APPENDIX:
Estimating the effect of intergroup contact over years: Evidence from a youth program in Israel

Contents

1 Descriptive Statistics

	Treatment	Control	
Treatment	60	11	
Control	9	58	
Missing	12	4	
	Jewish-Israeli	*Arab-Palestinian*	Std. mean difference
Demographics			
N	95	43	
Age	10.68	10.51	0.219
	(0.08)	(0.16)	
Female (N, %)	28	43	0.030
	(29.47%)	(100%)	
Family religiosity	2.21	1.53	0.055
	(0.04)	(0.08)	
religious (N, %)	1 (1.05%)	20 (46.51%)	0.016
traditional (N, %)	73 (76.84%)	23 (53.49%)	0.087
secular (N, %)	21 (22.11%)	0 (0%)	0.096

Table 1: Descriptive statistics: RCT sample

Covariate	*Location A*	*Location B*	*Location* C
N	43	28	67
Age	10.51	10.50	10.76
	(0.16)	(0.18)	(0.09)
Female (N, %)	43	28	0
	(100%)	(100%)	(0%)
Ethnicity (N, %)	1.00	0	0
	(100%)	(0%)	(0%)
Family religiosity	1.53	2.04	2.28
	(0.08)	(0.06)	(0.06)
religious (N, %)	20 (46.51%)	1 (3.57%)	0 (0%)
traditional (N, %)	23 (53.49%)	25 (89.29%)	48 (71.64%)
secular (N, %)	0 (0%)	2 (7.14%)	19 (28.36%)

Table 2: Descriptive statistics: RCT sample across the three locations

	Jewish-Israeli	*Arab-Palestinian*
N	299	342
Participants per year		
2015	51	88
2016	70	65
2017	87	79
2018	59	82
2019	32	28

	Jewish-Israeli	*Arab-Palestinian*
Demographics		
Age	11.93	11.53
	(0.12)	(0.10)
Female	229	225
	(76.59%)	(65.79%)
Years in program	1.973	2.982
	(0.076)	(0.108)
Family religiosity	2.124	1.944
	(0.029)	(0.017)
religious (N, %)	21 (7.02%)	24 (7.02%)
traditional (N, %)	220 (73.58%)	315 (92.11%)
secular (N, %)	58 (19.40%)	2 (0.58%)

Table 3: Descriptive statistics: Survey data

2 Research Design

We describe the research design in detail within the main text (*Research Design* section). In terms of data collection, participants had separate data collection sessions that were facilitated by Arab or Jewish coaches or other staff, in Arabic or Hebrew language respectively. The PIs attended several data collection sessions each year, and designed age appropriate survey instruments for self-administration with minimal facilitation by our field staff, which allowed for privacy and corresponds with current guidance on measuring sensitive topics (Tourangeau and Yan, 2007). In addition to outcome variables, the endline survey collected covariate information on parents' occupations, religious affiliations, and basic information about the family's immigration history. Coaches distributed these surveys to their players at the end of the season, during roughly the same time as the RCT data collection.

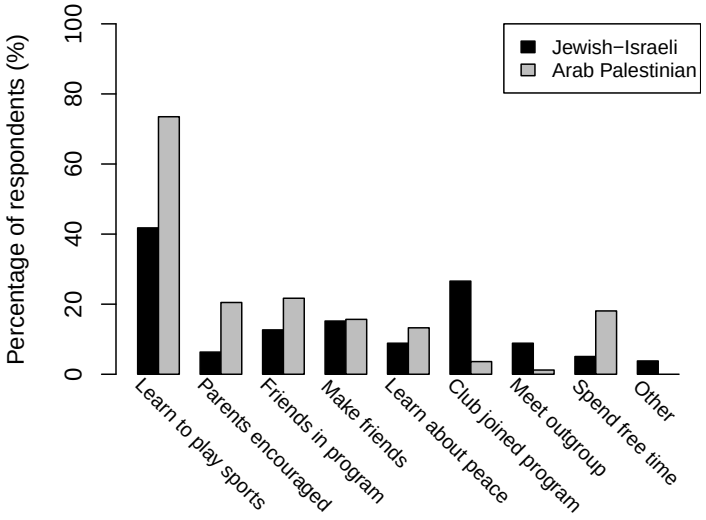

Figure 1: Main reasons for joining the program, as reported by participants in the 2017 survey. Participants were instructed to indicate all of the categories that apply to them.

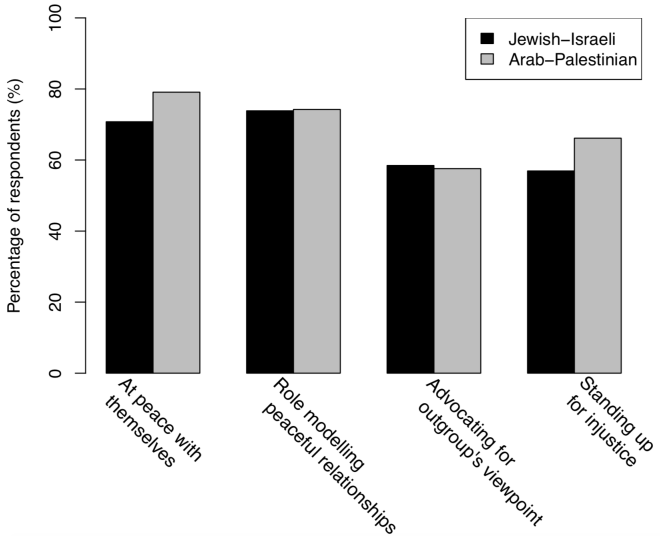

Figure 2: Responses to the question of best strategies that participants can engage in to promote peace (2016 survey, 136 participants).

3 Outcomes

Concept	*Indicator*	*References*
Outgroup Regard	Social distance	Bogardus, E. S. (1933). A social distance scale. Sociology and Social Research, 17, 265–271. Parrillo, V. N., & Donoghue, C. (2005). Updating the Bogardus social distance studies: A new national survey. The Social Science Journal, 42,257–271. http://dx.doi.org/10.1016/j.soscij.2005.03.011
Outgroup Regard	Support for peace process	Child-friendly version of the monthly Peace Index conducted by the Guttman Center for Public Opinion and Policy Research and the Israel Democracy Institute (https://en.idi.org.il/centers/1159/1520)
Outgroup Regard	Perspective taking	Developed based on Bruneau, E. Saxe, R. (2012). The power of being heard: The benefits of 'perspective giving' in the context of intergroup conflict. Journal of Experimental Psychology, 48(4), 855-866.
Outgroup Regard	Hostile attribution by subject	McGothlin, H, and Killen, M. (2010). How social experience is related to children's intergroup attitudes. Journal of Social Psychology: Special Issue: Children's Intergroup Attitudes, 40, 625-634. McGothlin, H, and Killen, M. (2006). Intergroup attitudes of European American children attending ethnically homogenous schools. Child Development, 77, 1375.1386.
Outgroup Regard	Hostile attribution by peers	Expansion of McGothlin and Killen (2006, 2010)
Outgroup Regard	Optimism about peace	Child-friendly version of the monthly Peace Index conducted by the Guttman Center for Public Opinion and Policy Research and the Israel Democracy Institute (https://en.idi.org.il/centers/1159/1520)
Outgroup Regard	Ingroup identity esteem	Luhtanen, R., & Crocker, J. (1992). A collective self-esteem scale: Self evaluation of one's social identity. Personality and Social Psychology Bulletin, 18, 302–318. http://dx.doi.org/10.1177/0146167292183006
Ingroup Regulation	Effort to persuade	Expansion of McGothlin & Killen (2006, 2010)
Ingroup Regulation	Ingroup censuring and policing	Huesmann, L. R., & Guerra, N. G. (1997). Children's normative beliefs about aggression and aggressive behavior. Journal of Personality and Social Psychology, 72, 408–419. Möller, I., & Krahé, B. (2009). Exposure to violent video games and aggression in German adolescents: A longitudinal analysis. Aggressive Behavior, 35, 75–89.
Ingroup Regulation	Willing to share outgroup perspective	Created based on qualitative interviews

Table 4: Outcome measures and associated references

4 Cross-Sectional Survey

A Regression Specifications

We analyzed survey data with the following specifications (all reported below):

(1) $Y_{ics} = \alpha + \beta_1$ Years in prog.$+ \epsilon_{ics}$

(2) $Y_{ics} = \alpha + \beta_1$ Years in prog. $+ \beta_2$ Age $+ \epsilon_{ics}$

(3) $Y_{ics} = \alpha + \beta_1$ Years in prog. $+ \beta_2$ Age $+ \beta_3$ Years in prog. X Arab $+ \epsilon_{ics}$

(4) $Y_{ics} = \alpha + \beta_1$ Years in prog. X Age $+ \beta_3$ Years in prog. X Arab $+\mu_c + \kappa_s + \epsilon_{ics}$

(5) $Y_{ics} = \alpha + \beta_1$ Years in prog. X Age $+ \beta_3$ Years in prog. X Arab $+\mu_c + \kappa_s + \tau_i + \epsilon_{ics}$,

where the parameter μ_c is a year-specific fixed effect, κ_s is a site-specific fixed effect and τ_i is an individual fixed effect; we control for these fixed effects using year dummy, site dummy and individual code dummy variables.

B Results

Variable	(1)	(2)	(3)	(4)	(5)
(Intercept)	16.93	13.17	14.42	12.73	
(SE)	(0.36)	(0.76)	(0.9)	(1.34)	
[p]	[0]	[0]	[0]	[0]	
Years in prog.	0.58	0.56	0.57	0.53	2.54
(SE)	(0.11)	(0.36)	(0.38)	(0.37)	(1.14)
[p]	[0]	[0.12]	[0.14]	[0.16]	[0.04]
Age - 8		1.23	1.07	0.54	
(SE)		(0.38)	(0.41)	(0.42)	
[p]		[0]	[0.01]	[0.2]	
Age started - 8		0.16	0.17	0.6	
(SE)		(0.44)	(0.47)	(0.49)	
[p]		[0.72]	[0.72]	[0.23]	
Years in prog. X Age started		-0.13	-0.13	-0.13	-0.52
(SE)		(0.06)	(0.06)	(0.06)	(0.22)
[p]		[0.04]	[0.05]	[0.04]	[0.03]
Arab			-2.03	-2.03	
(SE)			(0.73)	(0.74)	
[p]			[0.01]	[0.01]	
Years in prog. X Arab			0.23	0.29	0.03
(SE)			(0.24)	(0.24)	(0.4)
[p]			[0.36]	[0.22]	[0.94]
N	601	601	601	601	

Descriptive regressions for Social Distance. Column 4 includes fixed effects for year and location. Column 5 includes individual-level fixed effects. 63 respondents have between 2 to 4 repeated values in the data, while 503 respondents appear only once.

Table 5: (a) Social Distance

Variable	(1)	(2)	(3)	(4)	(5)
(Intercept)	2.57	1.76	2.01	1.64	
(SE)	(0.1)	(0.19)	(0.22)	(0.27)	
[p]	[0]	[0]	[0]	[0]	
Years in prog.	0.09	0.03	0.28	0.25	1.07
(SE)	(0.03)	(0.08)	(0.11)	(0.11)	(0.38)
[p]	[0.01]	[0.7]	[0.02]	[0.03]	[0.01]
Age - 8		0.56	0.39	0.33	
(SE)		(0.13)	(0.18)	(0.18)	
[p]		[0]	[0.03]	[0.07]	
Age started - 8		-0.25	-0.17	-0.16	
(SE)		(0.14)	(0.17)	(0.18)	
[p]		[0.07]	[0.34]	[0.36]	
Years in prog. X Age started		-0.02	-0.02	-0.02	-0.19
(SE)		(0.02)	(0.02)	(0.02)	(0.07)
[p]		[0.15]	[0.18]	[0.21]	[0.01]
Arab			-0.46	-0.56	
(SE)			(0.17)	(0.18)	
[p]			[0.01]	[0]	
Years in prog. X Arab			-0.19	-0.16	-0.07
(SE)			(0.09)	(0.08)	(0.31)
[p]			[0.03]	[0.06]	[0.84]
N	601	601	601	601	

Descriptive regressions for Perspective Sharing. Column 4 includes fixed effects for year and location. Column 5 includes individual-level fixed effects. 63 respondents have between 2 to 4 repeated values in the data, while 503 respondents appear only once.

Table 6: (b) (Narrative) Perspective Sharing

5 RCT Main & Moderation Effects

A ITT Specification

We estimate ITT effects based on the following regression (Model 1):

$$Y_{ics} = \alpha^{ITT} + \beta^{ITT} Z_{ics} + X'_{ics}\theta^{ITT} + \mu_c^{ITT} + \kappa_s^{ITT} + \epsilon_{ics}^{ITT},$$

where Z_{ics} takes the value 1 if individual i with outcome measured in year c at site s was assigned to the treatment and 0 if not, X'_{ics} is a vector of covariates for this individual, and Y_{ics} is an outcome. The parameter μ_c^{ITT} is a year-specific fixed effect and the κ_s^{ITT} is a site-specific fixed effect; we control for these fixed effects using year dummy variables and site dummy variables. The estimate of the coefficient on Z_{ics}, which we label $\hat{\beta}^{ITT}$, measures the average effect of the treatment. Covariates (X'_{ics}) include year and participant's location of origin, gender, age and ethnicity (the latter gets absorbed by the location indicators, since locations are specific to Jewish-Israeli vs. Arab-Palestinian subjects). Covariate specification follows Lin (2013) in including mean-centered covariates as well as their interactions with the treatment assignment indicator.

B Principal Component Analysis & Standardized Effects

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Social distance	0.947	-0.235	0.137	-0.071	0.037	0.059	0.139
Support for peace	0.152	-0.034	-0.046	0.008	0.038	-0.217	-0.962
Perspective taking	0.248	0.961	-0.097	-0.043	-0.036	0.045	-0.002
Hostile attribution (sub.)	0.024	0.068	0.164	0.765	0.618	-0.003	0.024
Hostile attribution (peer)	0.114	-0.120	-0.957	0.221	-0.013	0.065	0.055
Optimism about peace	0.055	0.028	-0.011	0.151	-0.202	-0.942	0.214
Ingroup identity esteem	0.044	-0.008	0.162	0.579	-0.757	0.237	-0.079

Table 7: (a) PC scores for all the items that form the composite index of Outgroup Regard

	PC1	PC2	PC3	PC4
Effort to persuade	0.018	0.000	0.065	-0.998
Ingroup censuring	0.990	0.140	-0.030	0.016
Ingroup policing	0.126	-0.948	-0.292	-0.017
Share outgroup perspective	0.068	-0.285	0.954	0.063

Table 8: (b) PC scores for all the items that form the composite index of Ingroup Regulation

	Outgroup Regard	Ingroup Regulation	Personal Resources
Pooled Sample	-0.078	-0.012	0.116
	(0.141)	(0.175)	(0.100)
Arab-Palestinian	-0.203	-0.268	0.175
	(0.277)	(0.350)	(0.159)
Jewish-Israeli	-0.003	0.027	0.123
	(0.248)	(0.208)	(0.170)

Table 9: Standardized effects for the composite indices. Standardization follows the procedure in Paluck et al. (2019) to ensure comparability.

	Social Distance	Perspective Sharing	Global Self-Esteem
Pooled Sample	0.066	0.139	0.063
	(0.144)	(0.142)	(0.089)
Arab-Palestinian	0.001	-0.283	-0.008
	(0.387)	(0.379)	(0.189)
Jewish-Israeli	0.105	0.419	0.059
	(0.264)	(0.218)	(0.138)

Table 10: Standardized effects for the main indicators

C Moderation by Ethnicity

We examine the extent to which effects vary by ethnicity, in line with previous research (Ditlmann and Samii, 2016). To do so, we modify our ITT specification (Model 2):

$$Y_{ics} = \alpha^{MOD} + \beta_1^{MOD} Z_{ics} + \beta_2^{MOD} Z_{ics} \mathrm{I(Arab)}_{ics} + X'_{ics}\theta^{MOD} + \mu_c^{MOD} + \kappa_s^{MOD} + \epsilon_{ics}^{MOD},$$

where β_1^{MOD} is the effect of the program on Jewish-Israeli participants, and β_2^{MOD} is how Arab-Palestinian participants differ from Jewish-Israeli participants in their effects.

Hypothesis H1b stipulates that the one-year program impact on outgroup regard should be stronger for Israeli-Jewish than for Arab-Palestinian participants. We present the results from the ITT analysis that tests for heterogeneity by ethnic membership. The effect for Arab-Palestinian participants is the sum of the program and the interaction term coefficients: for the overall Outgroup Regard index it is -0.58 SD (p=0.37), and for the social distance indicator it is 0.15 SD (p=0.94). With both interaction coefficients insignificant we cannot reject the null of homogeneity, but across all point estimates we see that program is more effective for Jewish-Israelis than Arab-Palestinian participants – a point we return to below.

	Outgroup Regard	Social Distance	Supp. Peace	Persp. Tk.	Host. Att. Self	Host Att. Peer	Opti. Peace	Ingroup Ident.
Program effect (Jewish-Israeli)	0.13	0.43	0.11	0.00	-0.50	0.44	0.10	-0.44
(SE)	(0.27)	(0.89)	(0.16)	(0.21)	(0.26)	(0.30)	(0.18)	(0.31)
[p]	[0.64]	[0.63]	[0.49]	[0.99]	[0.06]	[0.15]	[0.60]	[0.16]
[p FDR]	[0.64]	[0.74]	[0.74]	[0.99]	[0.38]	[0.38]	[0.74]	[0.38]
Program X Ethnicity interaction	-0.71	-0.27	0.00	-0.82	0.37	-2.24	-0.51	-0.20
(SE)	(0.77)	(2.51)	(0.52)	(0.67)	(0.80)	(0.92)	(0.61)	(1.05)
[p]	[0.36]	[0.91]	[0.99]	[0.23]	[0.65]	[0.02]	[0.41]	[0.85]
[p FDR]	[0.36]	[0.99]	[0.99]	[0.80]	[0.99]	[0.14]	[0.96]	[0.99]
N	138	138	138	138	138	138	138	138

Table 11: Moderation by ethnicity on outgroup regard

Our estimates, even accounting for the confidence intervals, are smaller than one would have expected given Paluck, Green, and Green (2019)'s meta analysis. The standardized effect size of our program on social distance (0.066) – the indicator most well-aligned with Paluck et al. (2019) – is less than a third of the effect size of contact detected in their

recent meta-analysis (0.39 SD as the average effect across included experiments). Moreover, confidence interval of our pooled effect (-0.216; 0.348) does not include their effect size.

D Moderation by Age

	Outgroup Regard	Social Distance	Supp. Peace	Persp. Tk.	Host. Att. Self	Host Att. Peer	Opti. Peace	Ingroup Ident.
Interaction team (M is age)	-0.01	0.03	0.01	-0.02	-0.04	-0.02	-0.01	-0.05
(SE)	(0.02)	(0.07)	(0.02)	(0.02)	(0.02)	(0.04)	(0.02)	(0.02)
[p]	[0.58]	[0.64]	[0.51]	[0.3]	[0.14]	[0.57]	[0.74]	[0.07]

Table 12: Moderation by age on outgroup regard

	Ingroup Regulation	Effort Persuade	Censuring	Policing	Perspective Sharing	Perspective Sharing: Just Narratives
Interaction team (M is age)	0	0	0.05	-0.04	-0.01	0.02
(SE)	(0.02)	(0.01)	(0.05)	(0.02)	(0.02)	(0.03)
[p]	[0.99]	[0.6]	[0.36]	[0.15]	[0.77]	[0.42]

Table 13: Moderation by age on ingroup regulation

E Fusion Analysis Subgroup Results on Outgroup Regard

	Year		
	1	2	3
Treated mean (se-bootstrap)	18.05 (0.64)	18.73 (0.69)	19.96 (0.8)
Control mean (se-bootstrap)	18.2 (0.82)	17.76 (0.68)	18.16 (0.75)
Treatment effect (se-bootstrap)	-0.16 (1.02)	0.98 (0.95)	1.8 (1.09)
p-bootstrap	[0.87]	[0.29]	[0.1]

Table 14: Fusion Analysis Subgroup Results: All

	Year		
	1	2	3
Treated mean (se-bootstrap)	16.52 (0.77)	17.15 (0.82)	18.28 (0.94)
Control mean (se-bootstrap)	17.32 (1.02)	16.93 (0.85)	17.34 (0.97)
Treatment effect (se-bootstrap)	-0.8 (1.27)	0.21 (1.13)	0.94 (1.34)
p-bootstrap	[0.51]	[0.85]	[0.46]

Table 15: Fusion Analysis Subgroup Results: Arab-Palestinian Girls

	Year		
	1	2	3
Treated mean (se-bootstrap)	19.11 (0.71)	19.8 (0.76)	21 (0.89)
Control mean (se-bootstrap)	19.25 (0.93)	18.67 (0.85)	19.01 (0.85)
Treatment effect (se-bootstrap)	-0.13 (1.17)	1.14 (1.14)	1.99 (1.2)
p-bootstrap	[0.93]	[0.34]	[0.08]

Table 16: Fusion Analysis Subgroup Results: Jewish-Israeli Boys

	Year		
	1	2	3
Treated mean (se-bootstrap)	17.76 (1.09)	18.48 (1.06)	19.84 (1.1)
Control mean (se-bootstrap)	17.4 (1.38)	17.09 (1.24)	17.6 (1.3)
Treatment effect (se-bootstrap)	0.36 (1.65)	1.39 (1.54)	2.24 (1.62)
p-bootstrap	[0.82]	[0.36]	[0.17]

Table 17: Fusion Analysis Subgroup Results: Jewish-Israeli Girls

F Fusion Analysis Subgroup Results on Perspective Sharing

	Year		
	1	2	3
Treated mean (se-bootstrap)	4.68 (0.24)	4.97 (0.22)	5.29 (0.24)
Control mean (se-bootstrap)	4.49 (0.24)	4.56 (0.21)	4.56 (0.22)
Treatment effect (se-bootstrap)	0.18 (0.34)	0.41 (0.31)	0.73 (0.32)
p-bootstrap	[0.62]	[0.19]	[0.02]

Table 18: Fusion Analysis Subgroup Results: All

	Year		
	1	2	3
Treated mean (se-bootstrap)	4.92 (0.48)	5.11 (0.41)	5.18 (0.41)
Control mean (se-bootstrap)	5.04 (0.39)	5.10 (0.41)	5.10 (0.42)
Treatment effect (se-bootstrap)	-0.12 (0.59)	0.01 (0.57)	0.08 (0.59)
p-bootstrap	[0.81]	[0.99]	[0.90]

Table 19: Fusion Analysis Subgroup Results: Arab-Palestinian Girls

	Year		
	1	2	3
Treated mean (se-bootstrap)	4.78 (0.22)	5.10 (0.24)	5.49 (0.27)
Control mean (se-bootstrap)	4.37 (0.27)	4.43 (0.24)	4.43 (0.26)
Treatment effect (se-bootstrap)	0.41 (0.35)	0.67 (0.35)	1.06 (0.36)
p-bootstrap	[0.26]	[0.06]	[0.00]

Table 20: Fusion Analysis Subgroup Results: Jewish-Israeli Boys

	Year		
	1	2	3
Treated mean (se-bootstrap)	4.31 (0.30)	4.63 (0.28)	5.08 (0.31)
Control mean (se-bootstrap)	4.21 (0.45)	4.28 (0.38)	4.28 (0.37)
Treatment effect (se-bootstrap)	0.10 (0.52)	0.36 (0.46)	0.79 (0.47)
p-bootstrap	[0.86]	[0.43]	[0.10]

Table 21: Fusion Analysis Subgroup Results: Jewish-Israeli Girls

G Alternative Estimations of Treatment Effects

G.1 Outgroup Regard: OLS Estimation

Stat.	Year 1	Year 2	Year 3
Treated mean	14.6	16.71	18.31
(se-HC2)	(1.17)	(1.25)	(1.82)
Control mean	13.5	13.02	14.49
(se-HC2)	(1.68)	(1.65)	(1.79)
Treatment effect	1.1	3.69	3.82
(se-HC2)	(0.89)	(1.7)	(2.76)
p-HC2	[0.22]	[0.03]	[0.17]

G.2 Outgroup Regard: IPW Estimation

Stat.	Year 1	Year 2	Year 3
Treated mean	18.4	19.87	19.44
(se-HC2)	(0.8)	(1.26)	(2.89)
Control mean	18.95	17.25	19.09
(se-HC2)	(1.27)	(0.84)	(1.33)
Treatment effect	-0.55	2.63	0.34
(se-HC2)	(1.5)	(1.51)	(3.18)
p-HC2	[0.71]	[0.08]	[0.91]

G.3 Outgroup Regard: Unadjusted Estimation

Stat.	Year 1	Year 2	Year 3
Treated mean	17.56	17.51	17.73
(se-HC2)	(0.72)	(1.14)	(1.83)
Control mean	18.95	17.11	19.67
(se-HC2)	(1.27)	(0.83)	(1.1)
Treatment effect	-1.4	0.4	-1.94
(se-HC2)	(1.46)	(1.41)	(2.13)
p-HC2	[0.34]	[0.78]	[0.36]

G.4 Perspective Sharing: OLS Estimation

Stat.	Year 1	Year 2	Year 3
Treated mean	4.57	5.65	5.98
(se-HC2)	(0.32)	(0.36)	(0.53)
Control mean	3.86	4.23	4.37
(se-HC2)	(0.45)	(0.42)	(0.47)
Treatment effect	0.71	1.42	1.6
(se-HC2)	(0.26)	(0.51)	(0.76)
p-HC2	[0.01]	[0.01]	[0.04]

G.5 Perspective Sharing: IPW Estimation

Stat.	Year 1	Year 2	Year 3
Treated mean	4.52	5.02	5.08
(se-HC2)	(0.29)	(0.37)	(0.61)
Control mean	4.38	4.98	4.99
(se-HC2)	(0.42)	(0.4)	(0.6)
Treatment effect	0.14	0.04	0.09
(se-HC2)	(0.51)	(0.54)	(0.86)
p-HC2	[0.78]	[0.93]	[0.92]

G.6 Perspective Sharing: Unadjusted

Stat.	Year 1	Year 2	Year 3
Treated mean	4.6	4.91	4.49
(se-HC2)	(0.24)	(0.3)	(0.38)
Control mean	4.38	4.93	4.87
(se-HC2)	(0.42)	(0.33)	(0.49)
Treatment effect	0.22	-0.02	-0.38
(se-HC2)	(0.49)	(0.45)	(0.62)
p-HC2	[0.66]	[0.96]	[0.54]

6 Mediation Analysis

As pre-registered, we tested whether outgroup regard and self-esteem mediate the effect of intergroup contact on ingroup regulating behaviors. We estimate the average causal mediation effect (ACME) and average natural direct effect (ADE) using the regression approach proposed by (Tingley et al., 2014). To estimate the effect of the treatment on the mediator, we use the same specification as within the Data Analysis section. To estimate the effect of the mediator on the outcome, we augment the above specification to include the mediator variable (M_{ics}) as follows:

$$Y_{ics} = \alpha^{ITT} + \beta^{ITT} Z_{ics} + \gamma^{ITT} M_{ics} + \phi^{ITT} Z_{ics} \times M_{ics} + X'_{ics}\theta^{ITT} + \mu_c^{ITT} + \kappa_s^{ITT} + \epsilon_{ics}^{ITT},$$

Then, the ACMEs and ADEs for the treatment group, control group, and then the average of the two are calculated as per the methods of Tingley et al (2014). We report all three sets of estimates, as well as the implied estimate for the proportion of the total effect of Z on Y that is mediated by M.

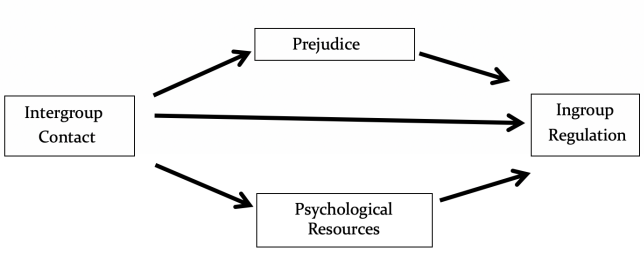

Figure 3: Mediation of the relationship between the Intergroup Contact and Ingroup Regulation by Outgroup Regard and Psychological Resources

A Mediation Results

We use the ACME and ADE estimators from Tingley et al. (2014), where inference is based on the non-parametric bootstrap. We use the indices to conduct the mediation analysis. Please note that the results are very unstable due to the near-zero estimates of the ITT effects. We are reporting these results in accordance with our pre-registration; however, the mediation analysis, and in particular the "proportion mediated" statistic, is not otherwise well-motivated given the lack of the main effect[1]:

	Coef.	CI lb	CI ub
Tr. ACME	-0.08	-0.44	0.25
Tr. ADE	0.11	-0.51	0.69
Tr. Prop. Mediated	2.26	-11.01	7.55
Co. ACME	-0.14	-0.73	0.41
Co. ADE	0.05	-0.69	0.74
Co. Prop. Mediated	3.94	-25.54	10.82
Avg. ACME	-0.11	-0.52	0.34
Avg. ADE	0.08	-0.64	0.69
Avg. Prop. Mediated	3.10	-18.27	8.80

Table 22: Estimates for Outgroup Regard as a mediator of Ingroup Regulation. ADE = average natural direct effect. Prop. Mediated = the proportion of the total effect on Ingroup Regulation that is mediated by effects on Outgroup Regard index.

	Coef.	CI lb	CI ub
Tr. ACME	0.07	-0.06	0.37
Tr. ADE	-0.06	-0.85	0.70
Tr. Prop. Mediated	-3.12	-3.06	2.55
Co. ACME	0.04	-0.09	0.17
Co. ADE	-0.09	-0.88	0.59
Co. Prop. Mediated	-1.78	-1.08	1.35
Avg. ACME	0.05	-0.06	0.22
Avg. ADE	-0.08	-0.87	0.64
Avg. Prop. Mediated	-2.45	-2.53	1.42

Table 23: Estimates for Personal Resources as a mediator of Ingroup Regulation. ADE = average natural direct effect. Prop. Mediated = the proportion of the total effect on Ingroup Regulation that is mediated by effects on Personal Resources index.

[1]What we refer to in the pre-analysis plan as "Promoting Peace" is here stated under a different term of "Ingroup Regulation" (similarly, we opted for Outgroup Regard" instead of "Prejudice"); the terms are conceptualized and operationalized in the same way.

7 Moderated-Mediator Effects

We also conducted a moderated-mediator analysis to look specifically at whether the mediation effects on ingroup regulation vary by ethnicity. To do this, we simply estimated the mediation effects separately for the two subsamples. Given low statistical power, we pre-registered this analysis as descriptive and report the results below.

	Coef.	CI lb	CI ub
Tr. ACME	-0.01	-0.63	0.54
Tr. ADE	0.10	-0.61	0.97
Tr. Prop. Mediated	-0.11	-3.28	10.14
Co. ACME	-0.01	-1.07	0.77
Co. ADE	0.10	-0.71	0.88
Co. Prop. Mediated	-0.16	-4.06	14.74
Avg. ACME	-0.01	-0.85	0.67
Avg. ADE	0.10	-0.64	0.95
Avg. Prop. Mediated	-0.14	-3.67	12.44

Table 24: Jewish-Israeli: Mediation effects of Outgroup Regard on Ingroup Regulation

	Coef.	CI lb	CI ub
Tr. ACME	-0.18	-0.76	0.39
Tr. ADE	-0.42	-1.55	0.71
Tr. Prop. Mediated	0.20	-1.23	2.62
Co. ACME	-0.47	-1.50	0.82
Co. ADE	-0.71	-1.64	0.76
Co. Prop. Mediated	0.53	-2.29	4.80
Avg. ACME	-0.33	-0.97	0.55
Avg. ADE	-0.57	-1.61	0.70
Avg. Prop. Mediated	0.36	-2.06	3.36

Table 25: Arab-Palestinian: Mediation effects of Outgroup Regard on Ingroup Regulation

	Coef.	CI lb	CI ub
Tr. ACME	0.10	-0.15	0.44
Tr. ADE	-0.08	-1.28	1.16
Tr. Prop. Mediated	-14.76	-4.01	1.13
Co. ACME	0.08	-0.20	0.41
Co. ADE	-0.10	-1.36	1.19
Co. Prop. Mediated	-11.47	-2.26	1.29
Avg. ACME	0.09	-0.14	0.38
Avg. ADE	-0.09	-1.30	1.17
Avg. Prop. Mediated	-13.11	-3.13	1.02

Table 26: Jewish-Israeli: Mediation effects of Personal Resources on Ingroup Regulation

	Coef.	CI lb	CI ub
Tr. ACME	-0.01	-0.32	0.69
Tr. ADE	1.00	-1.56	3.93
Tr. Prop. Mediated	-0.02	-0.57	0.56
Co. ACME	-0.29	-1.20	0.44
Co. ADE	0.72	-1.36	3.47
Co. Prop. Mediated	-0.41	-2.02	1.40
Avg. ACME	-0.15	-0.68	0.20
Avg. ADE	0.86	-1.45	3.75
Avg. Prop. Mediated	-0.22	-1.10	0.87

Table 27: Arab-Palestinian: Mediation of Personal Resources on Ingroup Regulation

8 Twinning Partner Analysis

As pre-registered, we are also interested in the ways in which one's own outcomes depend on whether one's experience in the program was more or less positive, which we measure through the nature of twinning partners. The assumption here is that participants had a more positive experience if their team twinned with a team in which members on average had low rather than high Outgroup Regard levels. Teams sometimes had multiple twinning partners, and so we constructed a twinning matrix for which each row shows the number of twinnings with a given other team. We note that teams with youth selected as part of the RCT assignment process typically twinned with teams that were fully outside the RCT, and so the twinning partner analysis requires that we use results from the general survey data.

For the twinning partner analysis, the first thing that we do is construct, for each individual, a measure of their exposure to other teams. This is simply a vector corresponding to the row for that individual's team in the twinning matrix. We can call this row the vector $W_{t[i]} = (W_{t[i],1}, ..., W_{t[i],T})'$, where $t[i]$ indexes subject i's team t, and we have T teams overall. Then, we have data on outcomes for members of subject i's twinning partner teams. For an outcome Y_{ics}, we can construct the team-level mean $\bar{Y}_t$, and the vector of team-level means is given by $\bar{Y} = (\bar{Y}_1, ..., \bar{Y}_T)'$. Then, subject i's exposure to such outcomes among twinning partners is measured in terms of the following linear combination:

$$TP(Y)_{t[i]} = W'_{t[i]}\bar{Y}.$$

For the analysis, we study how outcomes correlate with the exposure to twinning partner outcomes. We estimate the ITT equation, but instead of using the binary treatment variable (Z_{ics}), we use the $TP(Y)_t$ values.

	Social Distance	Self-Esteem	Willing. Discuss	In-gr. Censur.	In-gr. Police
Twinning effect	2.57	-4.99	-1.41	-1.09	0.13
(SE)	(4.11)	(4.73)	(0.8)	(1.08)	(1.44)
[p]	[0.56]	[0.31]	[0.1]	[0.33]	[0.93]
[p FDR]	[0.56]	[0.41]	[0.24]	[0.41]	[0.09]
Control mean	17.99	21.2	4.75	11.81	6.09
(Control SD)	(5.14)	(6.36)	(1.78)	(3.86)	(1.79)
N	83	83	83	64	64

Table 28: ITT Estimates with Twinning Exposure Measures

NB: The organization we collaborated with was not able to provide us with individual-level twinning data (i.e. number of twinnings that each individual participated in), which prevented us from furthering our twinning partner analysis.

9 Personal Resources

As pre-registered, we are also interested in the relationship between contact, personal resources and outcomes. Individuals can be effective at changing their peers' behaviors and opinions (Paluck, 2012; Paluck and Aronow, 2016; Tankard, 2016), but dissenting from one's peers to, for example, advocate on behalf of an outgroup, can be extremely cumbersome and costly. Group belonging provides various psychological benefits, particularly related to resilience and motivation (Chong, 2014), while deviating from valued groups can be threatening to one's self-esteem (Cialdini, 2004). In addition to changing attitudes, we should ask whether programs promoting peace also help build up psychological resources that participants need to advocate for peace vis-a-vis their ingroup peers. Only few studies so far test if such programs increase psychological resources, e.g., self-esteem (Ditlmann, Samii, and Zeitzoff, 2017). Increased personal resources may emerge from social interactions within the program and from doing well at sports.

Regulating peers, for example, through in-group censuring or policing, requires program participants to deviate from their communities. Given that conforming with members of the ingroup has many psychological benefits that participants lose when they deviate, they need to acquire the resources necessary for deviation through their participation in the program. Our pre-registered hypothesis stipulates that participating in an intergroup contact program within a conflict setting increases psychological resources. We test this hypothesis for three indicators of resources as well as one overall index that combines them.
We hypothesized that:

> H: Participating vs. not participating in a sports and peacebuilding program increases participants' self-esteem.

If, as past research suggests, minorities have a more pronounced need for respect and empowerment in intergroup interactions, they may benefit more in terms of resource building – if this need is fulfilled. This conjecture led us to test if this hypothesis is moderated by

ethnicity. Consistent with past research (Ditlmann and Samii, 2016; Ditlmann, Samii, and Zeitzoff, 2017), we propose that the program's impact on psychological resources should be stronger for Arab-Palestinian than Jewish participants and that it should grow stronger over time. We test these hypotheses for the overall Resources index and the self-esteem indicator (the below table shows principal component scores which informed our index creation).

	PC1	PC2	PC3
Confidence to persuade	0.032	-0.325	-0.945
Self-Esteem	0.995	-0.084	0.063
Appraised Risk of Intervening	0.099	0.942	-0.321

Table 29: PC scores for all the items that form the composite index of Personal Resources.

We initially also wanted to test how these combine to create change, gaining insights from the stress-and-coping frameworks in the psychology literature (Folkman et al., 1986; Van Zomeren, Leach, and Spears, 2012) that explain through what psychological processes people effectively manage adverse situations. We pre-specified testing of whether the more effectively a peace program increases participants' psychological resources, the more they engage in ingroup regulating behavior yet find no evidence of such a mediation.

Concept	*Indicator*	*Operationalization*
Personal Resources	Confidence to persuade	Ambiguous images question ("Do you think the kids agree...?")
Personal Resources	Self-esteem	10-items index; 4-point scales (survey 2015-2018)
Personal Resources	Appraised risk of intervening	In a vignette in which an ingroup member commits aggression toward an outgroup member, assessment that effort to stop aggression of the ingroup member is risky; 7-point scale.

Table 30: Personal Resources: Concepts and associated indicators

Sources for these operationalizations are as follows: the ambiguous images measure is an expansion of McGothlin & Killen (McGlothlin, 2006, 2010); the scale of self esteem is derived from (Heatherton and Polivy, 1991); and the final measure of appraised risk of intervening

is an expansion of (Huesmann, 1997), and (Möller and Krahé, 2009).

A Findings from the RCT: ITT Estimates and Ethnicity Moderator Effects

Below table presents the results of the effects of treatment on personal resources, from the ITT analysis pooling across members of both ethnic groups and controlling for a set of above-noted covariates (same as for the Outgroup Regard and Ingroup Regulation indices). Using the data pooled across ethnic groups, we find no evidence of a causal effect of the program on Personal Resources.

	Index	Conf. Pers	Self-Esteem	Appr. Risk
Program effect	0.12	0.43	0.40	0.12
(SE)	(0.11)	(0.27)	(0.57)	(0.27)
[p]	[0.25]	[0.11]	[0.48]	[0.66]
[p FDR]	[0.25]	[0.34]	[0.66]	[0.66]
Control mean	0.04	0.64	21.20	3.49
(Control SD)	(1.06)	(1.81)	(6.36)	(2.2)
N	138	138	138	138

Table 31: Intention-to-treat effect of treatment on the composite index of Personal Resources and all the corresponding items: confidence to persuade, self-esteem and risk appraisal.

Below we present the results from the ITT analysis that tests for heterogeneity by ethnic group controlling for a series of covariates and their interaction with the program. As we do for Outgroup Regard and Ingroup Regulation, we test whether ethnicity moderates the effect of the program on the overall Resources index and our main resource indicator: self-esteem. The results reveal a marginally significant program effect on the Personal Resources index for Jewish participants (SE=.20, p=.06) but the interaction does not reach significance (p=.44). For Arab-Palestinian participants, the program coefficient is -0.06 (p=0.85) for the Resource index and 1.29 (p=0.40) for self-esteem.

	Personal Resources	Self-Esteem
Program effect for Jewish-Israeli participants	0.20	0
(SE)	(0.11)	(0.72)
[p]	[0.06]	[0.99]
Program X Ethnicity interaction	-0.26	1.29
(SE)	(0.34)	(1.88)
[p]	[0.44]	[0.50]
N	138	138

Table 32: Ethnicity as moderating the effect of the program on composite index of Personal Resources and self-esteem scores.

B Findings from Fusion Analyses

The figure below shows the results of the fusion analysis for self-esteem across all participants and separately for the three locations. Since the error bars overlap with the point estimates for the treatment and control condition at all points in time, we conclude that there is no significant long-term effect of the program on self-esteem (at least within a 3-year time-frame). As in the RCT data, there is no substantial heterogeneity in the program's effect across the groups.

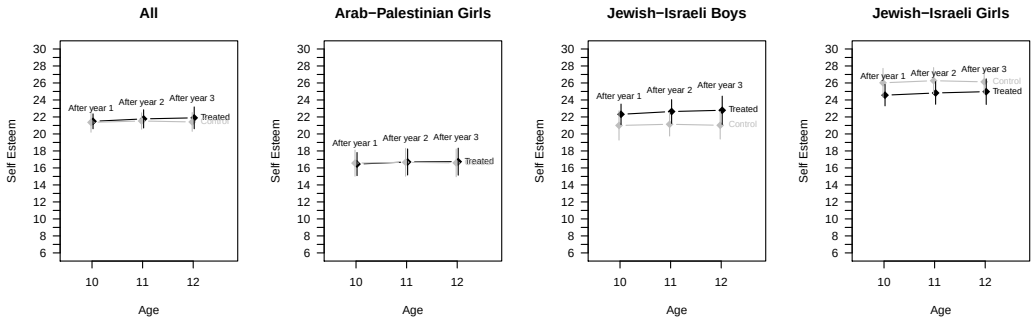

Figure 4: Results from the fusion analyses for the self-esteem indicator within the subgroups of Arab Palestinian girls, Jewish Israeli girls and Jewish-Israeli boys.

C Findings from Further Analysis of Observational Data

Finally, we also explore the observational data results from the regression of the self-esteem outcome on self-reported years in program. The results suggest that – even with some

selection presumably present in the cross-sectional survey – years spent in program are not associated with higher self-esteem once we include the covariates and fixed effects. However, we do find the predicted significant association between years in program and self-esteem scores in our small panel of 63 respondents (column 5).

Variable	(1)	(2)	(3)	(4)	(5)
(Intercept)	24.26	25.03	25.31	22.76	
(SE)	(0.29)	(0.62)	(0.69)	(0.81)	
[p]	[0]	[0]	[0]	[0]	
Years in prog.	0.17	0.2	0.09	0.24	3.57
(SE)	(0.09)	(0.25)	(0.3)	(0.3)	(1.14)
[p]	[0.06]	[0.42]	[0.77]	[0.43]	[0.01]
Age - 8		0.81	0.84	0.09	
(SE)		(0.42)	(0.45)	(0.44)	
[p]		[0.06]	[0.07]	[0.83]	
Age started - 8		-1.11	-1.15	-0.12	
(SE)		(0.45)	(0.47)	(0.47)	
[p]		[0.02]	[0.02]	[0.81]	
Years in prog. X Age started		0.01	0.01	-0.05	-0.46
(SE)		(0.05)	(0.05)	(0.05)	(0.22)
[p]		[0.78]	[0.82]	[0.33]	[0.05]
Arab			-0.45	-0.99	
(SE)			(0.56)	(0.58)	
[p]			[0.42]	[0.09]	
Years in prog. X Arab			0.17	0.37	-0.5
(SE)			(0.23)	(0.23)	(0.68)
[p]			[0.45]	[0.12]	[0.49]
N	554	554	554	554	554

Descriptive regressions for self-esteem indicator. Column 4 includes fixed effects for year and location. Column 5 includes individual-level fixed effects. 63 respondents have between 2 to 4 repeated values in the data, while 502 respondents appear only once.

Table 33: Global Self-Esteem

D Discussion of Resources Findings

Contrary to our hypothesis, we do not observe an impact of the program on Personal Resources in the RCT, fusion nor cross-sectional survey. While this finding does not support our hypothesis, it is not necessarily inconsistent with past literature. Even though it seems intuitive that there should be a link between athletic abilities and self-esteem, the empirical literature yields mixed conclusions (Noordstar et al., 2016) and attempting to link self-esteem to an intergroup program with a sports component is a novel contribution of our research. Within the peace curriculum that accompanied the interethnic sports experience, youth

learned about communal behavior with a large emphasis on team work. It is possible that this focus did not lend itself to fostering self-esteem - an individualistic concept (Rosenberg et al., 1995; Rosenberg and Simmons, 1972). For many Jewish participants, in particular, we might also be reaching a ceiling effect since their baseline self-esteem was already high.

Despite our mostly negative results, it seems worthwhile to further pursue this link in the future. First, we do observe a positive impact of years in program on self-esteem in our small panel, suggesting that there might be an effect for the most dedicated individuals who stayed in the program for more multiple years and repeatedly participated in our research. Second, the reported levels of global self-esteem for Arab-Palestinians in the league and league program were significantly higher than that of Arab-Palestinian participants who were not part of this group ($p = 0.04$). No significant difference was detected comparing the self-esteem of league vs. non-league Jewish-Israeli participants. This finding is consistent with the positive (albeit insignificant) interaction of years in program and being Arab-Palestinian we observe in the cross-sectional survey. While merely descriptive and tentative, this result speaks to past research on the importance for racial minorities to feel respected and seen as competent (Shelton et al., 2010); being part of a selected group playing within the all-star league or league program and taking part in all the public-facing activities of the NGO may have been instrumental in improving the perception of self for this group of minority participants.

Taken together, we find no strong evidence that the program in its current form fosters participants' self-esteem. Descriptively however, we observe benefits for the self-esteem of Arab-Palestinian league participants, suggesting an interesting avenue for future research.

10 Additional Ingroup Regulation Results

A Image Perspective Sharing Results

We operationalize perspective sharing through narrative (in the main text) and image sharing. The images shown to participants were shared with us by a local artist working with the program. The left one was shown to Jewish-Israeli and the right one to Arab-Palestinian participants, portraying different scenes from the violent conflict. Below are the images:

Figure 5: Image perspective sharing: survey question asked was "Would you share this image with ingroup members?"

It became clear to us that the complex nature of the events portrayed in the images may have led participants to respond in a way that goes beyond merely capturing one's willingness to share outgroup perspective with ingroup members (i.e. that something about the image itself affects users' willingness to share this particular image). Moreover, questions were translated from English language to Hebrew and Arabic by the local staff, and, due to nuances in translation, the teams were provided with slightly different phrasings of the question about image-sharing. Given some level of measurement error this may have induced, we focus more on interpreting the narrative perspective sharing measure in the main text. Nonetheless, we present the results associated with image sharing below:

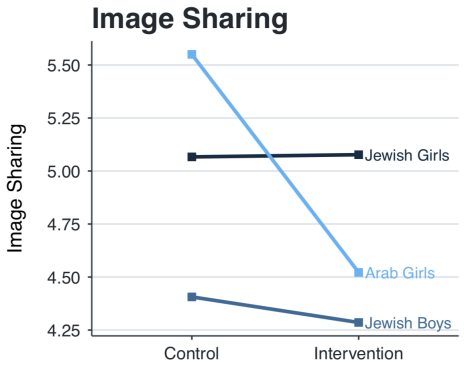

Figure 6: Willingness to share outgroup's perspective (image) with ingroup members.

B Main Effects and Moderation by Ethnicity

Hypothesis H2b stipulates that the impact of the year-long program on ingroup regulation should be stronger for Jewish-Israeli than Arab-Palestinian participants. As for our outgroup regard outcomes, with all interaction coefficients insignificant after FDR adjustments, we cannot reject the null hypothesis of homogeneity. Yet, the point estimates consistently point to greater program effectiveness for Jewish-Israeli participants, a point we return to below. Perspective sharing result includes willingness to share both the images and the narratives by or about the outgroup with ingroup members.

	Ingroup Regulation	Effort to Persuade	Censuring	Policing	Perspective Sharing
Program effect for Jewish-Israeli participants	0.59	0.17	0.73	-0.32	0.38
(SE)	(0.37)	(0.1)	(0.53)	(0.2)	(0.23)
[p]	[0.12]	[0.08]	[0.18]	[0.13]	[0.11]
[p FDR]	[0.12]	[0.17]	[0.18]	[0.17]	[0.17]
Program X Ethnicity interaction	-1.96	-0.38	-0.62	-0.25	-1.50
(SE)	(1.05)	(0.22)	(1.4)	(0.97)	(0.74)
[p]	[0.07]	[0.10]	[0.66]	[0.80]	[0.05]
[p FDR]	[0.07]	[0.20]	[0.80]	[0.80]	[0.20]
N	138	138	138	138	138

Table 34: Moderation by ethnicity (Model 2): Ingroup Regulation

C Censuring & Policing: Fusion Analysis

Below we present the fusion results on the two indicators within the family of ingroup regulation outcome (for which, as reported in the main text, we see no effect of the program):

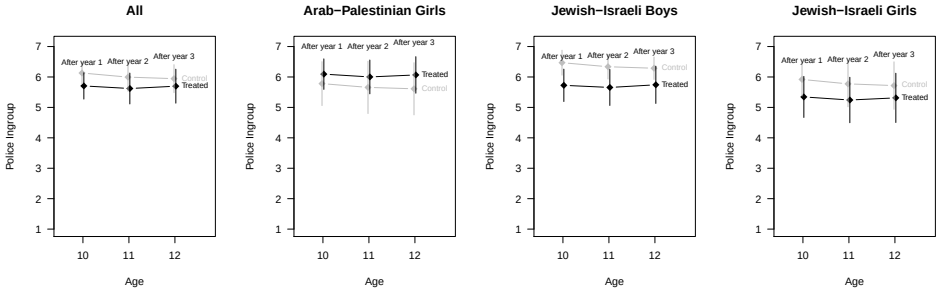

Figure 7: Fusion analysis of the policing measure

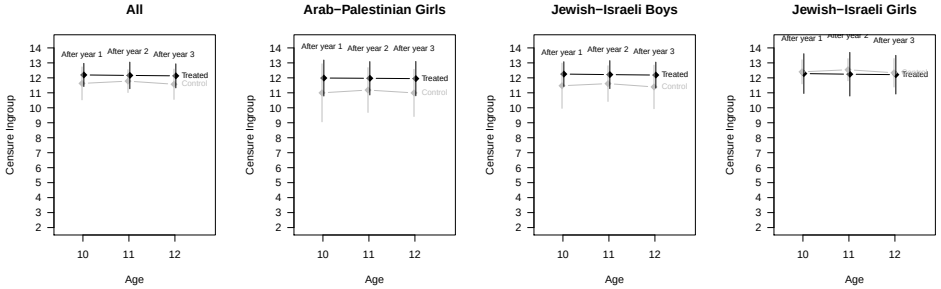

Figure 8: Fusion analysis of the censuring measure

11 Implementing the Fusion Analysis

The random forest allows us to make efficient use of our data while allowing for flexibly modeling of non-linearities and interaction effects. A fully non-parametric matching estimator would be very noisy, since our dataset is modestly sized (138 RCT observations plus 641 survey data observations). The way a random forest works is that it fits a large number of regression trees, and then averages over each of the trees. The `bartMachine` implementation incorporates methods based on a Bayesian model to help prevent overfitting. As Athey, Tibshirani, Wager, et al. (2019) show, an intuitive way to understand random forest models

is to think of them as "kernel" models that construct a prediction for the expected value at x^* with a weighted average of sample units' outcomes, where a unit's weight depends on how close its X value is to x^*. Such closeness is measured by the proportion of times that the regression trees used in the forest place the unit in a tree leaf that includes the value x^*. Figure 9 illustrates using simulated data.

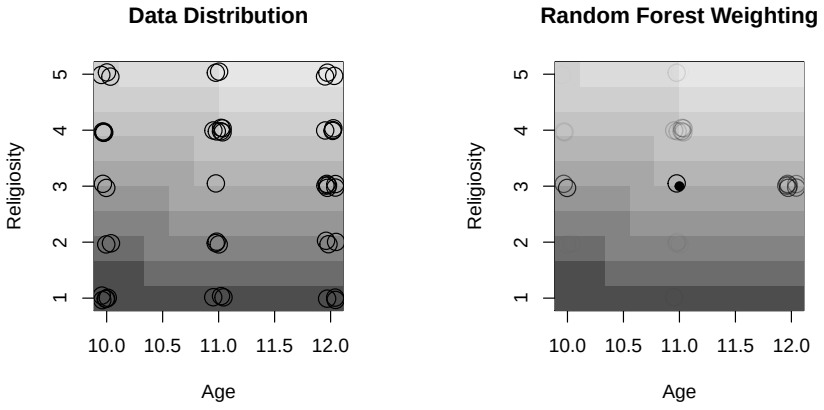

Figure 9: Illustration of random forest kernel weighting using simulated data. Suppose the data are distributed as shown at left, where the gray shading characterizes values of the outcome variable, and the x and y axes are different covariates. When we fit a random forest to these data and then generate a prediction for someone with age = 11 and religiosity index = 3, the prediction is based on a weighted average. The graph at right shows how observations are weighted in generating the predictions (darker outline means more weight).

12 Trimming Bounds for Attrition in the Fusion Analyses

Our fusion analysis attempts to predict the outcome trajectory after 1, 2, and 3 years in the program, and compare this to an outcome trajectory under control. To do this, we use the RCT data and then data from our general survey, where the latter gives us information on outcomes after 2 and 3 years in the program. A potential source of bias is that not all participants continue on for a second or third year. Table 35 is a cross-tabulation of the survey data by a subject's entering year and their number of years completed in the program at the time of data collection. We focus only on those who were in the program for 3 or fewer years. (The data contain subjects who participated for more than 3 years, but they are omitted from our analysis.) As is apparent, the numbers drop off after each year in the program. For example, our survey in 2015 captured 59 subjects who joined the program that year, but then when we conducted the survey the following year, we only captured 27 from that same 2015 cohort (who would have completed their second year in the program), and then only 15 in the year after that (third year in the program). We see similar patterns, sometimes less severe, sometimes more, for other cohorts.

Entering year:	2013	2014	2015	2016	2017	2018	2019	Total
1 year in program (N)	0	0	59	51	60	72	22	264
2 years in program (N)	0	36	27	46	21	8	0	138
3 years in program (N)	21	20	15	12	13	0	0	81

Table 35: Cross tabulation of survey respondents by their year of entry into the program and the number of program years completed at the time of data collection. Surveys were conducted in 2015-2019.

These patterns raise the issue of selection bias due to attrition. In many cases, such attrition is due to the fact that in a given locality and year, additional years of programming simply weren't offered. This explains much of the large drop-off between 2018 and 2019, for example. Our analysis controls for year and, at least in a coarse manner, locality, and so

selection biases associated with these sources of attrition are addressed by these variables being included in the random forest prediction algorithm. In other cases though, additional years of programming were offered, and participants choose whether or not to continue. Such choices are what we might expect as the source of selection bias.

Table 36 shows how the covariate profiles of RCT and survey participants vary for those with 1, 2, and 3 years in the program, compared to our target group (10 year old RCT participants). We see that those who stay in the program longer are substantially more likely to be Female and Arab, less likely to have parents born in Israel or to live in Jerusalem, and have parents' occupational prestige levels (a measure of household socio-economic status) that are somewhat lower for fathers but higher for mothers. The entropy weighting reduces some of these gaps, but given the relatively modest samples sizes, cannot close them entirely. This is one of the reasons that we prefer the non-parametric model-based approach using the random forest. Also, even if we balance these covariates, it leaves open the possibility of sample selection due to unobservables or outcomes.

	Target	Raw	Weighted	Raw	Weighted	Raw	Weighted
Years in program	0.55	1.00	1.00	2.00	2.00	3.00	3.00
Age	10.00	10.70	10.37	11.07	11.35	11.11	11.73
Year of data collection	2015.53	2016.50	2015.84	2016.63	2016.31	2016.53	2016.36
Female	0.55	0.63	0.58	0.60	0.53	0.64	0.59
Arab	0.26	0.44	0.35	0.52	0.31	0.73	0.48
Religiosity index	2.00	2.03	2.02	1.98	2.04	1.89	1.94
Father born in Israel	0.87	0.71	0.78	0.70	0.81	0.73	0.81
Mother born in Israel	0.72	0.64	0.67	0.66	0.71	0.71	0.71
Jerusalem	0.70	0.60	0.66	0.54	0.64	0.40	0.62
Father occup. prestige	2.32	2.06	2.17	1.92	2.12	2.22	2.29
Mother occup. prestige	2.09	2.29	2.22	2.17	1.94	2.44	2.32
N	47	242	242	89	89	45	45

Table 36: Covariate means for the fusion analysis target sample (10 year old RCT participants) and then for the raw and entropy-balancing reweighted samples of subjects 1 year (RCT treated participants plus survey participants with 1 year), 2 years (only survey participants), and 3 years (only survey participants) since program initiation.

Given these limitations, we follow Lee (2009) to use a trimming bounds approach to assess the robustness of our conclusions to selection bias. The idea is that our estimates for the treatment group represent how outcomes evolve among *those who stay in the program*. However, our estimates for the control group represent a mixture of those who would stay

and those who would drop out. Because the control group subjects are never treated, they contain a mix of types who would stay or drop out, and we cannot distinguish one from the other. We are seeking to characterize effects for subjects who stay, but our control group includes both subjects who would always stay and those who would drop out.

What we can do is to estimate the proportion of these "drop out" types and then trim the upper and lower tails of the control outcome distributions to get bounds on the mean outcome for the "stay" types. To estimate the proportion of dropout types, we use the treatment group data to calculate shares of subjects who stay from one year to the next, conditional on all of the covariates that we include in the fusion analysis. We then trim the conditional distributions in the control group by these conditional dropout rates, and aggregate to form the bounds.

With panel data, estimating the conditional dropout rates is simple because one can observe directly who drops out. In our case, we have repeated cross-section data from a population that differs from the RCT participants. Therefore, we need to use an indirect approach based on predicted missingness rates for individuals with covariates that match those of our RCT sample. Define the response indicator $S_{it} = 0, 1$ for whether unit i drops out or remains in the sample in $t = 1, 2, 3$ periods after treatment. Let $Z_i = 0, 1$ be an indicator for i is in the control or treatment condition. And finally suppose covariates are given by X_i. The quantity $p(S_{it} = 0 \mid Z_i = 1, X_i = x)$ is the conditional dropout rate for treated units in period t with $X_i = x$. This defines the share of control group observations with $X_i = x$ that we need to trim to get the bounds for period t. We can derive an expression for $p(S_{it} = 0 \mid Z_i = 1, X_i = x)$ based on the treatment propensity score—i.e., $p(Z_i = 1 \mid X_i = x)$. For our target sample (the RCT sample of 10 year olds), we observe $p(Z_i = 1 \mid X_i = x)$. For 2 and 3 years after treatment, we can estimate $p(Z_i = 1|X_i = x, S_{i2} = 1)$ and $p(Z_i = 1|X_i = x, S_{i3} = 1)$ with the data from the survey and the RCT control group. Since the control group is a random subsample of our target population, there is no missingness, implying what Lee (2009) refers to as "monotonic"

missingness. We can now observe,

$$\begin{aligned} p(Z_i = 1|X_i = x) &= p(Z_i = 1|X_i = x, S_{it} = 1)p(S_{it} = 1|X_i = x) + \underbrace{p(S_{it} = 0|X = x)}_{\text{(by monotonicity)}} \\ &= p(Z_i = 1|X_i = x, S_{it} = 1)[1 - p(S_{it} = 0|X_i = x)] + p(S_{it} = 0|X_i) \\ \Leftrightarrow p(S_{it} = 0|X_i = x) &= \frac{p(Z_i = 1|X_i = x) - p(Z_i = 1|X_i = x, S_{it} = 1)}{1 - p(Z_i = 1|X_i = x, S_{it} = 1)}, \end{aligned}$$

while by monotonicity again, we have,

$$p(S_{it} = 0|X_i = x) = p(S_{it} = 0|X_i = x, Z_i = 1)p(Z_i = 1|X_i = x).$$

Therefore,

$$p(S_{it} = 0|X_i = x, Z_i = 1) = \frac{p(Z_i = 1|X_i = x) - p(Z_i = 1|X_i = x, S_{it} = 1)}{p(Z_i = 1|X_i = x)[1 - p(Z_i = 1|X_i = x, S_{it} = 1)]}.$$

Intuitively, by examining how the covariate distribution of the survey participants after t years of treatment differs from the RCT participants, we can infer the conditional missingness rates.

The estimated $p(S_{it} = 0|X_i = x, Z_i = 1)$ values are used to construct the bounds. We use the random forest algorithm to estimate the conditional propensity scores, $p(Z_i = 1|X_i = x)$ and $p(Z_i = 1|X_i = x, S_{it} = 1)$. For each observation in our target sample (the RCT sample of 10 year olds) with $X_i = x$, we can use the random forest algorithm on the survey data to estimate a conditional outcome *density* after 2 or 3 years of treatment. We can then trim the bottom $p(S_{it} = 0|X_i = x, Z_i = 1)$ portion of this density and take the mean of this trimmed distribution to get an *upper bound* on the conditional outcome mean, and trim the upper portion of the density by the same amount and take the trimmed mean to get a *lower bound* on the conditional outcome mean. The upper and lower bounds on the average treatment effects for our target sample are simply the averages of these conditional upper and lower

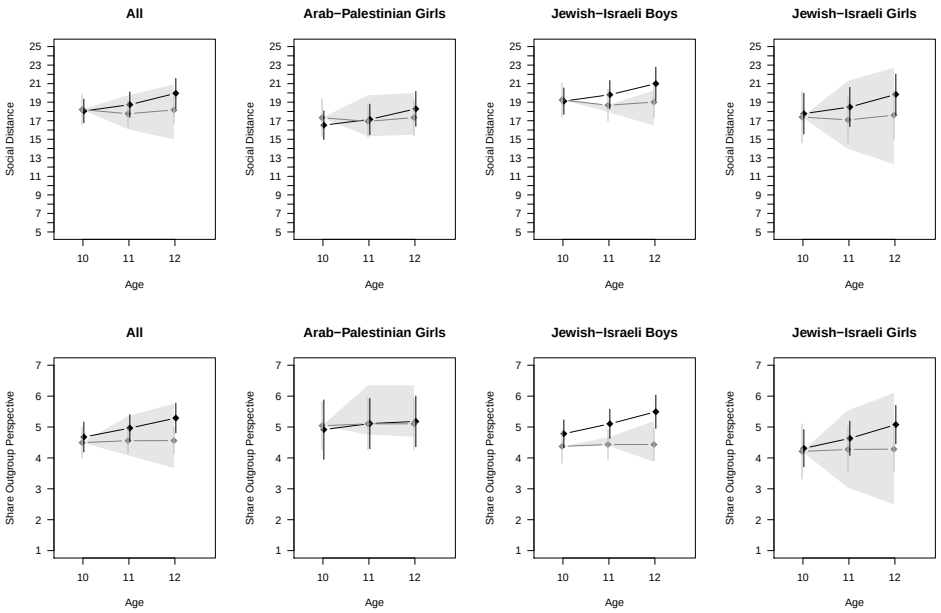

Figure 10: Trimming bounds on control group estimates to account for attrition. The graphs reproduce the point estimates for the treated and control trimming gray polygons show the trimming bounds for the control group trajectories.

bounds.

The results for the outgroup regard and perspective sharing outcomes (which also appear in the main text) are shown in Figure 10. The gray polygons show the range of the trimming bounds for the control group trajectories. When the gray area overlaps with the treatment group trajectory, it indicates that selection biases could imply that effects are either positive or negative. We see that the attrition results in such ambiguity for both Arab-Palestinian girls and Jewish-Israeli girls. The width of these bounds is a function of both the attrition rate (which determines how much of the control group distribution is trimmed) and then also the variance of the outcome distribution (which determines how much the trimming affects the mean estimate). The combined effect for Arab-Palestinian girls and Jewish-Israeli girls is substantial. For Jewish-Israeli boys, the original results are robust to attrition, in that even in the worst case scenario, we still see indication of a positive effect.

13 Comparison with the Pre-Analysis Plan

Pre-Specified Analyses (PAP)	**If not presented in the main text, why?**	**Additional Analysis/Changes**
Intent-to-Treat Analysis: 1. Outgroup Regard/Prejudice 2. Self-esteem/Personal Resources 3. Ingroup Regulation	Self-esteem/Personal Resources results in appendix to preserve space in main text for discussion of more important findings. We find no evidence of effects on Self-esteem/Personal Resources.	Distinguish between two ways of operationalizing outgroup perspective sharing (narratives vs. images).
Mediation analysis: 1. Prejudice as mediator of Ingroup Regulation 2. Self-Esteem as mediator of Ingroup Regulation	Results in the Appendix. We find no evidence of mediation. In main text we discuss implications for theory.	NA
Moderator analysis: 1. Ethnicity Moderator Effects on main indices 2. Moderated Mediator Effects 3. Twinning Partner Moderator Effects	Ethnicity Moderator Effects reported. As above for moderated mediator effects. The lack of availability of individual level twinning data prevents us from conducting the twinning partner moderator analysis.	We present and provide discussion of differences in effects at the sites where Jewish boys versus Jewish girls were enrolled, noting that other community level factors vary as well.
Twinning Analysis: 1. ITT Estimates with Twinning Exposure Measures 2. Simpler Twinning Analysis (number of twinnings)	Results in the Appendix. Our final dataset included team-level, rather than individual-level number of twinnings we hoped to obtain. Low variation in the number of twinnings (and the lack of reliable data on the number of twinnings on an individual level) makes the analysis less informative.	NA
Fusion Estimates of Developmental Effects: 1. Social Distance 2. Self-Esteem 3. Sharing Outgroup Perspective 4. Ingroup censuring 5. Ingroup policing 6. Appraised risk of intervening	4, 5 and 6 are in the Appendix to preserve space in main text to focus on larger contributions.	NA
Cross-sectional survey and descriptive results: 1. Comparison between league vs. non-league participants 2. Interaction between perceived levels of parental approval of the program and years in the program 3. Information about participants' discussions of outgroup's perspective with ingroup members 4. Sensitivity to attrition in the fusion analysis	All results discussed except 2, for which data could not be obtained.	NA

Table 37: Deviations from the Pre-Analysis Plan

14 Research Ethics

Our study adheres to the guidelines provided by the American Political Science Association ("Principles and Guidance for Human Subjects Research", 2020), with the well-being and safety of the subjects prioritized throughout our conduct of the study. Prior to initiating this research, we obtained the approval of the Institutional Review Board (IRB) under protocol number 13-9496. Data collection was overseen by individuals trained in the IRB-approved standards and protocols; these individuals include a graduate student fluent in both Arabic and Hebrew who administered the research instruments as well as staff members from the collaborating organization that facilitated access to subjects.

The research was embedded within the existing programming of the organization we partnered with. The organization successfully operated in the Middle East since 2005, leveraging sports to foster tolerance and facilitate meaningful interactions between Jewish and Arab youth. Consistent with Principle 10(b) of the *2020 APSA Principles and Guidance for Human Participants Research*, we were fully transparent with our partner organization about our research objectives, and have collaborated closely with their team throughout the study. This partnership was mutually beneficial: it provided us, the researchers, an opportunity to study the effects of intergroup contact within a conflict setting, while providing the organization with data-driven insights on how to best evaluate and enhance the impact that it aims to have on the communities it serves. The RCT component of the study was undertaken in a manner that minimized the disruption to the existing program: for example, control group participants were often admitted into the program after one season, and we have sometimes allowed crossovers if coaches or teachers had a strong preference that a certain child would benefit from being in the treatment group during a particular season. In fact, we have specifically developed the fusion analysis in order to not disrupt the program, but still maximize learning in our research. Doing so allowed us to overcome the limitations of the data we were able to obtain from the RCT component of the study (i.e., smaller sample size and inability to capture effects beyond one year of participation), while not affecting

the experiences of individuals directly engaged in the program.

As for consent with respect to the intervention, all participants in the sports program voluntarily applied with their parents. As for consent with respect to data collection, the organization with which we collaborated obtained parental permission forms for all the participants prior to the data collection events. The form was part of the IRB-approved protocol and clearly stated the aim of the research (*"Your child has been invited to take part in a research study to learn how contact between Israeli and Palestinian youth, or lack thereof, can affect youths' perceptions of themselves, their communities, and people from other communities and ethnicities."*), explained the expectations (*"If you give permission for your child's participation in this study, your child will be asked to take part in a survey interview, games, and computerized test that asks about his/her background (age, gender, education, etc) as well as perceptions about the future, friends, leaders in his/her community, and people of different ethnicities."*), and provided parents with information about who is conducting the study and all the steps taken to protect the privacy of their children.

We maintained confidentiality of all research subjects by provided each participant with a unique code that was linked to participants' responses. These codes are what we use to distinguish subjects in any data accessible to people outside the research team. In selecting our survey measures, we ensured that all the questions are age appropriate. In that spirit, we also refrained from any questions that would require participants to assess political actions or actors, given the tenuous political situation within the context we study. No financial incentives were provided to participants.

References

Athey, Susan, Julie Tibshirani, Stefan Wager, et al. (2019). "Generalized random forests". In: *Annals of Statistics* 47.2, pp. 1148–1178.

Chong, Dennis (2014). *Collective action and the civil rights movement*. University of Chicago Press.

Cialdini Robert B., and Noah J. Goldstein (2004). "Social influence: Compliance and conformity". In: *Annu. Rev. Psychol.* 55, pp. 591–621.

Ditlmann, Ruth and Cyrus Samii (2016). "Can intergroup contact affect ingroup dynamics? Insights from a field study with Jewish and Arab-Palestinian youth in Israel," in: *Peace and Conflict: Journal of Peace Psychology* 22.4, p. 380.

Ditlmann, Ruth, Cyrus Samii, and Thomas Zeitzoff (2017). "Addressing Violent Intergroup Conflict from the Bottom Up?" In: *Social Issues and Policy Review* 11.1, pp. 38–77.

Folkman, Susan, Richard S Lazarus, Christine Dunkel-Schetter, Anita DeLongis, and Rand J Gruen (1986). "Dynamics of a stressful encounter: cognitive appraisal, coping, and encounter outcomes." In: *Journal of personality and social psychology* 50.5, p. 992.

Heatherton, Todd F and Janet Polivy (1991). "Development and validation of a scale for measuring state self-esteem." In: *Journal of Personality and Social psychology* 60.6, p. 895.

Huesmann L. Rowell, and Nancy G. Guerra (1997). "Children's normative beliefs about aggression and aggressive behavior." In: *Journal of personality and social psychology* 72.2, p. 408.

Lee, David S. (July 2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects". In: *The Review of Economic Studies* 76.3, pp. 1071–1102. ISSN: 0034-6527. DOI: 10.1111/j.1467-937X.2009.00536.x. eprint: https://academic.oup.com/restud/article-pdf/76/3/1071/18350419/76-3-1071.pdf.

Lin, W. (2013). "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique". In: *The Annals of Applied Statistics* 7.1, pp. 295–318.

McGlothlin Heidi, and Melanie Killen (2006). "Intergroup attitudes of European American children attending ethnically homogeneous schools". In: *Child Development* 77.5, pp. 1375–1386.

— (2010). "How social experience is related to children's intergroup attitudes". In: *European Journal of Social Psychology* 40.4, pp. 625–634.

Möller, Ingrid and Barbara Krahé (2009). "Exposure to violent video games and aggression in German adolescents: A longitudinal analysis". In: *Aggressive Behavior: Official Journal of the International Society for Research on Aggression* 35.1, pp. 75–89.

Noordstar, Johannes J., Janjaap van der Net, Suzanne Jak, Paul J.M. Helders, and Marian J. Jongmans (2016). "Global self-esteem, perceived athletic competence, and physical activity in children: A longitudinal cohort study". In: *Psychology of Sport and Exercise* 22, pp. 83–90.

Paluck Elizabeth Levy, and Hana Shepherd (2012). "The salience of social referents: A field experiment on collective norms and harassment behavior in a school social network." In: *Journal of personality and social psychology* 103.6, p. 899.

Paluck Elizabeth Levy, Hana Shepherd and Peter M. Aronow (2016). "Changing climates of conflict: A social network experiment in 56 schools". In: *Proceedings of the National Academy of Sciences* 113.3, pp. 566–571.

Rosenberg, Morris, Carmi Schooler, Carrie Schoenbach, and Florence Rosenberg (1995). "Global self-esteem and specific self-esteem: Different concepts, different outcomes". In: *American sociological review*, pp. 141–156.

Rosenberg, Morris and Roberta G. Simmons (1972). "Black and White self-esteem: The urban school". In: *Child, American Sociological Association* 1.

Shelton, J. Nicole, Thomas E. Trail, Tessa V. West, and Hilary B. Bergsieker (2010). "From strangers to friends: The interpersonal process model of intimacy in developing interracial friendships". In: *Journal of Social and Personal Relationships* 27.1, pp. 71–90.

Tankard Margaret E., and Elizabeth Levy Paluck (2016). "Norm perception as a vehicle for social change". In: *Social Issues and Policy Review* 10.1, pp. 181–211.

Tingley Dustin, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai (2014). "Mediation: R package for causal mediation analysis". In.

Tourangeau, R. and T. Yan (2007). "Sensitive questions in surveys." In: *Psychological bulletin* 133.5, p. 859.

Van Zomeren, Martijn, Colin Wayne Leach, and Russell Spears (2012). "Protesters as "passionate economists" a dynamic dual pathway model of approach coping with collective disadvantage". In: *Personality and Social Psychology Review* 16.2, pp. 180–199.