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)

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))

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Exam Score		Final Course Score		FMCE Score				
	Male	Female	Male	Female	Male	Female	Exam Score	Course Score	FMCE Score
Values Affirmation	-0.246**	0.186	-0.191	0.108	-0.075	0.268	-0.246**	-0.191	-0.075
	[0.119]	[0.185]	[0.119]	[0.193]	[0.138]	[0.210]	[0.120]	[0.121]	[0.140]
Female							-0.751***	-0.608***	-0.667***
							[0.171]	[0.172]	[0.190]
Values Affirmation X Female							0.432**	0.299	0.343
							[0.219]	[0.221]	[0.246]
Constant	0.296***	-0.455***	0.243**	-0.365**	0.194*	-0.474***	0.296***	0.243**	0.194*
	[0.095]	[0.143]	[0.094]	[0.149]	[0.111]	[0.159]	[0.095]	[0.096]	[0.113]
Observations	283	116	283	116	212	96	399	399	308
R-squared	0.015	0.009	0.009	0.003	0.001	0.017	0.061	0.044	0.053

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

Table S2. Effect of values affirmation on student performance, by quartile of the distribution of ability (students in Miyake et al. (2010))

	(1)	(2)	(3)	(4)
	Distribution of Ability (Beginning-of-Semester FMCE Score)			
	1st quartile	2nd quartile	3rd quartile	4th quartile
		Panel A. Mean Exam Score		
Values Affirmation	-0.514**	-0.065	0.250	-0.153
	[0.248]	[0.268]	[0.239]	[0.179]
Female	-0.401	-0.582*	-0.544	-1.072***
	[0.348]	[0.328]	[0.358]	[0.243]
Female X Values Affirmation	**0.651**	**-0.215**	**0.001**	**0.457**
	[0.450]	**[0.439]**	**[0.436]**	**[0.320]**
Constant	-0.403**	0.073	0.274	1.168***
	[0.198]	[0.228]	[0.189]	[0.138]
		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))

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

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.

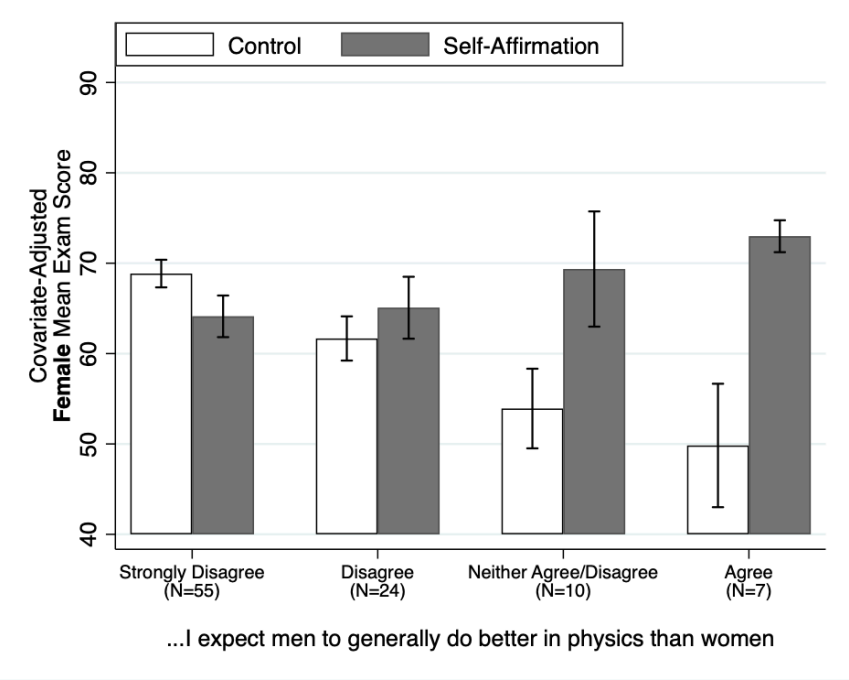
Female Performance, Stereotype Threat and Values Affirmation

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.

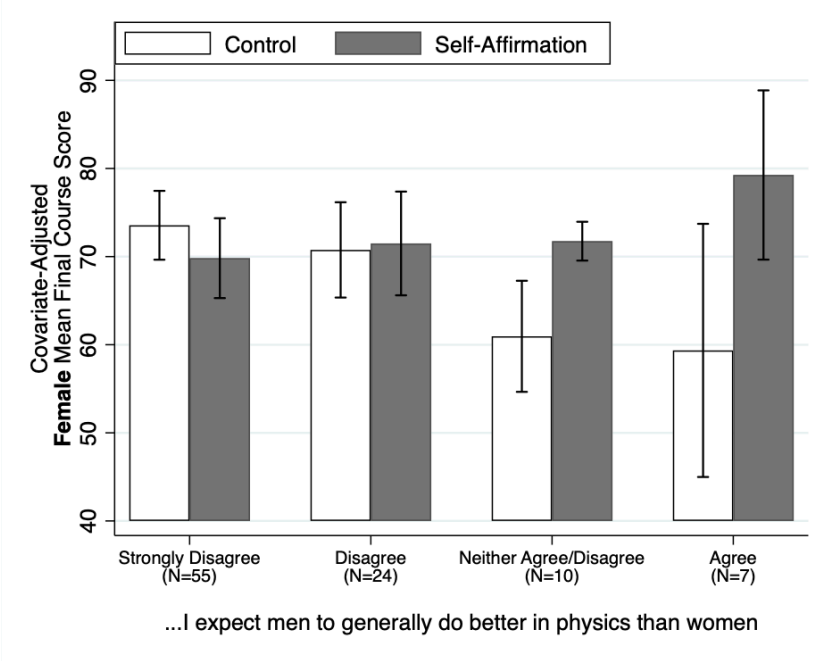
Figure S1 shows the covariate-adjusted effects of values affirmation on female students, by stereotype endorsement, and their confidence intervals. Panel A shows the effects on mean exam score, Panel B those on final course score and Panel C those on end-of-semester FMCE score. Panel A reveals that values 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

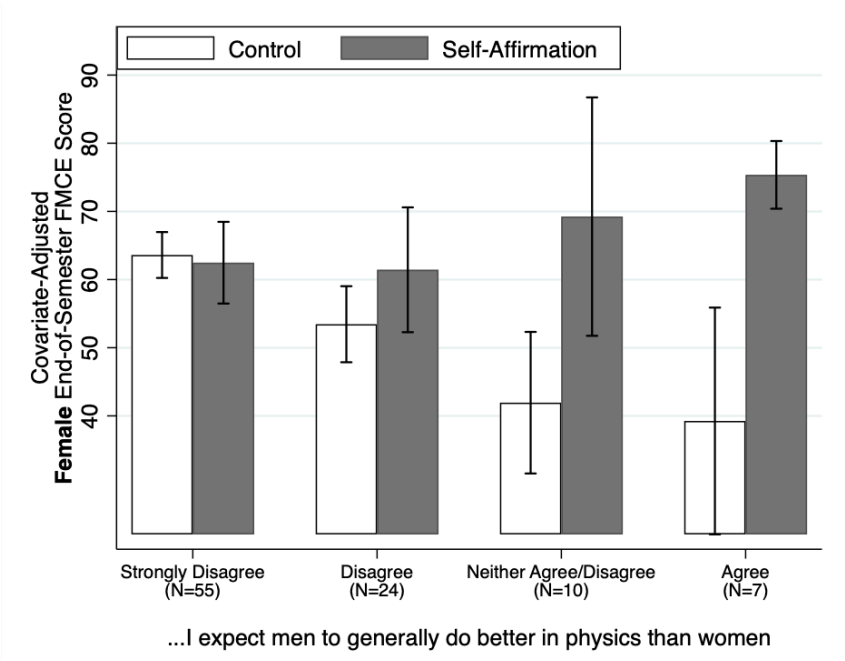
Panel A: Mean Exam Score



Panel B: Mean Course Score



Panel C: End-of-Semester FMCE Score



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

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:

$$Y_i = \beta + \beta_Z \times Z_i + \beta_F \times F_i + \beta_{FZ} \times F_i \times Z_i + \varepsilon_i, \qquad i=\{1,\ldots,N\}$$

The effect of the intervention on women is therefore:

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

while the corresponding effect for men is:

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

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

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

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:

$$\begin{aligned} Y_i = {} & \beta + \beta_Z \times Z_i + \beta_F \times F_i + \beta_{FZ} \times F_i \times Z_i + \\ & \beta_S \times S_i + \beta_{SZ} \times S_i \times Z_i + \beta_{SF} \times S_i \times F_i + \varepsilon_i, \qquad i=\{1,\ldots,N\} \end{aligned}$$

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:

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

The corresponding effect for men is:

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

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

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

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, $\chi2(4)=41.64$ $p<0.01$). 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

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

to the average effect for men

$$\Delta_Z(men) = \beta_Z + \beta_{ZS} \times E[S|men],$$

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 θ (we can call it something like "science aptitude"). Since θ is unobserved it is part of the error term, ε, in the regression model (because it is unobserved) so we may write the model as:

$$\begin{aligned} Y_i \;=\; & \beta + \beta Z \times Z_i + \beta F \times F_i + \beta FZ \times F_i \times Z_i + \\ & \beta S \times S_i + \beta SZ \times S_i \times Z_i + \beta SF \times S_i \times F_i + \alpha\theta_i + \phi_i \qquad\qquad i=\{1,\ldots,N\} \end{aligned}$$

where α 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 ξ_i is an error term, capturing the notion that test scores like SAT do not measure ability perfectly (measurement error). Unless the variance of ξ_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").

(4) 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).
(5) 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.
(6) 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).
(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.

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_i) and gender (F_i) jointly interacted with stereotype threat (T_i), we calculated the coefficient of the interaction effect as:

$$\hat{\beta} = \beta_{ZF} + \beta_{ZFT} \times E(T_i)$$

[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 tex) 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:

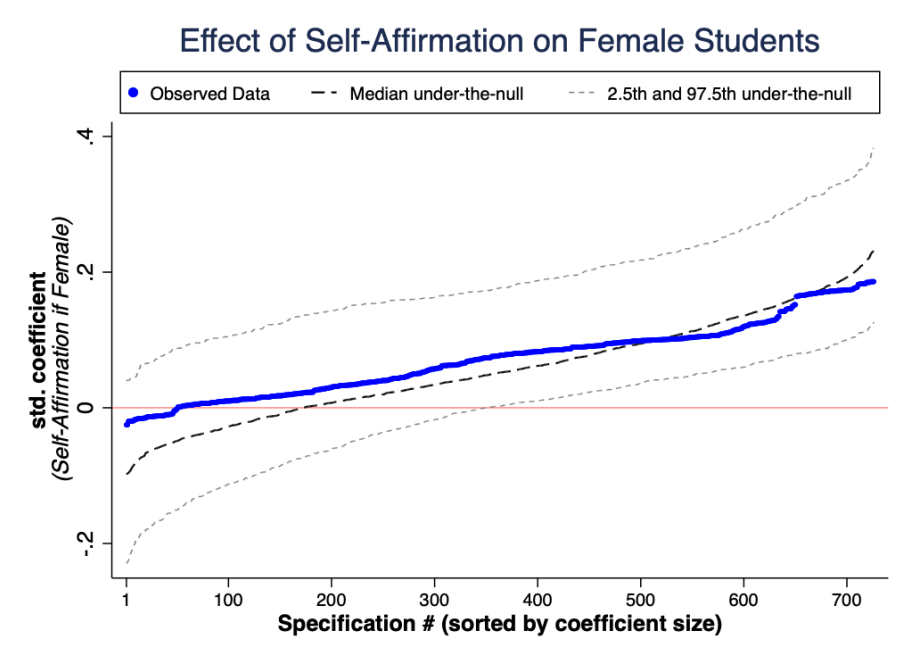
$$\hat{\beta} = \beta_Z + \beta_{ZS} \times E(S_i)$$

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 S4. Joint Inference Tests for the Specification Curve

	Observed Result	p-value (% of shuffled samples with as or more extreme results)
Values Affirmation Effect on Female Students		
1. Median coefficient	0.08	0.320
2. Share of results with predicted sign	0.93	0.222
3. Share of results with predicted sign & p<0.05	0.03	0.308

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), Dee (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 (β_t=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 (β_t=-0.33 SD, $t(66)$=-1.53, p=0.132). The coefficient of the interaction effect of gender and values affirmation, β_i, was on the limit of marginal significance (β_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 (β_t=-0.68 SD, $t(35)$=-2.39, p=0.022), while the

performance of women remained unaffected (β_i=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)

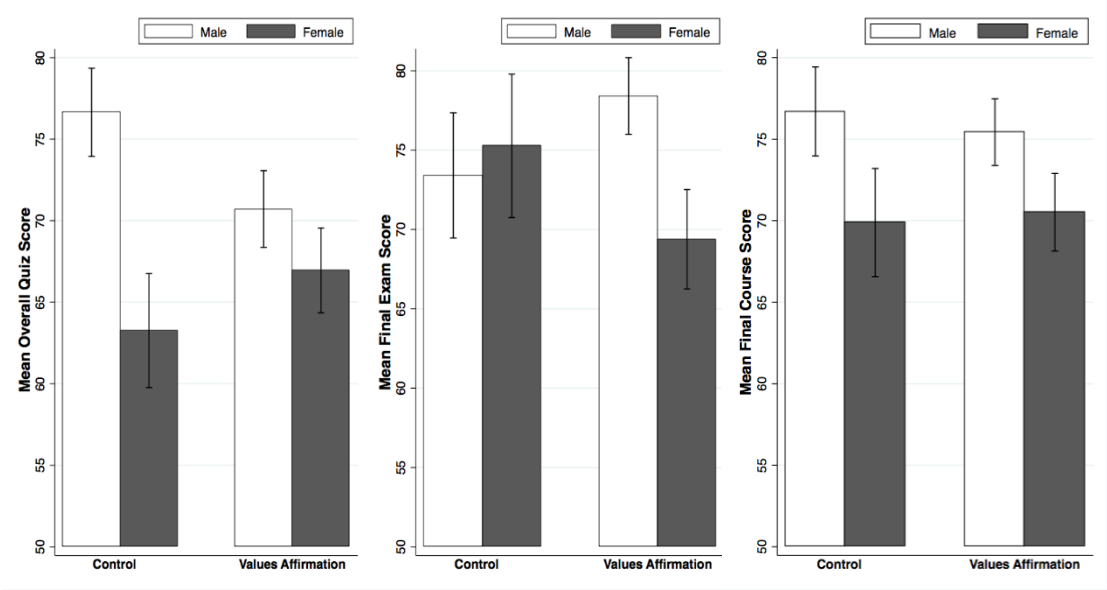


Fig. S3: Mean quiz scores, final exam scores and final course scores are reported in panels (a), (b) and (c), respectively. +/- 1 SE are indicated.

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 (β_i=-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 (β_i=-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 (β_i=-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, without covariates (whole sample)

	(1) Quiz: Male	(2) Quiz: Female	(3) Exam: Male	(4) Exam: Female	(5) Final Grade: Male	(6) Final Grade: Female	(7) Quiz	(8) Exam	(9) Final Grade
Values Affirmation	-0.328	0.204	0.255	-0.300	-0.108	-0.00577	-0.328	0.255	-0.108
	[0.215]	[0.240]	[0.226]	[0.274]	[0.207]	[0.235]	[0.224]	[0.246]	[0.218]
Female							-0.740***	0.0951	-0.415
							[0.260]	[0.286]	[0.253]
Values Affirmation X Female							0.532*	-0.556	0.102
							[0.321]	[0.353]	[0.312]
Constant	0.498***	-0.242	0.0630	0.158	0.381**	-0.0341	0.498***	0.0630	0.381**
	[0.177]	[0.192]	[0.186]	[0.219]	[0.170]	[0.188]	[0.184]	[0.203]	[0.179]
Observations	68	61	68	61	68	61	129	129	129
R-squared	0.034	0.012	0.019	0.020	0.004	0.000	0.070	0.039	0.044

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

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

	(1) Quiz: Male	(2) Quiz: Female	(3) Exam: Male	(4) Exam: Female	(5) Final Grade: Male	(6) Final Grade: Female	(7) Quiz	(8) Exam	(9) Final Grade
Values Affirmation	-0.676**	0.109	-0.201	-0.163	-0.523*	-0.00561	-0.676**	-0.201	-0.523*
	[0.283]	[0.314]	[0.283]	[0.358]	[0.275]	[0.300]	[0.306]	[0.332]	[0.295]
Female							-1.162***	-0.406	-0.879***
							[0.343]	[0.372]	[0.331]
Values Affirmation X Female							0.785*	0.0371	0.517
							[0.424]	[0.460]	[0.409]
Constant	0.862***	-0.300	0.412*	0.00568	0.744***	-0.135	0.862***	0.412	0.744***
	[0.232]	[0.249]	[0.232]	[0.285]	[0.226]	[0.239]	[0.252]	[0.273]	[0.243]
Observations	37	38	37	38	37	38	75	75	75
R-squared	0.140	0.003	0.014	0.006	0.094	0.000	0.174	0.048	0.129

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