Regression and controls

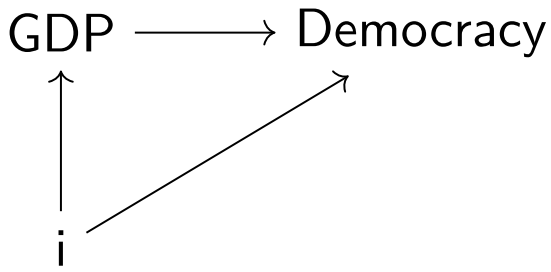
Felipe Balcazar

NYU

May , 2023

NEW YORK UNIVERSITY

Example of a regression



$$Dem_{it} = \alpha + \beta GDP_{it} + \mu_i + \varepsilon_{it}.$$

Interaction effects: regression form

$$Dem_{it} = \alpha + \beta_1 GDP_{it} + \beta_2 Oil_{it} + \beta_3 GDP_{it} \times OIL_{it} + \delta X_{it} + \mu_i + \gamma_t + \varepsilon_{it}.$$

- β_1 is the effect of the treatment conditional on $Oil = 0$:

  $$E(Dem_{it}|GDP_{it} = 1, Oil_{it} = 0, X) - E(Dem_{it}|GDP_{it} = 0, Oil_{it} = 0, X)$$

- $\beta_1 + \beta_3$ is the effect of the treatment conditional on $Oi; = 1$:

  $$E(Dem_{it}|GDP_{it} = 1, Oil_{it} = 1, X) - E(Dem_{it}|GDP_{it} = 0, Oil_{it} = 1, X)$$

- β_3 is the additional effect of the treatment for $Oil = 1$.

Interaction effects: regression form

$$Dem_{it} = \alpha + \beta_1 GDP_{it} + \beta_2 Oil_{it} + \beta_3 GDP_{it} \times OIL_{it} + \delta X_{it} + \mu_i + \gamma_t + \varepsilon_{it}.$$

- β_1 is the effect of the treatment conditional on $Oil = 0$:

  $$E(Dem_{it}|GDP_{it} = 1, Oil_{it} = 0, X) - E(Dem_{it}|GDP_{it} = 0, Oil_{it} = 0, X)$$

- $\beta_1 + \beta_3$ is the effect of the treatment conditional on $Oi; = 1$:

  $$E(Dem_{it}|GDP_{it} = 1, Oil_{it} = 1, X) - E(Dem_{it}|GDP_{it} = 0, Oil_{it} = 1, X)$$

- β_3 is the additional effect of the treatment for $Oil = 1$.

Interaction effects: regression form

$$Dem_{it} = \alpha + \beta_1 GDP_{it} + \beta_2 Oil_{it} + \beta_3 GDP_{it} \times OIL_{it} + \delta X_{it} + \mu_i + \gamma_t + \varepsilon_{it}.$$

- β_1 is the effect of the treatment conditional on $Oil = 0$:

  $$E(Dem_{it}|GDP_{it} = 1, Oil_{it} = 0, X) - E(Dem_{it}|GDP_{it} = 0, Oil_{it} = 0, X)$$

- $\beta_1 + \beta_3$ is the effect of the treatment conditional on $Oi; = 1$:

  $$E(Dem_{it}|GDP_{it} = 1, Oil_{it} = 1, X) - E(Dem_{it}|GDP_{it} = 0, Oil_{it} = 1, X)$$

- β_3 is the additional effect of the treatment for $Oil = 1$.

Interaction effects: regression form

$$Dem_{it} = \alpha + \beta_1 GDP_{it} + \beta_2 Oil_{it} + \beta_3 GDP_{it} \times OIL_{it} + \delta X_{it} + \mu_i + \gamma_t + \varepsilon_{it}.$$

- β_1 is the effect of the treatment conditional on $Oil = 0$:

  $$E(Dem_{it}|GDP_{it} = 1, Oil_{it} = 0, X) - E(Dem_{it}|GDP_{it} = 0, Oil_{it} = 0, X)$$

- $\beta_1 + \beta_3$ is the effect of the treatment conditional on $Oi; = 1$:

  $$E(Dem_{it}|GDP_{it} = 1, Oil_{it} = 1, X) - E(Dem_{it}|GDP_{it} = 0, Oil_{it} = 1, X)$$

- β_3 is the additional effect of the treatment for $Oil = 1$.

Why tables and graphs?

- To communicate ideas and your results more clearly.
- Researchers use these tools to do the same.
  - Good tables and graphs are powerful tools!
  - A scatter plot to show correlations.
  - A descriptive statics table to communicate the structure of your data.
  - A balance table to show random assignment.
  - A regression table to showcase your results in the paper.
  - A confidence interval graph to illustrate the previous results.
- Importantly, they must be self-contained.
  - Title and axis should be clear.
  - Labels should be descriptive but not long.
  - Footnotes should be explanatory.

Why tables and graphs?

- To communicate ideas and your results more clearly.
- Researchers use these tools to do the same.
  - Good tables and graphs are powerful tools!
  - A scatter plot to show correlations.
  - A descriptive statics table to communicate the structure of your data.
  - A balance table to show random assignment.
  - A regression table to showcase your results in the paper.
  - A confidence interval graph to illustrate the previous results.
- Importantly, they must be self-contained.
  - Title and axis should be clear.
  - Labels should be descriptive but not long.
  - Footnotes should be explanatory.

Why tables and graphs?

- To communicate ideas and your results more clearly.
- Researchers use these tools to do the same.
  - Good tables and graphs are powerful tools!
  - A scatter plot to show correlations.
  - A descriptive statics table to communicate the structure of your data.
  - A balance table to show random assignment.
  - A regression table to showcase your results in the paper.
  - A confidence interval graph to illustrate the previous results.
- Importantly, they must be self-contained.
  - Title and axis should be clear.
  - Labels should be descriptive but not long.
  - Footnotes should be explanatory.

Summary statistics table: Olken (2007)

SUMMARY STATISTICS

	Summary Statistics
Total project size (US$)	8,875 (4,401)
Share of total reported expenses:	
Road project	.766 (.230)
Ancillary projects (culverts, retaining walls, etc.)	.154 (.181)
Other projects (schools, bridges, irrigation, etc.)	.079 (.166)
Share of reported road expenses:	
Sand	.099 (.080)
Rocks	.484 (.143)
Gravel	.116 (.181)
Unskilled labor	.196 (.125)
Other	.105 (.164)
Percent missing:	
Major items in road project	.237 (.343)
Major items in roads and ancillary projects	.247 (.350)
Materials in road project	.203 (.395)
Unskilled labor in road project	.273 (.851)
Observations	538

NOTE.—Statistics shown are means, with standard deviations in parentheses. Data on expenditures are taken from the 538 villages for which percent missing in road and ancillary projects could be calculated. Exchange rate is Rp. 9,000 = US$1.00.

Summary statistics: Aguero et al. (2020)

Table 1
Descriptive statistics. Averages 2007–2012.

	Producing districts	Non-producing districts in producing provinces	Non-producing districts in non-producing provinces
A. Test scores (*Evaluación Censal de Estudiantes*)			
Average score in mathematics	517.06 (103.26) [53, 944]	534.04 (107.72) [53, 944]	509.18 (105.95) [53, 944]
Average score in reading	514.73 (85.7) [62, 814]	532.49 (91.19) [49, 814]	508.47 (91.17) [49, 814]
Number of students	*110,885*	*411,275*	*1,565,264*
B. Schools' characteristics (*Censo Escolar*), percentage			
Teachers with long-term contract	28.56 (36.29) [0, 100]	31.22 (40.02) [0, 100]	27.79 (38.48) [0, 100]
Number of schools	*5,811*	*24,534*	*83,330*
C. Districts' characteristics			
Mining production per-capita[a]	62 (224.99) [0, 2494.75]	0.00 (0) [0, 0]	0.00 (0) [0, 0]
Canon transfers per-capita[a]	0.37 (0.96) [0, 9.71]	0.12 (0.39) [0, 10.15]	0.04 (0.1) [0, 1.36]
Number of districts	128	554	1156

Note: standard deviation in parenthesis. Minimum and maximum values in brackets. Population figures for each district and year are obtained from Peru's National Bureau of Statistics (INEI).

Balance table: MHE (Ch 1.)

Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
			A. Health			
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
			B. Characteristics			
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

Results tables: MHE (Ch 1.)

OHP effects on insurance coverage and health-care use

Outcome	Oregon: Control mean (1)	Oregon: Treatment effect (2)	Portland area: Control mean (3)	Portland area: Treatment effect (4)
A. Administrative data				
Ever on Medicaid	.141	.256 (.004)	.151	.247 (.006)
Any hospital admissions	.067	.005 (.002)		
Any emergency department visit			.345	.017 (.006)
Number of emergency department visits			1.02	.101 (.029)
Sample size	74,922		24,646	
B. Survey data				
Outpatient visits (in the past 6 months)	1.91	.314 (.054)		
Any prescriptions?	.637	.025 (.008)		
Sample size	23,741			

Notes: This table reports estimates of the effect of winning the Oregon Health Plan (OHP) lottery on insurance coverage and use of health care. Odd-numbered columns show control group averages. Even-numbered columns report the regression coefficient on a dummy for lottery winners. Standard errors are reported in parentheses.

Example of stars in papers

	Dependent variable: Average years of education attained			
	(1)	(2)	(3)	(4)
(a) Full sample				
Stock of democracy	0.004	0.005	0.006*	0.009**
	(0.003)	(0.003)	(0.004)	(0.004)
Constant	9.959***	9.048***	8.974***	8.893***
	(0.476)	(0.521)	(0.513)	(0.502)
Discount factor (r)	0.01	0.03	0.06	0.10
R^2	0.435	0.437	0.441	0.446
Clusters	210	210	210	210
Observations	3078	3078	3078	3078
(b) Restricted sample				
Stock of democracy	−0.007	−0.007	−0.006	−0.005
	(0.005)	(0.005)	(0.006)	(0.007)
Constant	10.998***	10.908***	10.767***	10.584***
	(0.481)	(0.495)	(0.510)	(0.518)
Discount factor (r)	0.01	0.03	0.06	0.10
R^2	0.421	0.417	0.411	0.405
Clusters	168	168	168	168
Observations	2714	2714	2714	2714

Note: Standard errors are clustered by country and birth-cohort in parentheses. * Significant at 10 %, ** significant at 5 %, *** significant at 1 %. The stock of democracy is calculated from 6 to 18 years after a

Example of stars in papers (II)

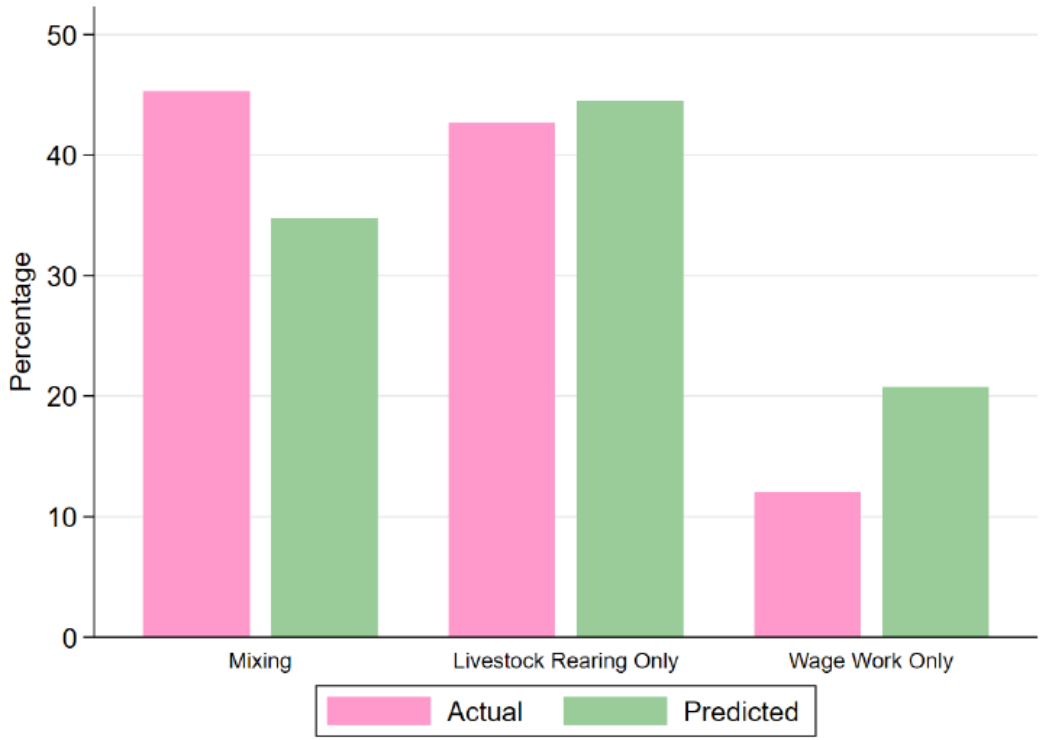
Table 5
Mechanisms: Role of health factors.

	Dependent variable				
	Individual experienced health complications in the past 4 weeks		Individual was sick in the past 4 weeks		Number of days individual couldn't work due to sickness in the past 4 weeks
	(All individuals)	(6–10 years of age)	(All individuals)	(6–10 years of age)	(14–65 years of age)
	(1)	(2)	(3)	(4)	(5)
Canon	−0.0956***	−0.0637	−0.0621**	−0.0776*	0.0067
	(0.0281)	(0.0543)	(0.0291)	(0.0468)	(0.0522)
***Canon* squared**	0.0096***	0.0105**	0.0050**	0.0077*	−0.0026
	(0.0023)	(0.0044)	(0.0023)	(0.0040)	(0.0041)
Observable characteristics	No	No	No	No	No
R-squared	0.0693	0.0558	0.0347	0.0568	0.0137
Number of observations	450,615	49,606	450,615	49,606	370,499

Note: Robust standard errors clustered at the district level in parentheses. * Significant at ten percent; ** significant at five percent; *** significant at one percent. All regressions include the value of mining production (constant USD, 2010 = 100), dummies for type of province (i.e., producer and non-producer in producing district), and fixed effects at the district level and by year. *Canon* corresponds to the value of *Canon* per-capita, in thousands of USD at constant prices of 2010.
Source: Authors' calculations based on ENAHO household survey and data from Peru's Ministry of Finance and Peru's Ministry of Mines and Energy.

Bar graph: Balboni et al. (2020)

Figure 11: Predicted vs. actual occupation in Year 4



Notes: The pink bars show the observed distribution across occupations (specialization in livestock rearing, specialization in wage labor, engaging in both occupations) in year 4 for those of the 64% of ultra-poor individuals for whom individual-level parameters can be calibrated using baseline and/or year 2 data (as described in the text) who report positive labor hours at year 4. The green bars show, for the same individuals, model-implied optimal occupational choices at each individual's observed year 4 capital level.

Line graph: Royer et al. (2015)

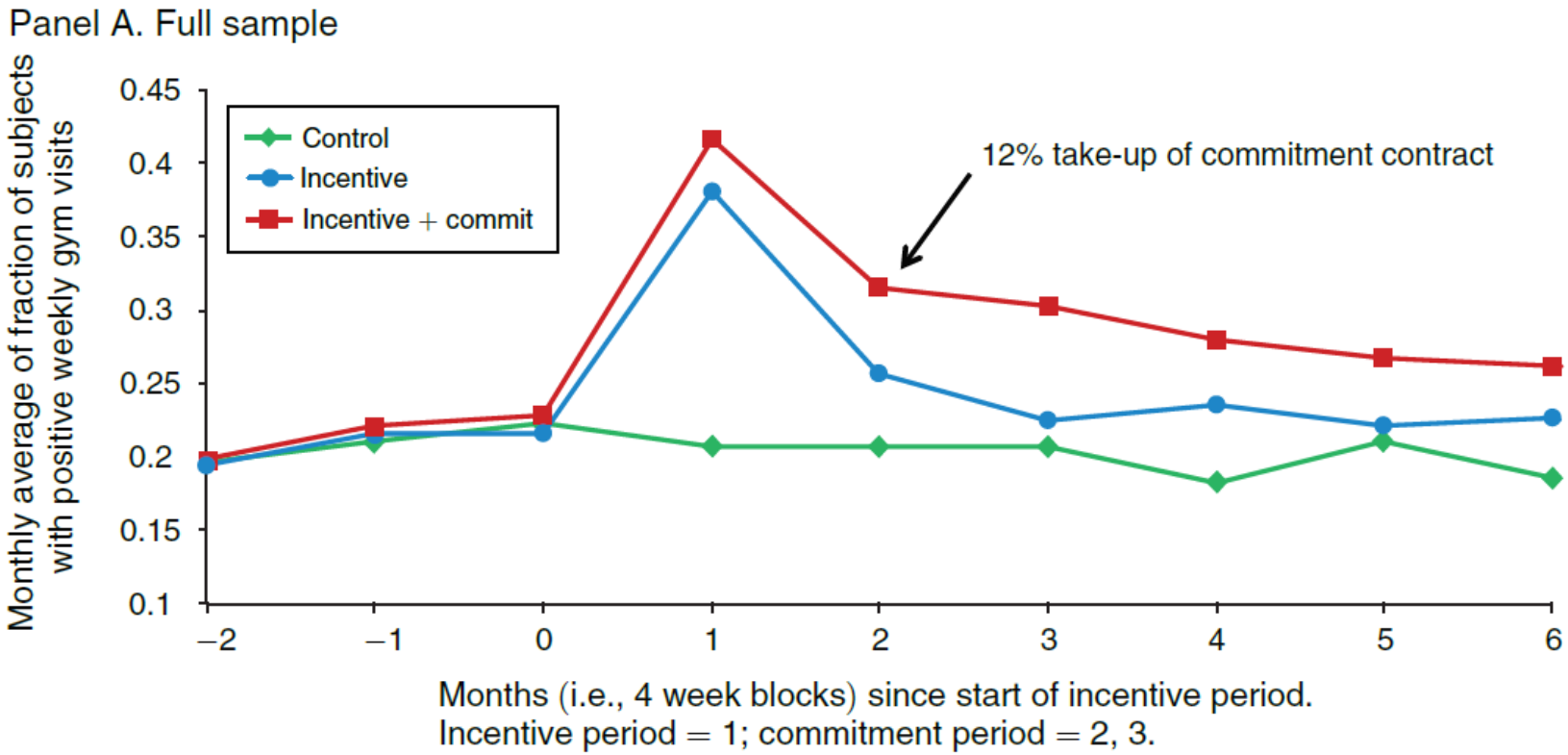


FIGURE 1. FRACTION WITH POSITIVE GYM VISITS BY TREATMENT

Results table: Bidwell et al. (202))

WINNING MPS: IMPACTS OF DEBATE PARTICIPATION ON POLICY

	Control Mean (1)	Treatment Effect (2)	Standard Error (3)	Naive p (One-Sided) (4)	N (5)
	A. Hypothesis-Level Policy Effect				
Mean effects index (nine outcomes)	.000	.298**	.159	.037	28
	B. Estimates for Individual Policy Outcomes				
Development spending verified in field (percent of 2012 CFF)	35.560	54.738**	31.707	.050	27
Total constituency visits	2.915	1.316**	.619	.022	28
Total public meetings held with constituents	1.018	1.089**	.606	.043	28
Percent of 2012–13 sittings attended (out of 57 total)	76.692	3.371	3.003	.137	28
Total public comments in parliamentary sittings, 2012–13	4.286	−1.569	2.224	.878	28
Total committee membership	3.929	.524	.625	.206	28
Total public comments in priority sector agenda items	.154	−.170	.166	.842	27
Membership in priority sector committee	.231	.201	.187	.147	27
Constituent assessment of focus on priority sector	.571	−.343	.150	.984	27

NOTE.—This table leverages the constituency-level randomization to estimate the effects of participating in a debate as a candidate on the subsequent performance of the elected MP in office. Significance levels are based on one-sided tests in the direction prespecified in the preanalysis plan in col. 4. Hypothesis-level mean effects indices are constructed following Kling, Liebman, and Katz (2007) and expressed in standard deviation units, with missing values for component measures imputed at random assignment group means. Estimates for individual outcomes are expressed in units natural to the measure. The standard error presented is the maximum value of conventional ordinary least squares and bias-corrected HC2 estimators from MacKinnon and White (1985), following discussion by Angrist and Pischke (2009). Specifications include gender, previous elected office experience, and stratification bins for the constituency (three bins of ethnic-party bias). Missing values for priority sector outcomes are from one control MP who did not provide a preelection priority, and missing values for development spending are from one treated MP who did not take office until December 2013 (1 year after the election) and thus did not receive the 2012 CFF.

** $p < .05$.

Results table: Beath et al. (2017)

Table 5: Effect on Security as Function of Distance to Pakistan

	Occurrence of at Least One Security Incident	
	Regions Bordering Pakistan	
	(1)	(2)
Treatment Effect at Midline	1.042	0.679
	(0.356)***	(0.316)**
Treatment Effect at Endline	0.902	0.539
	(0.263)***	(0.306)*
Distance to Pakistan × Treatment Effect at Midline	-29.056	-20.076
	(10.152)***	(9.344)**
Distance to Pakistan × Treatment Effect at Endline	-24.559	-15.579
	(7.761)***	(8.991)*
Distance to Pakistan	-18.827	-17.842
	(10.800)*	(7.365)**
Dependent Variable at Baseline		0.389
		(0.102)***
Matched pair-survey fixed effects	Yes	Yes
Observations	200	200
R-squared	0.867	0.896

Note: Dependent variables are unweighted average or measures for different radii (Kling, Leibman, and Katz, 2007). Midline referes to the period from the start of the program in October 2010 until the completion of the Midline survey in September 2009; Endline refers to the period from the completion of the midline survey until the completion of the endline survey in September 2011. Distance is measured in thousands of kilometers. Robust standard errors adjusted for clustering at the village-cluster level in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%.

Confidence interval graph: Mousa (2020)

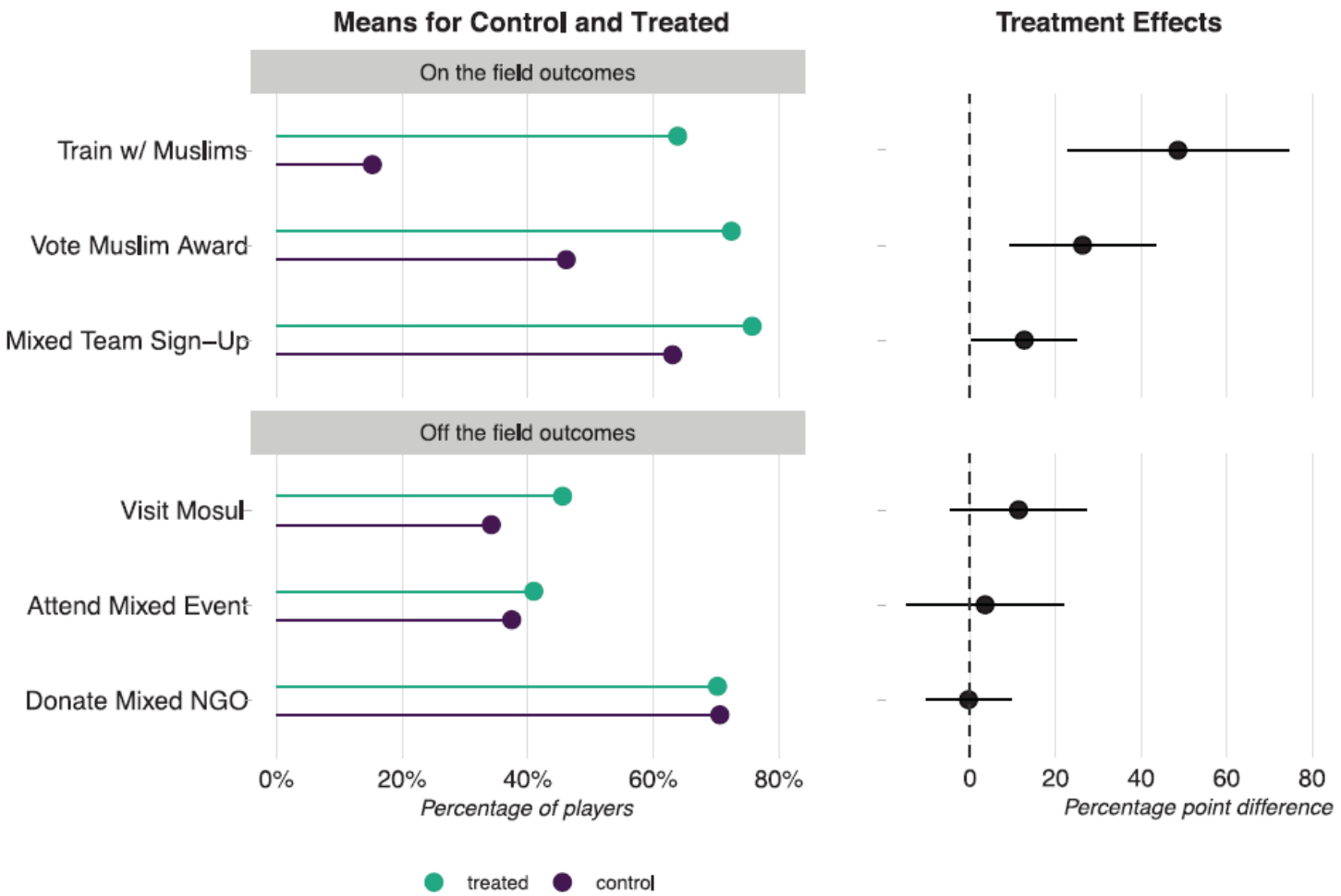
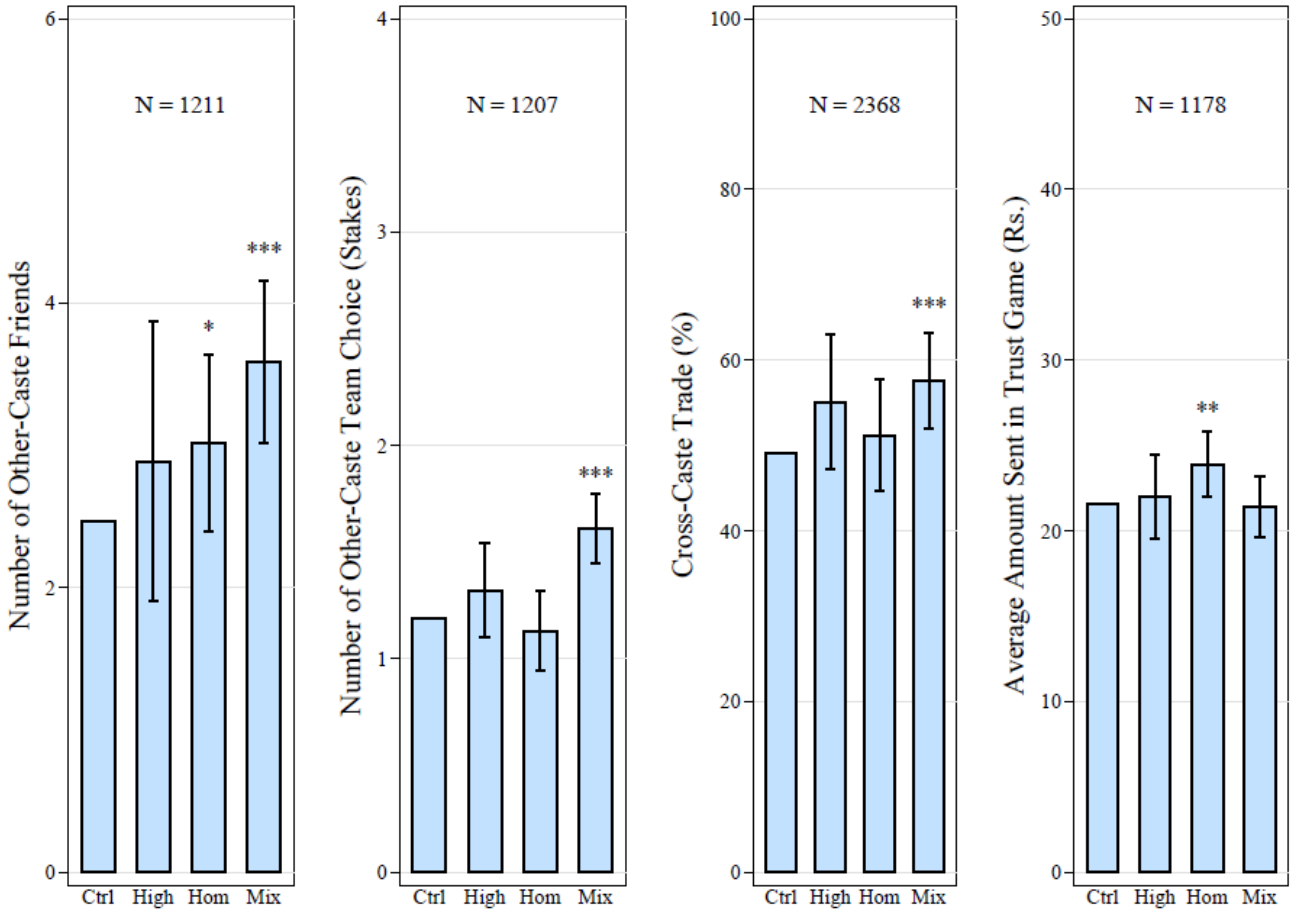


Fig. 1. Behavioral results. The intervention consistently improved on-the-field behavioral outcomes, with no detectable effects on off-the-field outcomes. The left panel shows covariate-adjusted mean outcomes for treated and control players, with covariates held at median or modal values. The right panel shows the difference between treated and control players, with 95% confidence intervals.

Confidence interval graph: Lowe et al. (2020)

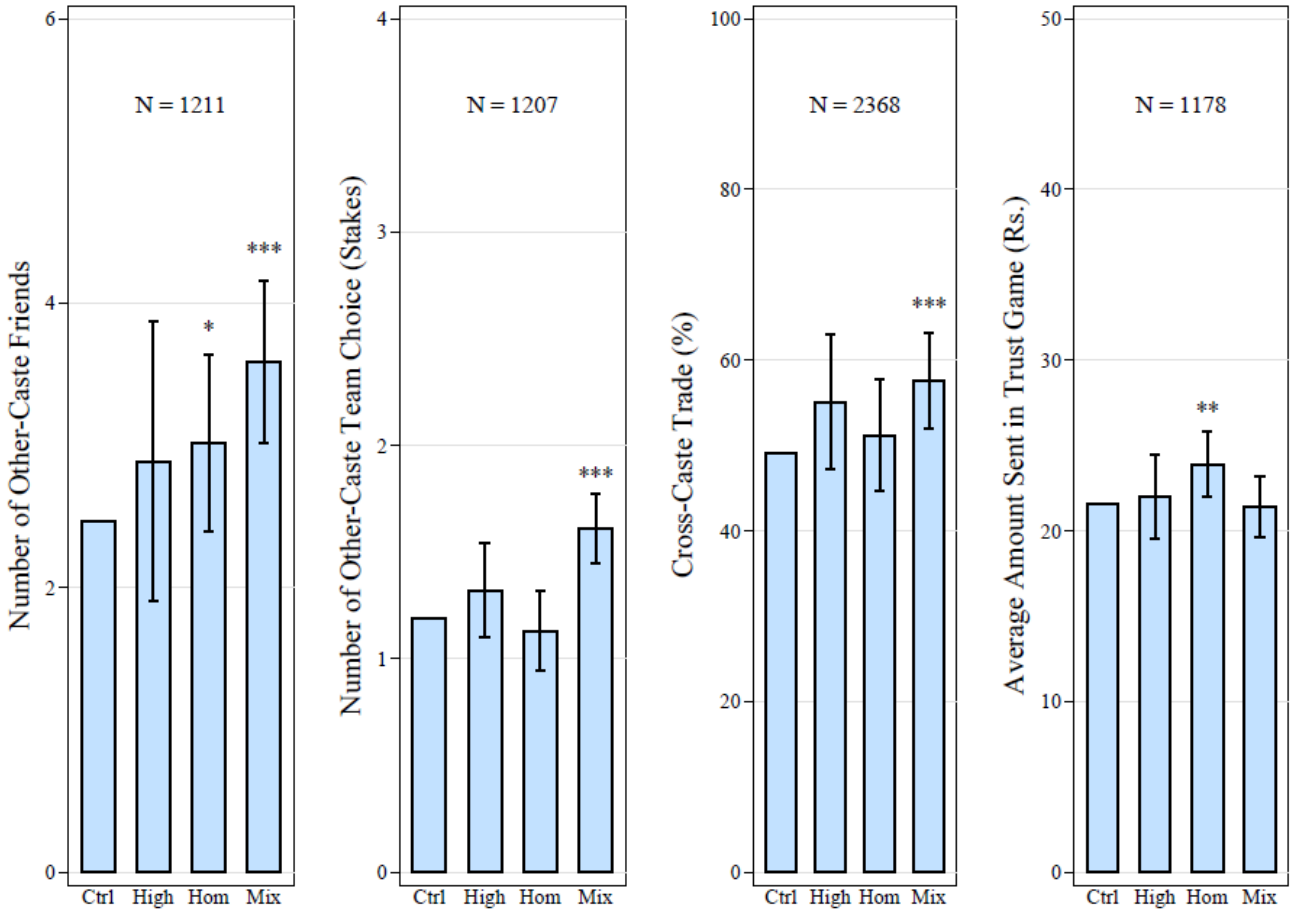
Figure 7: League Participation Reduces Intergroup Differences



Notes: The figure shows treatment effects and significance levels of Homog. Team (Hom), High Backup (High), and Mixed Team (Mix) relative to the low-priority backups (Ctrl), drawing on estimates from equation 2. From left-to-right the outcomes are: (1) number of other-caste men participant considers friends, (2) number of other-caste men chosen as teammates for future match with stakes, (3) percentage of cross-caste trade, and (4) average amount sent in the trust game to the three Recipients. For the cross-caste trade outcome, the regression additionally includes the Trade and Color-Switch Bonus dummy variables. *** p<0.01, ** p<0.05, * p<0.1.

Margins plot

Figure 7: League Participation Reduces Intergroup Differences



Notes: The figure shows treatment effects and significance levels of Homog. Team (Hom), High Backup (High), and Mixed Team (Mix) relative to the low-priority backups (Ctrl), drawing on estimates from equation 2. From left-to-right the outcomes are: (1) number of other-caste men participant considers friends, (2) number of other-caste men chosen as teammates for future match with stakes, (3) percentage of cross-caste trade, and (4) average amount sent in the trust game to the three Recipients. For the cross-caste trade outcome, the regression additionally includes the Trade and Color-Switch Bonus dummy variables. *** p<0.01, ** p<0.05, * p<0.1.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. A
there was some unprocessed data that should have been added
the final page this extra page has been added to receive it.
If you rerun the document (without altering it) this surplus page
will go away, because LaTeX now knows how many pages to expe
for this document.