INSTITUTE for
REPLICATION

No. 24
I4R DISCUSSION PAPER SERIES

Replication: Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India

Lenka Fiala

Erlend M. Fleisje

Tore Adam Reiremo

March 2023

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 24

Replication: Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India

Lenka Fiala[1], **Erlend M. Fleisje**[2], **Tore Adam Reiremo**[2]

[1]University of Bergen, Dept. of Economics, Bergen/Norway
[2]University of Oslo, Dept. of Economics, Oslo/Norway

MARCH 2023

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research

Hohenzollernstraße 1-3
45128 Essen/Germany

www.i4replication.org

ISSN: 2752-1931

Replication: Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India

Lenka Fiala, Erlend M. Fleisje, and Tore Adam Reiremo[*]

March 14, 2023

Abstract

Dhar et al. (2022) examine the effect of a gender attitude change program in secondary schools in India. In their preferred specification, the authors show that the program made the students report more gender-egalitarian attitudes by 0.18 of a standard deviation, and shifted self-reported behaviors to be more aligned with gender-progressive norms by 0.20 standard deviations (both significant at 1% level). In contrast, they found no effect on girls' aspirations, as these were already high before the intervention. The effects did not attenuate between the first end-line (right after the programme was completed) and the second (two years later). To put the paper's results in perspective, we first comment on the authors' deviations from their pre-registration and pre-analysis plans, provide detailed power calculations, and add multiple-hypothesis-testing-adjusted standard errors. Second, we show that the paper's results are perfectly reproducible. Third, we show that the results are robust to excluding control variables, and alternative ways of constructing indices and dealing with non-response.

1 Introduction

The paper by Dhar et al. (2022) tested the effect of an interactive gender attitude change program in secondary schools in India on student attitudes, aspirations, and behaviors. In their randomized controlled trial, adolescents in 150 schools received a two-year program aimed at raising awareness about gender inequality, and changing views about gender social norms. 164 schools

[*] Fiala: Department of Economics, University of Bergen (lenka.fiala@uib.no); corresponding author. Fleisje: Department of Economics, University of Oslo (e.m.fleisje@econ.uio.no), Reiremo: Department of Economics, University of Oslo (t.a.reiremo@econ.uio.no). We thank the organizers of the Oslo's Replication Games for the opportunity to replicate this paper, and their advice and support on this project. We have received no payment for this replication, and we have no conflict of interests to report.

served as a control group. Students were surveyed at baseline (mid 2013 to early 2014, prior to receiving the program), at an end line shortly after the program ended (late 2016 to early 2017), and second end line, two and a half years after the program ended (first half of 2019). A subsample of parents was also surveyed at the baseline.

The paper makes three key claims about outcomes of the school program: First, the authors state that *"the intervention made gender attitudes more progressive"* (p. 912), specifically, students in the treatment schools report 0.18 sd units increase in gender attitude (progressiveness) index. Second, the program *"did not affect girls' aspirations"* (p. 913), as shown by an effect size of 0.03 sd. And third, the program significantly shifted (self-reported) behavior towards more *"gender-progressive norms by 0.20 sd"* (p. 913).

We re-evaluate these results by first commenting on some analysis choices made by the authors, particularly given their original pre-registration. Building on this discussion, we then reproduce the paper results and note a few discrepancies between the paper and the code. Finally, we provide several robustness checks, addressing some of the shortcomings from the previous two sections.

While we do find some omissions and discrepancies between the published version of the paper and the code and/or the pre-registration, all three main results are robust.

Section 2 discusses discrepancies between the pre-analysis plan and the published paper, and calculates the estimated power of the experimental design. Section 3 reproduces the regression tables in the original paper, and points out some minor discrepancies between the procedures presented in the original paper and the code. Section 4 provides robustness checks with respect to adjusting p-values for multiple hypothesis testing, removing control variables, and allowing alternative ways of constructing indices and dealing with non-response. Section 5 concludes.

2 Methodology Discussion

Before discussing the paper's results, we feel that three methodological notes are in order.

First, while we commend the authors on detailing their deviations from the pre-analysis plan, we think it is likewise important for any pre-registered paper to note departures from the pre-registration of the design. We list these deviations here.

In the AEARCTR-0000072 document, as of October 27, 2022, the following statement mismatches the published paper:

- According to the pre-registration abstract, the objective of the intervention was to decrease support for sex-selective abortion and measure spillovers on the participants' siblings. While these outcomes will be measured in future follow-ups (as confirmed to us by the authors), the current paper does not address this issue. For the future, we would recommend that researchers either pre-register which outcomes will be measured at which points in time, or make it clearer in the pre-analysis plan why not all pre-registered outcomes are going to be collected (yet).

Regarding deviations from the pre-analysis plan (not mentioned in the report that details these deviations), the following four statements mismatch the published paper:

- The paper does not present results without control variables, even though it was mentioned in the pre-analysis plan as a robustness check.
- The pre-analysis plan from end line 1 mentions that the gender behavior index will "average" responses and that primary outcomes will be inverse variance weighted averages, following Anderson (2008, J. Am. Stat. Assoc., Appendix A). Such indexes indeed average each respondent's responses, with weights computed to "maximise the amount of information captured in the index". However, the pre-analysis plan dated November 2016 specifies that results from unweighted "simple" averages should also be reported, which is not the case in the published article. The authors' published code does however calculate the unweighted indices.
- The self-efficacy index mentioned in the pre-analysis plan was shortened and turned into a self-esteem index for girls. This change is not mentioned anywhere.
- The paper does not compare the two versions of scholarship applications that are mentioned in the pre-analysis plan; the data appendix omits that a secondary analysis was planned.

Second, neither the pre-registration nor the pre-analysis plan detail ex-ante power calculations. Since these are important for putting a paper's result in perspective (Maniadis et al., 2014), we do so here in Table 1. Importantly, given that the pre-registration does not specify any blocking, we do not take it into account.

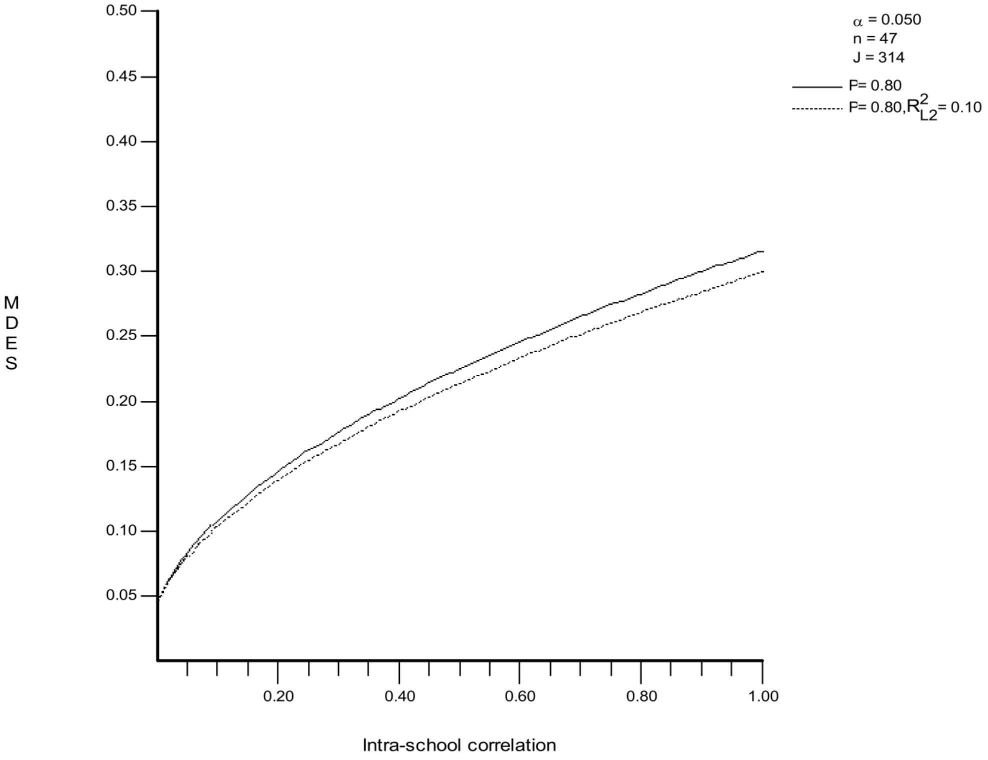
Since it is ex-ante unclear what intra-school correlation of student outcomes would be appropriate, we also include a sensitivity chart for different values of this correlation (keeping all other variables that enter this calculation the same as in Table 1). See Figure 1.

Table 1: Ex-ante Power Calculations

	(1)	(2)	(3)
School-level	✓		
Student-level, cluster		✓	✓
School-level controls			✓
MDES	0.3175	0.2282	0.2158

We report ex-ante calculations of the minimum detectable effect size (MDES) under alternative assumptions, using the sample size information provided in the AEA pre-registration plan. Throughout we assume significance $\alpha = 0.05$ and power of 80%. Column (1) is calculated using G*Power 3.1.9.2 (Faul et al., 2009) for a two-sided t-test with schools as independent observations. Columns (2)-(3) are calculated using Optimal Design 3.01 (Raudenbush et al., 2011), assuming 47 children per school and intra-school correlation of 0.5. Analysis with school-level controls assumes that these controls explain 10% of variation in outcomes. All Optimal Design calculations assume an equal number of schools in treatment and control, and so these calculations should be seen as approximations.

Figure 1: Minimum Detectable Effect Size for Student-level Data



For completeness, we also calculated the "ex-post" MDES (see Table 2), using the actual average number of children with non-missing data, the realized intra-school correlation for each outcome, and the actual explanatory power of variables used for stratification at baseline: district FEs, co-ed status of school, school size, and distance to the district headquarters. The article leaves some ambiguity in how background variables were used to stratify schools. To quote, *"The randomization was stratified by district, co-ed status of the school, school size, and distance to the district headquarters"* (p. 906). The final two variables, school size and distance to headquarters, take on many different values in the sample and we were not able to locate a specification of what thresholds were used to stratify on them. To approximate the stratification, for the purpose of calculating the MDES, the number of thresholds was chosen to reproduce the number of strata, and the threshold values were determined so as to yield strata of similar numbers of schools.

To evaluate both the ex-ante and the "ex-post" MDES calculation: the RCT seems well-powered to detect an effect commonly accepted as "small" (Serdar et al., 2021). The realized low intra-school correlations relative to our ex-ante analysis are unlikely to be a concern, as these have been shown in other studies, such as Lam et al. (2002); Sammons et al. (1993); Stockford (2009).

Table 2: Ex-post Power Calculations

	Gender attitudes index	Girls' aspirations index	Self-reported behavior index
Student-level, cluster	✓	✓	✓
Stratification controls	✓	✓	✓
MDES	0.0937	0.0876	0.1360

We report ex-post calculations of the minimum detectable effect size (MDES) for the three main outcome variables. Throughout we assume $\alpha = 0.05$ and power of 80%. Values are calculated using Optimal Design 3.01 (Raudenbush et al., 2011), using 45/25/44 children per school, realized intra-school correlation of 0.0652, 0.0358, 0.1709, and shares of 5.05, 4.47, and 4.44% of variation explained by the stratification variables for each 1st end line outcome. All Optimal Design calculations assume an equal number of schools in treatment and control, and so these calculations should be seen as approximations.

And third, we believe that the judgement whether there is a *"parsimonious"* (p. 912) set of outcomes and heterogeneity analyses should be left to the reader of the paper; therefore, throughout our analysis, we report both un-adjusted p-values and multiple-hypothesis-testing-adjusted p-values within tables. Again, we commend the authors for using the Bonferroni correction in Appendix Tables 9, 13, 14 (and 27, 28) for gender attitudes, aspirations, and self-reported behavior indices separately as a first step.

3 Reproducibility

As the first step in our replication, we show that the paper is reproducible; i.e., using the author-provided code, data, and attached ado-files while following the Readme text file instructions produces the same results as reported in the paper.[1]

The full set of regressions is omitted here, but is included as part of the robustness exercises presented in Section 4.1.

Our replication did not include a complete re-coding of the analysis from scratch. Thus, to note coding errors or discrepancies between article text and code, we depended on inspecting the code. Below is a list of locations where the code deviated from what was specified in the article, or where the code clearly deviated from the authors' intent. No deviation materially affects the results.

- Table 6, Panel B and Table 12, Panel A present similar regressions on outcomes at end line 1 and 2. However, in contrast to the regression tables presented in the rest of the paper (and the other two panels), the parameters in these two panels condition on non-attrition at the wrong end line:
  - Table 6, Panel B (boys). The authors condition on no attrition on end line 2, although this is an end line 1 regression.
  - Table 12, Panel A (girls). The authors condition on no attrition on end line 1, although this is an end line 2 regression.
- Table 13:
  - The authors drop the baseline control in the first regression, which contradicts the note under the table in the paper.
  - The authors take the absolute value of the control group mean of the 4th outcome variable, switching the sign.

We present the original and corrected tables below: [2]

[1] Our additional coding for reproducibility and robustness will be available at https://github.com/ermafl/djj2022_replic

[2] Two of the p-values in the original article's Table 6, Panel C are also mildly affected. We thank the authors for pointing this out to us. No p-values in the original Table 12 are affected.

Table 3: (Boys) Treatment Effects on Perceptions of Social Norms (end line 1)
Authors' table and corrected table (Table 6, Panel B in the original paper)

	Social norms toward work			Social norms toward education		
	Student agrees:			Student agrees:		
	Women should be allowed to work	Community thinks women should be allowed to work	Women should be allowed to work and thinks community will not oppose them	Women should be allowed to study in college even if it is far away	Community thinks women should be allowed to study in college even if it is far away	Women should be allowed to study in college and thinks community will not oppose them
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1: Conditioning on non-missing at endline 2						
Treated	0.1960	0.0848	0.1196	0.1450	0.1019	0.1291
	(0.0200)	(0.0199)	(0.0199)	(0.0156)	(0.0191)	(0.0198)
Constant	0.6446	0.3552	0.3374	0.7693	0.6000	0.6033
	(0.0273)	(0.0216)	(0.0202)	(0.0203)	(0.0292)	(0.0272)
Control group mean	0.4965	0.3370	0.3155	0.7576	0.5572	0.5706
Number of students	2,863	2,691	2,672	2,995	2,847	2,833
Panel 2: Corrected						
Treated	0.1896	0.0828	0.1140	0.1420	0.1056	0.1315
	(0.0195)	(0.0195)	(0.0194)	(0.0153)	(0.0188)	(0.0195)
Constant	0.5293	0.2907	0.4566	0.7623	0.5798	0.5963
	(0.0208)	(0.0213)	(0.0294)	(0.0203)	(0.0241)	(0.0268)
Control group mean	0.4975	0.3353	0.3167	0.7568	0.5563	0.5682
Number of students	2,988	2,803	2,784	3,174	3,015	3,000

Table 4: Treatment Effects on Perceptions of Social Norms (end line 1)
Authors' table and corrected table (Table 6, Panel C in the original paper)

	Social norms toward work			Social norms toward education		
	Student agrees:			Student agrees:		
	Women should be allowed to work	Community thinks women should be allowed to work	Women should be allowed to work and thinks community will not oppose them	Women should be allowed to study in college even if it is far away	Community thinks women should be allowed to study in college even if it is far away	Women should be allowed to study in college and thinks community will not oppose them
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1: Conditioning on non-missing at endline 2						
Girls = Boys p-value	0.0000	0.0253	0.0026	0.0000	0.0013	0.0000
Panel 2: Corrected						
Girls = Boys p-value	0.0000	0.0328	0.0030	0.0000	0.0003	0.0000

Table 5: (Girls) Treatment Effects on Perceptions of Social Norms (end line 2)
Authors' table and corrected table (Table 12, Panel A in the original paper)

	Social norms toward work			Social norms toward education		
	Student agrees:			Student agrees:		
	Women should be allowed to work	Community thinks women should be allowed to work	Women should be allowed to work and thinks community will not oppose them	Women should be allowed to study in college even if it is far away	Community thinks women should be allowed to study in college even if it is far away	Women should be allowed to study in college and thinks community will not oppose them
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1: Conditioning on non-missing at endline 1						
Treated	0.0128	0.0053	0.0060	0.0113	-0.0092	-0.0106
	(0.0058)	(0.0187)	(0.0178)	(0.0076)	(0.0187)	(0.0170)
Constant	0.9785	0.6595	0.7217	0.9442	0.6424	0.6801
	(0.0075)	(0.0226)	(0.0262)	(0.0094)	(0.0228)	(0.0291)
Control group mean	0.9648	0.6432	0.7071	0.9503	0.6490	0.7120
Number of students	3,590	3,435	3,418	3,542	3,403	3,378
Panel 1: Corrected						
Treated	0.0132	0.0033	0.0043	0.0110	-0.0099	-0.0096
	(0.0057)	(0.0182)	(0.0174)	(0.0076)	(0.0186)	(0.0171)
Constant	0.9778	0.6622	0.7156	0.9564	0.6351	0.6760
	(0.0076)	(0.0220)	(0.0254)	(0.0092)	(0.0314)	(0.0293)
Control group mean	0.9637	0.6435	0.7072	0.9490	0.6477	0.7087
Number of students	3,693	3,529	3,512	3,629	3,487	3,461

Table 6: Treatment Effects on Other Secondary Outcomes (end line 2)
Authors' table and corrected table (Table 13 in the original paper)

	Girls' self-esteem	Girls' education	Marriage and fertility aspirations (girls)	Marriage and fertility aspirations (boys)	Girls' experienced harassment	Boys' perpetrated harassment (school-grade level)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1: Missing baseline control in regression (1) and wrong sign of control group mean in regression (4)						
Treated	**0.0858**	0.0580	0.0524	0.0470	0.0626	0.0601
	(0.0259)	(0.0329)	(0.0291)	(0.0279)	(0.0295)	(0.0623)
Constant	**-0.0312**	-0.0895	0.1099	-0.1592	-0.0559	0.0924
	(0.0422)	(0.0401)	(0.0461)	(0.0378)	(0.0434)	(0.0609)
Control group mean	0.0000	0.0000	0.1433	**0.1692**	0.0000	-0.0026
Number of students	7,341	7,566	7,369	5,919	7,314	504
Panel 1: Corrections in bold						
Treated	**0.0843**	0.0580	0.0524	0.0470	0.0626	0.0601
	(0.0260)	(0.0329)	(0.0291)	(0.0279)	(0.0295)	(0.0623)
Constant	**-0.0327**	-0.0895	0.1099	-0.1592	-0.0559	0.0924
	(0.0423)	(0.0401)	(0.0461)	(0.0378)	(0.0434)	(0.0609)
Control group mean	0.0000	0.0000	0.1433	**-0.1692**	0.0000	-0.0026
Number of students	7,341	7,566	7,369	5,919	7,314	504

4 Robustness

As the second step in our replication, we explore the robustness of the results to alternative regression specifications and hypothesis tests. Broadly, these checks can be arranged into four categories: (i) Use of different sets of control variables, (ii) multiple hypothesis testing, (iii) use of different imputation procedures, and (iv) alternative ways of constructing outcome indices.

In general, the authors' results are robust to our alternative specification and hypothesis tests. However, the robustness tests also reveal that the treatment is only a relatively minor input into what forms the subjects' views and shapes their behaviors. This is documented in low R^2 in regressions where we omit controls. To illustrate the somewhat weak relationships between treatment and outcomes particularly for students with the most non-egalitarian views, we have included kernel density plots of the main outcomes by treatment status (see Figure 2).

Section 4.1 documents robustness with respect to removing control variables and adjusting for multiple hypothesis testing within tables. This section also documents the weak relationship between outcomes and treatment. Section 4.2 reports the main results when indices are constructed after imputing individual survey questions at the school average level rather than the district level. Lastly, Section 4.3 reports main results using alternative index construction procedures.

For transparency and ease of orientation, Table 7 documents which regressions use constructed indices as outcome variables and which controls are used in the original paper. Indices are constructed from individual survey questions using a relevant subset of survey responses. Whenever respondents fail to answer some of these questions, the authors impute these at the district level and include indicator variables to control for non-response to these questions in their regressions. In addition, each regression controls for analogous outcomes at the baseline whenever these are available. Gender-specific district fixed effects and gender-specific grade effects are present in all regressions.

Table 7: All regressions and control variables

Regression					Control variables				
Table	Regression	Outcome is an index	End line	Sample	Baseline outcome	Missing flags	Parental × other controls	District × Gender	Grades × Gender
2	(1)	✓	1	All	✓	✓		✓	✓
2	(2)	✓	1	Girls	✓	✓		✓	✓
2	(3)	✓	1	All	✓	✓		✓	✓
3	(1)	✓	1	All	✓	✓		✓	✓
3	(2)	✓	1	Girls	✓	✓		✓	✓
3	(3)	✓	1	All	✓	✓		✓	✓
4	(1)	✓	1	Girls	✓	✓		✓	✓
4	(2)	✓	1	Boys	✓	✓		✓	✓
4	(3)	✓	1	Girls	✓	✓		✓	✓
4	(4)	✓	1	Boys	✓	✓		✓	✓
5	(1)	✓	1	All	✓	✓	✓	✓	✓
5	(2)	✓	1	Girls	✓	✓	✓	✓	✓
5	(3)	✓	1	All	✓	✓	✓	✓	✓
6	(1)		1	By gender				✓	✓
6	(2)		1	By gender				✓	✓
6	(3)		1	By gender				✓	✓
6	(4)		1	By gender				✓	✓
6	(5)		1	By gender				✓	✓
6	(6)		1	Girls				✓	✓
7	(1)	✓	1	Girls	✓	✓		✓	✓
7	(2)	✓	1	By gender		✓		✓	✓
7	(3)		1	By gender	✓	✓		✓	✓
7	(4)		1	By gender	✓	✓		✓	✓
8	(1)	✓	2	All	✓	✓		✓	✓
8	(2)	✓	2	Girls	✓	✓		✓	✓
8	(3)	✓	2	All	✓	✓		✓	✓
8	(4)	✓	2	Girls				✓	✓
8	(5)	✓	2	All				✓	✓
9	(1)	✓	2	All	✓	✓		✓	✓
9	(2)	✓	2	Girls	✓	✓		✓	✓
9	(3)	✓	2	All	✓	✓		✓	✓
9	(4)	✓	2	Girls				✓	✓
9	(5)	✓	2	All				✓	✓
10	(1)	✓	2	Girls	✓	✓		✓	✓
10	(2)	✓	2	Boys	✓	✓		✓	✓
10	(3)	✓	2	Girls	✓	✓		✓	✓
10	(4)	✓	2	Boys	✓	✓		✓	✓
10	(5)		2	Girls				✓	✓
10	(6)		2	Boys				✓	✓
11	(1)		2	Girls				✓	✓
11	(2)		2	Girls				✓	✓
11	(3)		2	Girls				✓	✓
12	(1)		2	By gender				✓	✓
12	(2)		2	By gender				✓	✓
12	(3)		2	By gender				✓	✓
12	(4)		2	By gender				✓	✓
12	(5)		2	By gender				✓	✓
12	(6)		2	By gender				✓	✓
13	(1)	✓	2	Girls	✓	✓		✓	✓
13	(2)	✓	2	Girls		✓		✓	✓
13	(3)	✓	2	All				✓	✓
13	(4)	✓	2	All				✓	✓
13	(5)	✓	2	All		✓		✓	✓
13	(6)		2	All				✓	✓

Note: This table gives an overview of all of the regressions presented in the original paper. Missing flags are sets of control variables indicating that answers to individual survey questions were missing and imputed at the district×gender average. All regressions use measurements at end line 1 or end line 2 as outcomes and include analogous measurements at the baseline as control variables when available. District×gender and grades×gender controls are included in all regressions, or just grade and district when regressions are performed within one gender. The regressions in Table 5 investigate interactions between treatment and parental attitudes at the baseline and add interaction terms between baseline parental attitudes and all other control variables.

4.1 Robustness: Control Variables and Multiple Hypothesis Testing

Tables 8 to 22 reproduce all the regression tables in the original paper supplemented with estimates without any additional controls and with adjustments for multiple hypothesis testing.

Control variables: If treatment is successfully randomized, adding control variables may increase precision but not significantly alter the point estimates. Similarly, if survey non-response and attrition are random, adding controls for non-response should not alter the coefficients in the regression. To address the concern of randomization of treatment and attrition/non-response, we inspect the robustness of the results to removing all control variables. As mentioned previously, this is a robustness check the authors pre-registered but did not provide. Figure 3 in the Appendix shows that most missing-value-flags at end line are not predictable from baseline indices, which is some evidence that attrition rates are mostly independent of initial attitudes. Of course, this is only a preliminary test of possible attrition bias.

Multiple hypothesis testing: The paper includes results from a total of 54 regressions (excluding tables in the appendices). If we for simplicity assume that the treatment does not affect outcomes and that all outcomes are independent, the probability of falsely rejecting at least one null-hypothesis using a significance level of 5% level is $1 - 0.95^{54} = 94\%$. To address this concern, we use two distinct methods to adjust for multiple hypothesis testing. First, we follow List et al. (2019) by constructing adjusted p-values that adjust for familywise error rate (FWER), meaning the probability of making any type I error.[3] Secondly, we follow Young (2019) by directly testing the null hypothesis of no treatment effect across regressions.[4] The results of these alternative specification tests are also reported in Section 4.1.

The p-values in parentheses are "traditional" p-values, while the p-values in brackets are adjusted to account for multiple hypothesis testing (FWER p-values). These adjustments are made horizontally, meaning that in regressions with multiple parameters of interest, the p-value corrections are performed independently for each estimator (e.g., the adjusted p-values under *Treated* in Table 8 account for testing treatment in three different outcomes, but does not account for testing the two other variables).

In addition, cross-regression hypothesis testing is reported at the end of each table. These p-values correspond to the joint hypothesis that none of the estimated parameters are significant, i.e., $\beta_1 = \beta_2 = \beta_3 = 0$, where β_i is the parameter of a single variable across the different regressions. These p-values are calculated using the Stata package mhtreg.

Traditional and adjusted p-values, as well as the cross-regression hypothesis tests are clustered at the school level. Both methods used to adjust for multiple hypothesis testing rely on bootstrapping. We perform $10,000$ bootstrap

[3] We use the Stata package mhtreg developed by Steinmayr (2020).

[4] We use the Stata package randcmd, also developed by Young.

replications, using the same seed within tables.

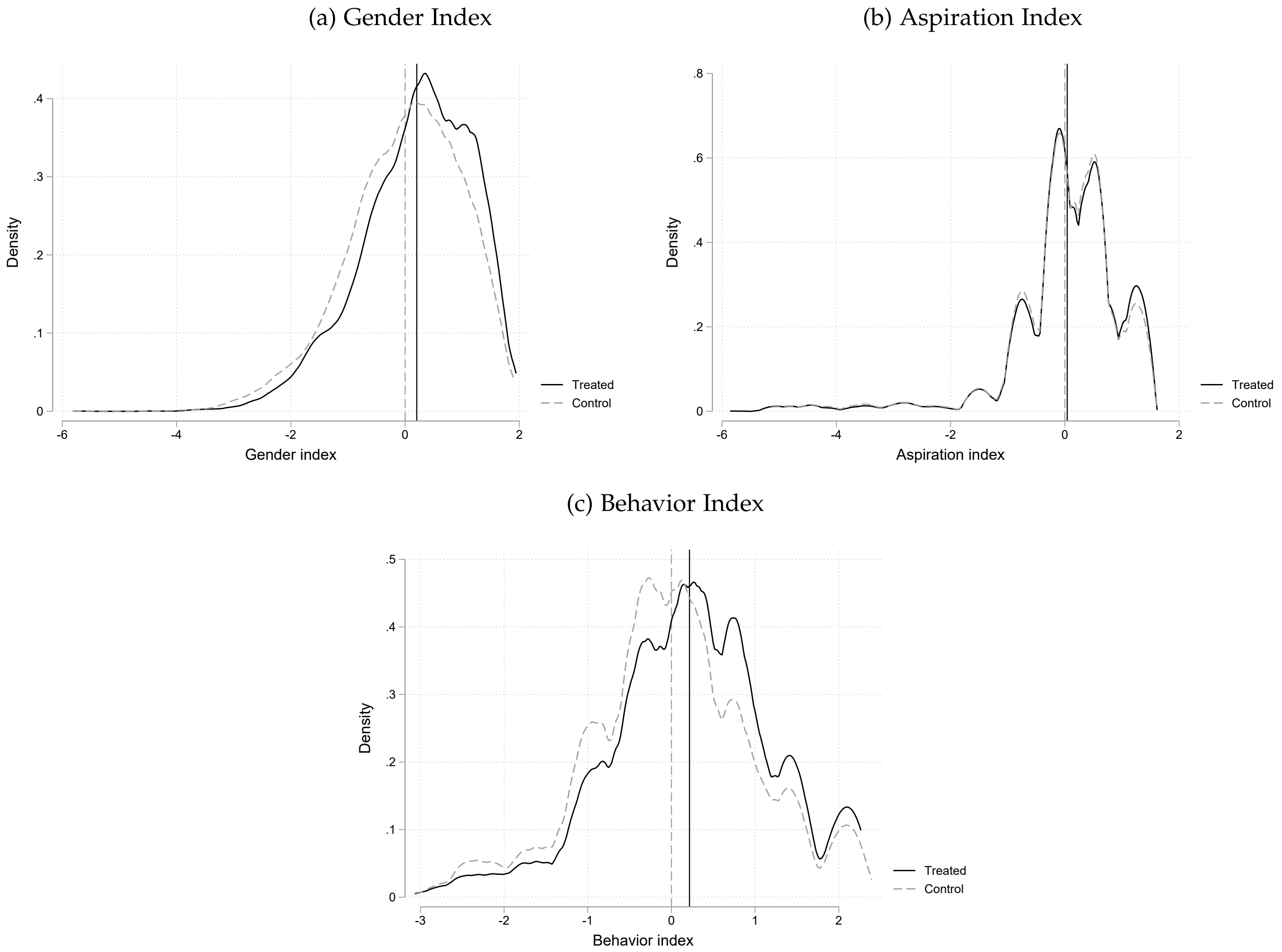
Weak relationship between treatment and outcome: The somewhat weak relationships between treatment and outcomes are illustrated by the low R^2 in the tables below. In general, treatment explains between 1 and 3 percent of indices, with some larger values when additional interactions are added. The distribution of the three main outcome indices used in Table 8 are presented in Figure 2. The figure shows clear differences in outcomes by treatment status, except for the aspiration index.

Table 8: Treatment Effects on Attitudes, Aspirations, and Behavior (end line 1) (Table 2 in the original paper)

	Gender attitudes index	Girls' aspirations index	Self-reported behavior index
	(1)	(2)	(3)
Panel 1: Without additional controls			
Treated	0.2040	0.0410	0.2161
	(0.0000)	(0.1932)	(0.0000)
	[0.0001]	[0.1944]	[0.0001]
Constant	-0.0000	0.0000	0.0000
	(1.0000)	(1.0000)	(1.0000)
Number of students	13,987	7,767	13,974
R^2	0.0107	0.0004	0.0117
P-value, joint hypothesis	0.0001		
Panel 2: With author's controls			
Treated	0.1797	0.0295	0.1964
	(0.0000)	(0.2222)	(0.0000)
	[0.0001]	[0.2306]	[0.0001]
Constant	0.0953	0.0294	-0.7999
	(0.4556)	(0.3584)	(0.0001)
Number of students	13,987	7,767	13,974
R^2	0.1097	0.2313	0.3428
P-value, joint hypothesis	0.0001		

Note: This table reproduces Table 2 in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* is the p-value corresponding to the joint null hypothesis of no treatment effect in any of the regressions in the panel. Traditional and FWER-adjusted p-values as well as the joint hypothesis test are clustered at the school level.

Figure 2: Distribution of Attitude-, Aspiration-, and Behavior index by treatment status (end line 1)



(a) Gender Index

(b) Aspiration Index

(c) Behavior Index

These figures show the full distribution of the three main index at end line 1 by treatment status. The vertical lines are average values. I.e., the difference between the horizontal line is equal to the point estimates in Table 8, Panel 1 (without controls). All figures are constructed using kernel density plots with Epanechnikov kernels, using a bandwith of 0.15.

Table 9: Robustness Check for Social Desirability Bias (end line 1)
(Table 3 in the original paper)

	Gender	Girls’	Self-reported
	attitudes index	aspirations index	behavior index
	(1)	(2)	(3)
Panel 1: Without additional controls			
Treated	0.2158	0.0424	0.2174
	(0.0000)	(0.2675)	(0.0000)
	[0.0001]	[0.2674]	[0.0001]
High social desirability (Soc. D) score	0.1602	0.1013	0.0612
	(0.0000)	(0.0042)	(0.0103)
	[0.0001]	[0.0085]	[0.0114]
Treated X High soc. D score	-0.0283	-0.0012	-0.0020
	(0.3970)	(0.9822)	(0.9541)
	[0.7802]	[0.9841]	[0.9981]
Constant	-0.0606	-0.0405	-0.0231
	(0.0053)	(0.1276)	(0.2412)
p-value: Treated + Treated x High Soc. D=0	0.0000	0.3434	0.0000
Number of students	13,987	7,767	13,974
R^2	0.0159	0.0029	0.0126
P-value, joint hypotheses:			
- Treated	0.0001		
- High social desirability (Soc. D) score	0.0001		
- Treated X High soc. D score	0.9999		
Panel 2: With author’s controls			
Treated	0.1900	0.0179	0.1964
	(0.0000)	(0.5410)	(0.0000)
	[0.0001]	[0.5439]	[0.0001]
High social desirability (Soc. D) score	0.1057	0.0616	0.0596
	(0.0000)	(0.0406)	(0.0016)
	[0.0001]	[0.0433]	[0.0031]
Treated X High soc. D score	-0.0244	0.0316	0.0014
	(0.4107)	(0.4678)	(0.9585)
	[0.7932]	[0.7195]	[0.9558]
Constant	0.0607	0.0058	-0.8205
	(0.6364)	(0.8620)	(0.0000)
p-value: Treated + Treated x High Soc. D=0	0.0000	0.1713	0.0000
Number of students	13,987	7,767	13,974
R^2	0.1118	0.2328	0.3437
P-value, joint hypotheses:			
- Treated	0.0001		
- High social desirability (Soc. D) score	0.0001		
- Treated X High soc. D score	0.9999		

Note: This table reproduces Table 3 in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypotheses* are the p-values corresponding to the joint null hypotheses of no treatment effect in any of the regressions in the panel, performed independently for the three tests. Traditional and FWER-adjusted p-values as well as the joint hypotheses tests are clustered at the school level.

Table 10: Gender-Specific Treatment Effects on Attitudes, Aspirations, and Behavior (end line 1)
(Table 4 in the original paper)

	Gender attitudes index		Self-reported behavior index	
	Girls	Boys	Girls	Boys
	(1)	(2)	(3)	(4)
Panel 1: Without additional controls				
Treated	0.1736	0.2106	0.1400	0.3304
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	[0.0001]	[0.0001]	[0.0001]	[0.0001]
Constant	0.2368	-0.2831	-0.0857	0.1025
	(0.0000)	(0.0000)	(0.0000)	(0.0001)
Number of students	7,802	6,185	7,794	6,180
R^2	0.0090	0.0110	0.0072	0.0203
P-value, joint hypothesis	0.0000			
Panel 2: With author's controls				
Treated	0.1611	0.2036	0.1419	0.2595
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	[0.0001]	[0.0001]	[0.0001]	[0.0001]
Constant	0.6930	-0.0613	-0.1854	-0.6520
	(0.0000)	(0.7813)	(0.4793)	(0.0313)
Number of students	7,802	6,185	7,794	6,180
R^2	0.0602	0.0427	0.2539	0.3891
P-value, joint hypothesis	0.0000			

Note: This table reproduces Table 4 in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* is the p-value corresponding to the joint null hypothesis of no treatment effect in any of the regressions in the panel. Traditional and FWER-adjusted p-values as well as the joint hypothesis test are clustered at the school level.

Table 11: Heterogeneous Effects by Parent Attitudes (end line 1)
(Table 5 in the original paper)

	Gender attitudes index	Girls' aspirations index	Self-reported behavior index
	(1)	(2)	(3)
Panel 1: Without additional controls			
Treated	0.2025	0.0713	0.2112
	(0.0000)	(0.0715)	(0.0000)
	[0.0001]	[0.0695]	[0.0001]
Treated x baseline parent attitudes	0.0197	-0.0112	-0.0138
	(0.4763)	(0.7296)	(0.5826)
	[0.8466]	[0.7356]	[0.8294]
Baseline Parent Gender Attitudes Index	0.0825	0.0515	0.0612
	(0.0000)	(0.0390)	(0.0003)
Constant	0.0205	0.0093	0.0128
	(0.3901)	(0.7583)	(0.5496)
Number of students	5,718	3,231	5,717
R^2	0.0197	0.0036	0.0143
P-value, joint hypotheses:			
- Treated	0.0001		
- Treated × baseline parent attitudes	0.0423		
Panel 2: With author's controls			
Treated	0.1741	0.0535	0.1795
	(0.0000)	(0.1068)	(0.0000)
	[0.0001]	[0.1118]	[0.0001]
Treated x baseline parent attitudes	0.0256	0.0003	-0.0391
	(0.2984)	(0.9899)	(0.0622)
	[0.4964]	[0.9891]	[0.1874]
Constant	0.4292	0.1485	0.3240
	(0.0577)	(0.0061)	(0.2202)
Number of students	5,718	3,231	5,717
R^2	0.1296	0.2183	0.3476
P-value, joint hypotheses:			
- Treated	0.0001		
- Treated × baseline parent attitudes	0.8725		

Note: This table reproduces Table 5 in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypotheses* are the p-values corresponding to the joint null hypotheses of no treatment effect in any of the regressions in the panel, performed independently for treatment and the interaction. Traditional and FWER-adjusted p-values as well as the joint hypotheses tests are clustered at the school level.

Table 12: (Girls) Treatment Effects on Perceptions of Social Norms (end line 1)
(Table 6, Panel A in the original paper)

	Social norms toward work			Social norms toward education		
	Student agrees:			Student agrees:		
	Women should be allowed to work	Community thinks women should be allowed to work	Women should be allowed to work and thinks community will not oppose them	Women should be allowed to study in college even if it is far away	Community thinks women should be allowed to study in college even if it is far away	Women should be allowed to study in college and thinks community will not oppose them
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1: Without additional controls						
Treated	0.0842	0.0267	0.0369	0.0365	0.0149	0.0138
	(0.0000)	(0.1340)	(0.0291)	(0.0000)	(0.4148)	(0.4321)
	[0.0001]	[0.3016]	[0.0906]	[0.0001]	[0.5396]	[0.4468]
Constant	0.8477	0.5175	0.5865	0.9350	0.6229	0.6949
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of students	3,874	3,661	3,625	3,900	3,737	3,717
R^2	0.0180	0.0007	0.0014	0.0074	0.0002	0.0002
P-value, joint hypothesis	0.0000					
Panel 2: With author's controls						
Treated	0.0829	0.0279	0.0399	0.0376	0.0153	0.0152
	(0.0000)	(0.1069)	(0.0115)	(0.0000)	(0.3904)	(0.3730)
	[0.0001]	[0.2486]	[0.0391]	[0.0001]	[0.4052]	[0.4964]
Constant	0.8814	0.4977	0.6844	0.9566	0.6066	0.6776
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of students	3,874	3,661	3,625	3,900	3,737	3,717
R^2	0.0268	0.0055	0.0098	0.0112	0.0044	0.0051
P-value, joint hypothesis	0.0000					

Note: This table reproduces Table 6, Panel A in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* is the p-value corresponding to the joint null hypothesis of no treatment effect in any of the regressions in the panel. Traditional and FWER-adjusted p-values as well as the joint hypothesis test are clustered at the school level.

Table 13: (Boys) Treatment Effects on Perceptions of Social Norms (end line 1)
(Table 6, Panel B in the original paper)

	Social norms toward work			Social norms toward education		
	Student agrees:			Student agrees:		
	Women should be allowed to work	Community thinks women should be allowed to work	Women should be allowed to work and thinks community will not oppose them	Women should be allowed to study in college even if it is far away	Community thinks women should be allowed to study in college even if it is far away	Women should be allowed to study in college and thinks community will not oppose them
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1: Without additional controls						
Treated	0.1922	0.0842	0.1158	0.1425	0.1056	0.1308
	(0.0000)	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	[0.0001]	[0.0002]	[0.0001]	[0.0001]	[0.0001]	[0.0001]
Constant	0.4975	0.3353	0.3167	0.7568	0.5563	0.5682
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of students	2,988	2,803	2,784	3,174	3,015	3,000
R^2	0.0377	0.0075	0.0143	0.0347	0.0116	0.0183
P-value, joint hypothesis	0.0000					
Panel 2: With author's controls						
Treated	0.1896	0.0828	0.1140	0.1420	0.1056	0.1315
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	[0.0001]	[0.0001]	[0.0001]	[0.0001]	[0.0001]	[0.0001]
Constant	0.5293	0.2907	0.4566	0.7623	0.5798	0.5963
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of students	2,988	2,803	2,784	3,174	3,015	3,000
R^2	0.0536	0.0202	0.0305	0.0390	0.0167	0.0268
P-value, joint hypothesis	0.0000					

Note: This table reproduces Table 6, Panel B in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* is the p-value corresponding to the joint null hypothesis of no treatment effect in any of the regressions in the panel. Both panels condition on non-attrition at end line 1 and not at end line 2 as in the original paper. See section 3. Traditional and FWER-adjusted p-values as well as the joint hypothesis test are clustered at the school level.

Table 14: (Girls) Treatment Effects on Other Secondary Outcomes (end line 1) (Table 7, Panel A in the original paper)

	Girls' self-esteem	Awareness of gender-based discrimination	IAT: associates girls with positive words	IAT: associates women with market work
	(1)	(2)	(3)	(4)
Panel 1: Without additional controls				
Treated	0.1025	0.1062	0.0016	-0.1000
	(0.0000)	(0.0031)	(0.9725)	(0.2216)
	[0.0001]	[0.0001]	[0.0001]	[0.0001]
Constant	0.0000	0.0989	0.4077	0.0002
	(1.0000)	(0.0001)	(0.0000)	(0.9972)
Number of students	7,788	7,777	1,676	1,830
R^2	0.0028	0.0031	0.0000	0.0024
P-value, joint hypothesis	0.0000			
Panel 2: With author's controls				
Treated	0.1037	0.0534	-0.0062	-0.0789
	(0.0000)	(0.0105)	(0.8948)	(0.2884)
	[0.0001]	[0.0001]	[0.0001]	[0.0001]
Constant	0.0142	0.4203	0.4379	0.4237
	(0.6758)	(0.0000)	(0.0000)	(0.0000)
Number of students	7,788	7,777	1,676	1,830
R^2	0.0138	0.4785	0.0255	0.0562
P-value, joint hypothesis	0.0000			

Note: This table reproduces Table 7, Panel A in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* is the p-value corresponding to the joint null hypothesis of no treatment effect in any of the regressions in the panel.

Table 15: (Boys) Treatment Effects on Other Secondary Outcomes (end line 1) (Table 7, Panel B in the original paper)

	Awareness of gender-based discrimination	IAT: associates girls with positive words	IAT: associates women with market work
	(1)	(2)	(3)
Panel 1: Without additional controls			
Treated	0.0573	0.0152	-0.0071
	(0.0782)	(0.7536)	(0.9145)
	[0.0001]	[0.0001]	[0.0001]
Constant	-0.1184	-0.5145	-0.0003
	(0.0000)	(0.0000)	(0.9954)
Number of students	6,162	1,250	1,368
R^2	0.0008	0.0001	0.0000
P-value, joint hypothesis	0.3637		
Panel 2: With author's controls			
Treated	0.0065	0.0144	-0.0045
	(0.7473)	(0.7638)	(0.9438)
	[0.0001]	[0.0001]	[0.0001]
Constant	0.4333	-0.5670	-0.0264
	(0.0000)	(0.0000)	(0.7442)
Number of students	6,162	1,250	1,368
R^2	0.5398	0.0158	0.0290
P-value, joint hypothesis	0.9819		

Note: This table reproduces Table 7, Panel B in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* is the p-value corresponding to the joint null hypothesis of no treatment effect in any of the regressions in the panel. Traditional and FWER-adjusted p-values as well as the joint hypothesis test are clustered at the school level.

Table 16: Treatment Effects on Attitudes, Aspirations, and Behavior (end line 2) (Table 8 in the original paper)

	Gender attitudes index	Girls' aspirations index	Self-reported behavior index	Applied to scholarship	Signed petition
	(1)	(2)	(3)	(4)	(5)
Panel 1: Without additional controls					
Treated	0.1882	-0.0070	0.2230	0.0308	0.0147
	(0.0000)	(0.8188)	(0.0000)	(0.0947)	(0.1643)
	[0.0001]	[0.8196]	[0.0001]	[0.2367]	[0.2923]
Constant	0.3326	-0.0000	0.0000	0.4078	0.1500
	(0.0000)	(1.0000)	(1.0000)	(0.0000)	(0.0000)
Number of students	13,679	7,560	13,677	7,347	13,303
R^2	0.0097	0.0000	0.0122	0.0010	0.0004
P-value, joint hypothesis	0.0000				
Panel 2: With author's controls					
Treated	0.1597	-0.0246	0.2271	0.0314	0.0121
	(0.0000)	(0.1943)	(0.0000)	(0.0669)	(0.1614)
	[0.0001]	[0.2005]	[0.0001]	[0.1820]	[0.2916]
Constant	0.2157	0.2205	0.1196	0.4782	0.1884
	(0.0003)	(0.0000)	(0.0257)	(0.0000)	(0.0000)
Number of students	13,679	7,560	13,677	7,347	13,303
R^2	0.1076	0.4642	0.0513	0.0107	0.0317
P-value, joint hypothesis	0.0000				

Note: This table reproduces Table 8 in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* is the p-value corresponding to the joint null hypothesis of no treatment effect in any of the regressions in the panel. Traditional and FWER-adjusted p-values as well as the joint hypothesis test are clustered at the school level.

Table 17: Robustness Check for Social Desirability Bias (end line 2)
(Table 9 in the original paper)

	Gender attitudes index	Girls' aspirations index	Self-reported behavior index	Applied to scholarship	Signed petition
	(1)	(2)	(3)	(4)	(5)
Panel 1: Without additional controls					
Treated	0.1728	-0.0097	0.2330	0.0330	0.0219
	(0.0000)	(0.7846)	(0.0000)	(0.1031)	(0.0567)
	[0.0001]	[0.7842]	[0.0001]	[0.1958]	[0.1618]
High social desirability (Soc. D) score	0.1146	0.0947	0.0646	0.0194	0.0178
	(0.0000)	(0.0056)	(0.0080)	(0.2720)	(0.0332)
	[0.0001]	[0.0243]	[0.0229]	[0.2790]	[0.0611]
Treated X High soc. D score	0.0435	0.0089	-0.0258	-0.0051	-0.0191
	(0.1813)	(0.8547)	(0.4591)	(0.8383)	(0.1388)
	[0.5422]	[0.8531]	[0.8439]	[0.9747]	[0.5132]
Constant	0.2892	-0.0380	-0.0245	0.4000	0.1433
	(0.0000)	(0.1403)	(0.1799)	(0.0000)	(0.0000)
p-value: Treated + Treated x High Soc. D=0	0.0000	0.9841	0.0000	0.2592	0.8341
Number of students	13,679	7,560	13,677	7,347	13,303
R^2	0.0145	0.0024	0.0129	0.0013	0.0007
P-value, joint hypotheses:					
- Treated	0.0000				
- High social desirability score	0.0000				
- Treated × High Soc. D score	0.0000				
Panel 2: With author's controls					
Treated	0.1502	-0.0339	0.2353	0.0338	0.0200
	(0.0000)	(0.1525)	(0.0000)	(0.0798)	(0.0421)
	[0.0001]	[0.1561]	[0.0001]	[0.1544]	[0.1299]
High social desirability (Soc. D) score	0.0697	0.0294	0.0585	0.0168	0.0140
	(0.0014)	(0.2320)	(0.0155)	(0.3387)	(0.0904)
	[0.0086]	[0.4259]	[0.0635]	[0.3473]	[0.2368]
Treated X High soc. D score	0.0277	0.0243	-0.0209	-0.0057	-0.0208
	(0.3650)	(0.4783)	(0.5415)	(0.8189)	(0.1019)
	[0.8441]	[0.8644]	[0.7906]	[0.8233]	[0.4124]
Constant	0.1915	0.2083	0.0998	0.4712	0.1837
	(0.0013)	(0.0000)	(0.0605)	(0.0000)	(0.0000)
p-value: Treated + Treated x High Soc. D=0	0.0000	0.7281	0.0000	0.2301	0.9458
Number of students	13,679	7,560	13,677	7,347	13,303
R^2	0.1094	0.4646	0.0519	0.0109	0.0319
P-value, joint hypotheses:					
- Treated	0.0000				
- High social desirability score	0.0000				
- Treated × High Soc. D score	0.0000				

Note: This table reproduces Table 9 in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypotheses* are the p-value corresponding to the joint null hypotheses of no treatment effect in any of the regressions in the panel, performed independently for the three tests. Traditional and FWER-adjusted p-values as well as the joint hypotheses tests are clustered at the school level.

Table 18: Gender-Specific Treatment Effects on Attitudes, Aspirations, and Behavior (end line 2)
(Table 10 in the original paper)

	Gender attitudes index		Self-reported behavior index		Signed petition	
	Girls	Boys	Girls	Boys	Girls	Boys
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1: Without additional controls						
Treated	0.1246	0.2408	0.1636	0.3097	0.0182	0.0048
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.2096)	(0.6525)
	[0.0001]	[0.0001]	[0.0001]	[0.0001]	[0.2119]	[0.6478]
Constant	0.5621	0.0630	-0.0672	0.0790	0.1895	0.1037
	(0.0000)	(0.0062)	(0.0001)	(0.0013)	(0.0000)	(0.0000)
Number of students	7,562	6,117	7,563	6,114	7,347	5,956
R^2	0.0050	0.0148	0.0105	0.0164	0.0005	0.0001
P-value, joint hypotheses:						
- Girls	0.0001					
- Boys	0.0001					
Panel 2: With author's controls						
Treated	0.1115	0.2179	0.1580	0.3112	0.0194	0.0031
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.1404)	(0.7415)
	[0.0001]	[0.0001]	[0.0001]	[0.0001]	[0.1444]	[0.7368]
Constant	0.5750	-0.1244	0.0697	0.2091	0.2649	0.0764
	(0.0000)	(0.1821)	(0.0964)	(0.0000)	(0.0000)	(0.0000)
Number of students	7,562	6,117	7,563	6,114	7,347	5,956
R^2	0.0609	0.0652	0.0795	0.0317	0.0156	0.0171
P-value, joint hypotheses:						
- Girls	0.0001					
- Boys	0.0001					

Note: This table reproduces Table 10 in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* are the p-values corresponding to the joint null hypotheses of no treatment effect in any of the regressions in the panel, performed independently for girls and boys. Traditional and FWER-adjusted p-values as well as the joint hypotheses tests are clustered at the school level.

Table 19: Unpacking the Treatment Effect on Scholarship Applications (end line 2) (Table 11 in the original paper)

	Applied to scholarship		
	(1)	(2)	(3)
Panel 1: Without additional controls			
Treated	0.0287	0.0139	-0.0255
	(0.1142)	(0.4906)	(0.3777)
	[0.2215]	[0.4907]	[0.5541]
Treated X BL aspiration index	0.0242		
	(0.0317)		
	[0.0500]		
Girls' aspirations index	0.0351		
	(0.0000)		
Treated x Above-median BL aspirations		0.0380	
		(0.1223)	
		[0.1263]	
B_Saspiration_index2_abm		0.0598	
		(0.0013)	
Treated x Has discussed educ goals with parent			0.0708
			(0.0124)
			[0.0344]
Student has discussed education goals with parent or adult relative			0.0029
			(0.8857)
Constant	0.4070	0.3847	0.4055
	(0.0000)	(0.0000)	(0.0000)
Number of students	7,347	7,347	7,347
R^2	0.0104	0.0074	0.0027
P-value, joint hypotheses:			
- Treated	0.0000		
- Interactions	0.0000		
Panel 2: With author's controls			
Treated	0.0291	0.0142	-0.0234
	(0.0873)	(0.4590)	(0.3835)
	[0.1760]	[0.4550]	[0.5732]
Treated X BL aspiration index	0.0222		
	(0.0446)		
	[0.0682]		
Treated x Above-median BL aspirations		0.0396	
		(0.1038)	
		[0.1078]	
Treated x Has discussed educ goals with parent			0.0683
			(0.0141)
			[0.0390]
Constant	0.4766	0.4543	0.4692
	(0.0000)	(0.0000)	(0.0000)
Number of students	7,347	7,347	7,347
R^2	0.0195	0.0160	0.0128
P-value, joint hypotheses:			
- Treated	0.0000		
- Interactions	0.0000		

Note: This table reproduces Table 11 in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypotheses* are the p-values corresponding to the joint null hypotheses of no treatment effects in any of the regressions in the panel, performed independently for treatment and the interactions. Traditional and FWER-adjusted p-values as well as the joint hypotheses tests are clustered at the school level.

Table 20: (Girls) Treatment Effects on Perceptions of Social Norms (end line 2)
(Table 12, Panel A in the original paper)

	Social norms toward work			Social norms toward education		
	Student agrees:			Student agrees:		
	Women should be allowed to work	Community thinks women should be allowed to work	Women should be allowed to work and thinks community will not oppose them	Women should be allowed to study in college even if it is far away	Community thinks women should be allowed to study in college even if it is far away	Women should be allowed to study in college and thinks community will not oppose them
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1: Without additional controls						
Treated	0.0128	0.0024	0.0036	0.0115	-0.0114	-0.0100
	(0.0293)	(0.8954)	(0.8362)	(0.1298)	(0.5515)	(0.5729)
	[0.1538]	[0.8930]	[0.9478]	[0.4224]	[0.9039]	[0.8701]
Constant	0.9637	0.6435	0.7072	0.9490	0.6477	0.7087
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of students	3,693	3,529	3,512	3,629	3,487	3,461
R^2	0.0014	0.0000	0.0000	0.0008	0.0001	0.0001
P-value, joint hypothesis	0.0000					
Panel 2: With author's controls						
Treated	0.0132	0.0033	0.0043	0.0110	-0.0099	-0.0096
	(0.0222)	(0.8554)	(0.8054)	(0.1456)	(0.5957)	(0.5753)
	[0.1261]	[0.8544]	[0.9294]	[0.4743]	[0.8952]	[0.9252]
Constant	0.9778	0.6622	0.7156	0.9564	0.6351	0.6760
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of students	3,693	3,529	3,512	3,629	3,487	3,461
R^2	0.0041	0.0038	0.0027	0.0013	0.0052	0.0053
P-value, joint hypothesis	0.0000					

Note: This table reproduces Table 12, Panel A in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* is the p-value corresponding to the joint null hypothesis of no treatment effect in any of the regressions in the panel. Both panels condition on non-attrition at end line 2 and not at end line 1 as in the original paper. See section 3. Traditional and FWER-adjusted p-values as well as the joint hypothesis test are clustered at the school level.

Table 21: (Boys) Treatment Effects on Perceptions of Social Norms (end line 2) (Table 12, Panel B in the original paper)

	Social norms toward work			Social norms toward education		
	Student agrees:			Student agrees:		
	Women should be allowed to work	Community thinks women should be allowed to work	Women should be allowed to work and thinks community will not oppose them	Women should be allowed to study in college even if it is far away	Community thinks women should be allowed to study in college even if it is far away	Women should be allowed to study in college and thinks community will not oppose them
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1: Without additional controls						
Treated	0.1215	0.0732	0.0959	0.0525	0.0303	0.0416
	(0.0000)	(0.0004)	(0.0000)	(0.0000)	(0.0973)	(0.0223)
	[0.0001]	[0.0012]	[0.0001]	[0.0001]	[0.1006]	[0.0376]
Constant	0.7472	0.5756	0.5774	0.8658	0.7080	0.7186
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of students	3,043	2,945	2,935	2,902	2,808	2,801
R^2	0.0232	0.0056	0.0097	0.0070	0.0011	0.0022
P-value, joint hypothesis	0.0000					
Panel 2: With author's controls						
Treated	0.1187	0.0700	0.0923	0.0505	0.0268	0.0378
	(0.0000)	(0.0003)	(0.0000)	(0.0001)	(0.1252)	(0.0268)
	[0.0001]	[0.0007]	[0.0001]	[0.0001]	[0.1278]	[0.0419]
Constant	0.6858	0.5953	0.4943	0.8365	0.7457	0.7763
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Number of students	3,043	2,945	2,935	2,902	2,808	2,801
R^2	0.0383	0.0184	0.0262	0.0168	0.0097	0.0135
P-value, joint hypothesis	0.0000					

Note: This table reproduces Table 12, Panel B in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* is the p-value corresponding to the joint null hypothesis of no treatment effect in any of the regressions in the panel. Traditional and FWER-adjusted p-values as well as the joint hypothesis test are clustered at the school level.

Table 22: Treatment Effects on Other Secondary Outcomes (end line 2)
(Table 13 in the original paper)

	Girls' self-esteem	Girls' education	Marriage and fertility aspirations (girls)	Marriage and fertility aspirations (boys)	Girls' experienced harassment	Boys' perpetrated harassment (school-grade level)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel 1: Without additional controls						
Treated	0.0878	0.0525	0.0529	0.0413	0.0611	0.0601
	(0.0008)	(0.1237)	(0.0735)	(0.1810)	(0.0413)	(0.3382)
	[0.0066]	[0.3120]	[0.2497]	[0.3359]	[0.1764]	[0.2852]
Constant	-0.0000	-0.0000	0.1433	-0.1692	-0.0000	-0.0026
	(1.0000)	(1.0000)	(0.0000)	(0.0000)	(1.0000)	(0.9526)
Number of students	7,341	7,566	7,369	5,919	7,314	504
R^2	0.0021	0.0007	0.0008	0.0004	0.0009	0.0020
P-value, joint hypothesis	0.0000					
Panel 2: With author's controls						
Treated	0.0843	0.0580	0.0524	0.0470	0.0626	0.0601
	(0.0013)	(0.0787)	(0.0734)	(0.0935)	(0.0346)	(0.3356)
	[0.0112]	[0.2540]	[0.2037]	[0.1802]	[0.1696]	[0.2876]
Constant	-0.0327	-0.0895	0.1099	-0.1592	-0.0559	0.0924
	(0.4405)	(0.0265)	(0.0178)	(0.0000)	(0.1982)	(0.1303)
Number of students	7,341	7,566	7,369	5,919	7,314	504
R^2	0.0071	0.0132	0.0034	0.0096	0.0162	0.0107
P-value, joint hypothesis	0.0000					

Note: This table reproduces Table 13 in the original paper, with additional regressions without controls and with hypothesis testing accounting for multiple hypotheses. P-values in parentheses are classical p-values, while p-values in brackets adjust for familywise error rates (FWER). *P-value, joint hypothesis* is the p-value corresponding to the joint null hypothesis of no treatment effect in any of the regressions in the panel. Regression (1) in Panel 2 does include baseline control, in contrast to the original paper. See section 3. Traditional and FWER-adjusted p-values as well as the joint hypothesis test are clustered at the school level.

4.2 Alternative Imputations

To compute the weights for the inverse variance indices, the authors make imputations for missing survey values, setting values equal to district-gender means. For the end lines, the imputations are also separated by treatment status. Different levels of imputation have their benefits and drawbacks, and it is uncertain which result in the most accurate estimates. This makes testing alternative imputations a natural robustness test. Changing the imputation to school-gender means instead of district-gender means makes virtually no difference to the main results (Tables 2 and 8 in the article), as can be seen from our Tables 23 and 24. The only appreciable difference is that the sample size for one index, the end line 1 behavior index, declines from nearly 14,000 to fewer than 10,000 because some combinations of school, gender and treatment status had no observed values to impute from.

Table 23: School vs district (original) imputations (end line 1)
(Table 2 in the original paper)

	Gender attitudes index		Girls' aspirations index		Self-reported behavior index	
Imputation	District (1)	School (2)	District (3)	School (4)	District (5)	School (6)
Treated	0.180*** [0.020]	0.180*** [0.020]	0.030 [0.024]	0.030 [0.024]	0.196*** [0.021]	0.243*** [0.029]
Control group mean	0.000	0.000	0.000	0.000	0.000	0.000
Basic controls	✓	✓	✓	✓	✓	✓
Number of students	13,987	13,987	7,767	7,767	13,974	9,607

Table 24: School vs district (original) imputations (end line 2)
(Table 8 in the original paper)

	Gender attitudes index		Girls' aspirations index		Self-reported behavior index	
Imputation	District (1)	School (2)	District (3)	School (4)	District (5)	School (6)
Treated	0.160*** [0.019]	0.160*** [0.019]	-0.024 [0.019]	-0.025 [0.019]	0.228*** [0.025]	0.227*** [0.025]
Control group mean	0.332	0.333	0.000	0.000	0.000	0.000
Basic controls	✓	✓	✓	✓	✓	✓
Number of students	13,679	13,679	7,560	7,560	13,677	13,677

4.3 Alternative Indices

The article uses inverse variance weights to compute indices, as was specified in the pre-analysis plan, to be the main outcomes of interest. Table 25 shows that the main results from the first end line are qualitatively consistent when estimated on principal factor indices (StataCorp (2017), entry on "mvfactor"), and hence remain robust to this alternative method of index construction. Table 26 shows that the second end line results are similarly qualitatively consistent. Those indices that were associated with statistically significant coefficients in the original article remain significant also with the principal factor indices. That the coefficients change somewhat in magnitude should be expected, and furthermore the changes are in different directions - some become larger, others shrink.

Table 25: Robustness of main results to factor-indices (end line 1)
(Table 2 in the original paper)

	Gender attitudes		Girls' aspirations		Behaviour	
	Authors' index	Factor index	Authors' index	Factor index	Authors' index	Factor index
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	0.180***	0.300***	0.030	0.036	0.225***	0.118***
	[0.020]	[0.023]	[0.024]	[0.027]	[0.024]	[0.014]
Control group mean	0.000	0.000	0.000	0.000	0.000	0.000
Basic controls	✓	✓	✓	✓	✓	✓
Number of students	13,987	13,987	7,767	7,802	13,974	13,987

Note: Indices are first principal factors. All indices standardised to mean 0 and standard deviations of 1, for comparability of coefficients.

Table 26: Robustness of main results to factor-indices (end line 2)
(Table 8 in the original paper)

	Gender attitudes		Girls' aspirations		Behaviour	
	Authors' index	Factor index	Authors' index	Factor index	Authors' index	Factor index
	(1)	(2)	(3)	(4)	(5)	(6)
Treated	0.160***	0.233***	-0.025	-0.004	0.227***	0.300***
	[0.019]	[0.023]	[0.019]	[0.031]	[0.025]	[0.031]
Control group mean	0.333	0.000	0.000	0.000	0.000	-0.000
Basic controls	✓	✓	✓	✓	✓	✓
Number of students	13,679	13,685	7,560	7,566	13,677	13,685

Note: All indices are standardised to means of 0 and standard deviations of 1, for comparability of coefficients. This also explains that the control group mean for the end line 2 gender attitudes is very near to zero, while the inverse weight index mean is 0.333 as reported in the article, because the authors compute this with the weights from the first end-line.

5 Conclusion

The key conclusion we draw from our work is that the paper of Dhar et al. (2022) is both reproducible and robust to alternative specifications. We commend the authors for their rigorous work and their commitment to open science by providing all documentation necessary for replications.

We did find a number of small inconsistencies or errors, but as we show, none of these materially affects the results or changes the interpretation of the paper. We further strengthen the paper's results by adding the authors' original pre-registered robustness check (regressions without controls), and show results remain unchanged even with alternative imputations of missing values or alternative definitions of indices.

To further aid the interpretation of the paper's results, we provide detailed power calculations and a short discussion cautioning that while the experiment is high-powered and the treatment effect significant, the treatment does explain relatively little variation in students' attitudes and behaviors. This may be related to the limited persistence of measured individual attitudes. Appendix Table 27 displays the correlation coefficients for each of the three main indices, between the baseline and the two endlines. Particularly the behavior index exhibits low correlation. This may suggest the measured attitudes are not as persistent as expected (cf. the authors' 2016 pre-analysis plan).

At any rate, we would welcome more work on the topic to gain more insights on the formation of social and gender norms and their stability over time.

Due to time constraints at the Replication Games, we did not pursue three things that we initially pre-registered: We did not check the reproducibility and robustness of results in the original paper's appendix, we did not explore heterogeneity of treatment results by background characteristics, and we did not use raw data as opposed to dichotomized variants of some variables produced by the authors. Nevertheless, we believe these would still be worthwhile exercises for future replication teams.

We also deviated from our pre-registration for practical reasons: we used the List et al. (2019) p-value correction for multiple hypothesis testing as opposed to the List et al. (2021) approach. The procedure proposed by List et al. (2019) is implemented as a Stata-package by Steinmayr (2020), allowing for clustering and variation in regressors across regressions.

References

Dhar, D., T. Jain, and S. Jayachandran (2022). Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in india. *American Economic Review 112*(3), 899–927.

Faul, F., E. Erdfelder, A. Buchner, and A.-G. Lang (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods 41*(4), 1149–1160.

Lam, Y. R., K.-c. Wong, and L.-m. Ho (2002). School effectiveness of a streamed-school system: A multilevel modelling of the hong kong secondary schools. *Australian Journal of Education 46*(3), 287–304.

List, J. A., A. M. Shaikh, A. Vayalinkal, et al. (2021). Multiple testing with covariate adjustment in experimental economics. Technical report, The Field Experiments Website.

List, J. A., A. M. Shaikh, and Y. Xu (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics 22*(4), 773–793.

Maniadis, Z., F. Tufano, and J. A. List (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review 104*(1), 277–90.

Raudenbush, S. W., J. Spybrook, R. Congdon, X. Liu, A. Martinez, H. Bloom, and C. Hill (2011). Optimal design software for multi-level and longitudinal research (version 3.01)[software].

Sammons, P., D. Nuttall, and P. Cuttance (1993). Differential school effectiveness: results from a reanalysis of the inner london education authority's junior school project data. *British Educational Research Journal 19*(4), 381–405.

Serdar, C. C., M. Cihan, D. Yücel, and M. A. Serdar (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia medica 31*(1), 27–53.

StataCorp (2017). *Stata Statistical Software: Release 15*. StataCorp LLC.

Steinmayr, A. (2020, October). MHTREG: Stata module for multiple hypothesis testing controlling for FWER. Statistical Software Components, Boston College Department of Economics.

Stockford, S. M. (2009). *Meta-analysis of intraclass correlation coefficients from multilevel models of educational achievement*. Arizona State University.

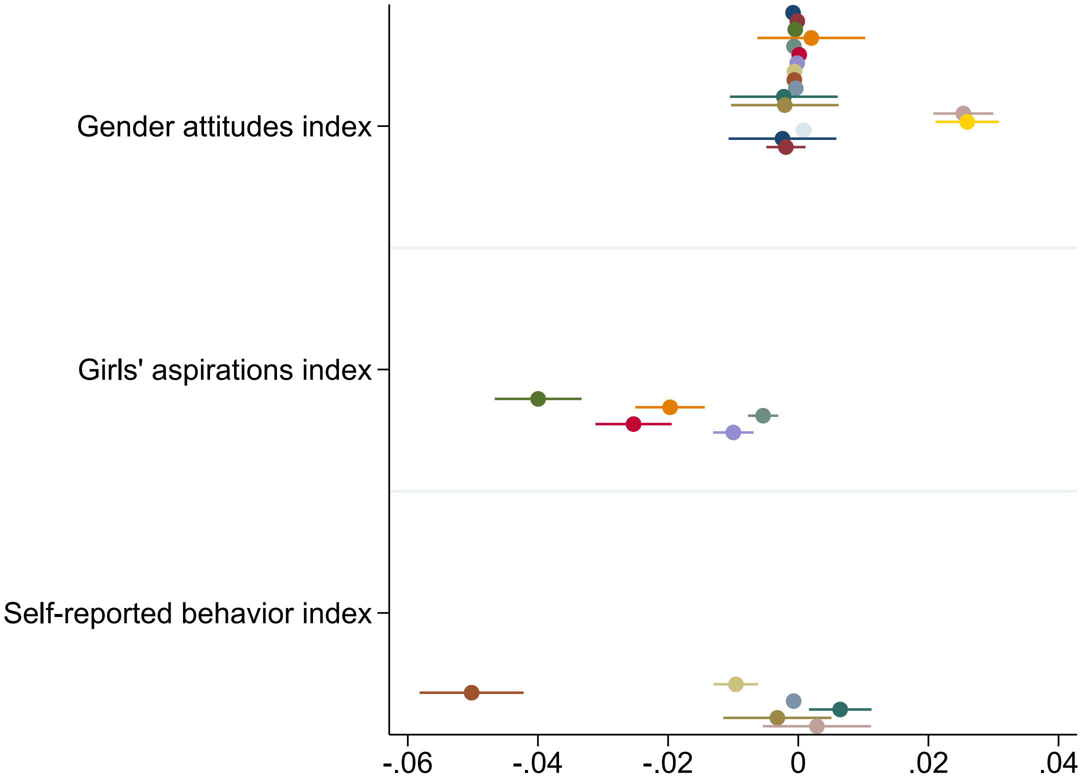
Young, A. (2019). Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. *The Quarterly Journal of Economics 134*(2), 557–598.

6 Appendix

Table 27: Intra-individual correlation ("persistence") of main indices

	Baseline	End Line 1	End Line 2
Panel 1: Gender index			
Baseline	1		
End line 1	0.1843	1	
End line 2	0.1951	0.3120	1
Panel 2: Aspiration index			
Baseline	1		
End line 1	0.2061	1	
End line 2	0.1701	0.3240	1
Panel 3: Behaviour index			
Baseline	1		
End line 1	0.1105	1	
End line 2	0.0932	0.2482	1

Figure 3: Baseline attitudes and end line 1 missing values



Note: This figure displays graphically the results of three sets of regressions: The coefficients are from single-variable regressions where the dependent variable is a missing-value-flag on an end line 1 survey item, and the independent variable is the corresponding baseline index. The most extreme values (i.e., the survey variables whose absence (flag) is the strongest related to index values) are, for each of the three indices, "Difference between boys' and girls' appropriate age to marry" (0.0259), "Expect to score above median at 10th marks" (-0.0400), "Willing to sit next to someone of opposite gender", (-0.0502).

7 Pre-analysis Plan

This document pre-specifies analyses for the *Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India* paper published in the AER (Dhar et al., 2022). The document was written and timestamped prior to the analyses conducted at the Oslo's Replication Games on Oct 27, 2022.

We split our replication into three parts: an analysis of reproducibility, a robustness analysis using the authors' own definitions and pre-specifications, and a robustness analysis using alternative definitions of concepts.

7.1 Reproducibility

We will re-run the code provided by the authors at the AEA website and make a note whether we are able to reproduce all results in the paper and the appendix; this includes all tables and in-text results.

We will also conduct power calculations based on the ex-ante pre-registered documents provided by the authors, showing sensitivity to different assumptions. We will contrast these to ex-post power calculations, taking into consideration the realized data.

Finally, we will note all deviations from the pre-registration, pre-analysis plan, and the document written by authors that lists all deviations.

7.2 Robustness

Throughout, we will report two sets of p-values: unadjusted p-values as used by the authors, and p-values that correct for multiple hypothesis outcomes as used by List et al. (2021).

7.2.1 Authors' Own Definitions

We will begin by re-running all analyses without control variables, as this is not reported in the paper.

Then, we will re-run all analyses using imputed district-gender mean, replacing it with school-gender mean.

Finally, we will use non-parametric kernel regression to analyse heterogeneity of the treatment effects. Within the survey data are background variables which plausibly lead to heterogeneous treatment effects. We will primarily be interested in parental attitudes to gender roles.

7.2.2 Alternative Definitions

We will re-run all analyses that use indexes, using factor analysis as an alternative method of aggregating survey items.

Additionally, survey respondents' answers to some questions were originally recorded on an ordinal scale of several values, but are "dichotomised" in the analysis. We will use the original values where possible to see if this materially affects results.