr>
<tr>
<td>2. Share of results with predicted sign</td>
<td>0.93</td>
<td>0.222</td>
</tr>
<tr>
<td>3. Share of results with predicted sign &amp; p&lt;0.05</td>
<td>0.03</td>
<td>0.308</td>
</tr>
</tbody>
</table>
Suggestive Replication Study

We started this study by attempting to replicate the values affirmation intervention in an introductory physics class for engineering students, at the University of California, San Diego. We used the same materials as Miyake et al. (2010) and followed their procedures as closely as possible. Miyake et al. (2010) focused on performance in the final exam, on a standardized physics test (FMCE) and final course grade. Instead of completing a standardized physics test (FMCE), students at UC San Diego had to complete weekly quizzes in class, which counted towards their final grade. Hence, we focused on three outcomes: (1) average quiz score, (2) final exam score and (3) final course score.

In total 129 students participated in the study, 44 in the control condition (22 females and 22 males) and 85 in the values affirmation condition (39 females and 46 males). As in the original study, the first values affirmation exercise took place in the first tutorial session, while students were invited to complete the second one online. Seventy-five students completed the second exercise. In the analysis we report the results considering all students and report any differences in the results when only students who completed both exercises are included. A detailed description of the course, the procedures and the sample is provided below. We focus on the effect of values affirmation on raw (covariate-unadjusted) means.

The sample size in this replication was smaller than in the original study. This is an important limitation of the replication study, which is underpowered and should therefore be considered suggestive. Nevertheless, we include the results here to disclose the data we collected. The results here should be viewed in conjunction with large-scale replication studies such as Hoffman and Kurtz-Cortes (2019), De Jong et al. (2016), De (2015), Bratter, Rowley and Chukhray (2016), Hanselman et al. (2017), Borman (2012), and Lauer et al. (2013). There were a number of differences compared to Miyake et al. (2010) that are worth noting. The percentage of male students was lower (53% compared to 74% in the original study), the evaluation format (quizzes and exams) was different, and students did not have access to a Peer Instruction curriculum, a resource to help them improve their learning, but access on campus to night tutorials 5 nights a week.

Fig. S3 summarizes the outcomes by condition and gender. The complete regression results are shown in Table S5. Table S6 presents the same analysis focusing only on subjects who completed the values affirmation exercise twice.

Fig. S3(a) displays the average quiz score. Females obtained a mean quiz score 63.25 (s.e.: 3.49) in the control and 66.95 (s.e.: 2.60) in the values affirmation condition. The difference in quiz scores across treatments was not significant ($\beta$=0.20 standard deviations (SD), $t(59)=0.85$, $p=0.40$). We cannot reject that this result differs to that found in the original study (without covariate adjustment), where FMCE scores showed a directional increase of 0.27 SD, that also lacked statistical significance ($p$-value=0.2).

Male students obtained a mean quiz score of 76.64 (s.e.:2.70) in the control condition and 70.71 (s.e.: 2.35) in the values affirmation condition. The difference in quiz scores was again not significant ($\beta$=-0.33 SD, $t(66)=-1.53$, $p=0.132$). The coefficient of the interaction effect of gender and values affirmation, $\beta_i$, was on the limit of marginal significance ($\beta_i=0.53$ SD, $t(125)=1.66$, $p=0.10$), owing to the joint non-significant changes in female and male quiz scores.

For male students who completed both affirmation exercises, values affirmation had a significantly negative effect on quiz score ($\beta$=-0.68 SD, $t(35)=-2.39$, $p=0.022$), while the
performance of women remained unaffected ($\beta=0.11$ SD, $t(36)=0.35$, $p=0.73$). Hence, if at all, the only significant effect of values affirmation was negative on male students.

(a) (b) (c)

**Fig. S3:** Mean quiz scores, final exam scores and final course scores are reported in panels (a), (b) and (c), respectively. +/- 1 SE are indicated.

In terms of the final exam performance, neither female nor male performance was affected significantly by values affirmation. Female final exam scores in the control and values affirmation conditions were 75.27 (s.e.: 3.14) and 69.38 (s.e.: 4.52), respectively. Male final exam scores were 73.41 (s.e.: 2.42) in the control condition and 78.41 (s.e.: 3.94) in the values affirmation condition. This led to a directional *increase* in the gender gap in the values affirmation condition. The interaction term was not significant ($\beta=-0.56$ SD, $t(125)=-1.57$, $p=0.118$). Considering students who participated in both affirmation exercises, the effects were also not significant.

The performance in the course of males and females – which was based 60% on the quiz score, 35% on the final exam score and 5% on class participation – was consequently not significantly affected by the values affirmation intervention. Considering only students who completed both values affirmation exercises, we found a marginally significant negative effect of the exercise on male students ($\beta=-0.52$ SD, $t(35)=-1.90$, $p=0.065$), owing to the negative effect of the intervention on their quiz scores, and we found no significant effect on female students ($\beta=-0.01$ SD, $t(36)=-0.02$, $p=0.985$).

Overall, our results revealed no significant effect of values affirmation on female students on any dimension, the same as we found when reanalyzing the data by Miyake et al. (2010). We observed a directionally negative effect of the values affirmation intervention on male students, at least on some dimensions of their performance, which was also reported by Miyake et al. (2010).
Details of the Suggestive Replication Study

A. Description of the course. For the replication study, the values affirmation intervention was conducted in an introductory physics course at UCSD. This class was intended for physical science and engineering majors. It was a calculus-based science-engineering general physics course covering vectors, motion in one and two dimensions, Newton’s first and second laws, work and energy, conservation of energy, linear momentum, collisions, rotational kinematics, rotational dynamics, equilibrium of rigid bodies, oscillations, gravitation. The class took place in the Fall Quarter of the academic year 2012-2013. Lectures took place on Mondays, Wednesdays and Fridays. Additionally, there were discussion sessions, on Monday evenings and Tuesday evenings.

The grading of the course was computed based on the average grade of the best six quizzes out of eight, which counted 60% towards the final course grade, and the final exam grade, which counted 40% towards the final course grade. Each quiz consisted of 4 multiple-choice questions and was conducted during the discussion session each week. Hence, the grades for each quiz were 0, 25, 50, 75 or 100. The final exam consisted of 12 multiple-choice questions and grades ranged from 0 to 100. The number of students who took the final exam was 321.

B. Procedures. We conducted the values affirmation exercise twice, following the procedures of Miyake et al. (2010). We used exactly the same materials that they used in both interventions. The exact documents used, which are those that were shared by Miyake et al. (2010), can be obtained from the authors.

All students were asked to consent to participation in the study, following the procedures of Miyake et al. (2010) and in line with the IRB regulations at UCSD. Consent of the students was requested during the first values affirmation exercise. Students who did not sign consent forms during the first exercise were requested to consent during the second exercise. Only students who consented are included in the study.

Following Miyake et al. (2010), we trained the teaching assistant at UCSD, who lead the first affirmation exercise. The second exercise was conducted online. During the first affirmation exercise, which took place in the first review session, there were personnel from the research team to monitor the administration. Instructors remained blind to the intervention by use of manila envelopes, which were sealed by students upon completion of the exercise. These envelopes were collected immediately after the administration by personnel from the research team.

C. Sample. The first values affirmation exercise was conducted during the first review session in the quarter, on Tuesday, October 2nd, 2012. Out of 201 students who were present, 178 respondents returned the exercises. 3 14 students did not provide their student identification number, 2 additional students did not provide their gender, while 16 dropped the class and did not complete the final exam. Out of the remaining students 17 did not provide consent

---

3 There were 201 participants in the session. 17 students returned the exercises completely blank, four rejected to participate and two requested and completed an alternative assignment.
to participate in the study. This leaves 129 students who completed the first exercise, finished the class (took the final exam) and consented to participate in the study. The attrition was not differential by condition ($\chi^2(1)=0.3383$, $p=0.561$).

Of the 129 students (68 males, 61 females), 22 males and 22 females were randomly assigned to the control group. 46 males and 39 females were randomly assigned to the treatment group.

The second values affirmation exercise was conducted online on November 7th, 2012. This was the 6th week of the course, in the middle between the beginning of classes (September, 28) and the final exam (December, 12). A total of 43 students replied within the first week. To increase participation a reminder was sent on November 26th. In total, 76 students participated in the 2nd administration. Of these 75 finished the class and consented to participate in the study. There were 37 males and 38 females. There was no significant difference in gender composition among students who completed both administrations and those who only completed the first one ($\chi^2(1)=0.8211$, $p=0.365$).

During this course, unlike in the course considered in Miyake et al. (2010), the instructors conducted no survey as part of the class. We hence chose not to add a separate survey eliciting gender stereotype endorsement among students, so as not to interfere with the potential effects of the values affirmation exercise. Adding a survey only with questions about gender stereotypes or adding these questions together with the values affirmation exercises could have potentially compromised the effectiveness of the intervention.
Table S5. Effect of values affirmation on student performance (whole sample).

<table>
<thead>
<tr>
<th>Grade</th>
<th>Quiz (1)</th>
<th>Exam (2)</th>
<th>Male Final Grade: (3)</th>
<th>Female Final Grade: (4)</th>
<th>Male Exam: (5)</th>
<th>Female Exam: (6)</th>
<th>Male Quiz: (7)</th>
<th>Female Quiz: (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Values Affirmation is a dummy variable that takes value 1 if the subject completed the values affirmation exercise, and 0 if the subject completed the control exercise. Female is a dummy that takes value 1 if the subject is a female, 0 if male. ** p<0.01, *** p<0.001. Standard errors in brackets.

Observations: 129
Table S6. Effect of values affirmation on student performance, without covariates (students who completed both exercises)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Quiz</th>
<th>Exam</th>
<th>Final Grade</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.676**</td>
<td>0.109</td>
<td>-0.201</td>
<td>-0.163</td>
<td>-0.523*</td>
</tr>
<tr>
<td>Female</td>
<td>0.112</td>
<td>0.0571</td>
<td>0.517</td>
<td>0.332</td>
<td>0.306</td>
</tr>
<tr>
<td>Male</td>
<td>0.785*</td>
<td>0.0371</td>
<td>0.517</td>
<td>0.0371</td>
<td>0.306</td>
</tr>
<tr>
<td>Female</td>
<td>0.785*</td>
<td>0.0371</td>
<td>0.517</td>
<td>0.0371</td>
<td>0.306</td>
</tr>
</tbody>
</table>

Note: Values Affirmation is a dummy variable that takes value 1 if the subject completed the values affirmation exercise and 0 if the subject completed the control exercise. Female is a dummy variable that takes value 1 if the subject is a female, 0 if male. 

*** p < 0.01, ** p < 0.05, * p < 0.1. Standard errors in brackets.