Political Science Research and Methods (2024), page 1 of 19
doi:10.1017/psrm.2024.8

ORIGINAL ARTICLE

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

Nejla Asimovic[1], Ruth K. Ditlmann[2] and Cyrus Samii[3]

[1]University of Pennsylvania, Philadelphia, PA, USA, [2]Hertie School, Berlin, Germany and [3]Department of Politics, New York University, New York, NY, USA
Corresponding author: Nejla Asimovic; Email: nejla.asimovic@gmail.com

(Received 9 June 2022; revised 13 September 2023; accepted 19 September 2023)

Abstract

We study how an intervention combining youth intergroup contact and sports affects intergroup relations in the context of an active conflict. We first conduct a randomized controlled trial (RCT) of one-year program exposure in Israel. To track effects of a multiyear exposure, we then use machine-learning techniques to fuse the RCT with the observational data gathered on multiyear participants. This analytical approach can help overcome frequent limitations of RCTs, such as modest sample sizes and short observation periods. Our evidence cannot affirm a one-year effect on outgroup regard and ingroup regulation, although we estimate benefits of multiyear exposure among Jewish-Israeli youth, particularly boys. We discuss implications for interventions in contexts of active conflict and group status asymmetry.

Keywords: civil/domestic conflict; ethnicity and nationalism; experimental research; field experiments; quantitative methods

1. Introduction

Governments, international organizations, and private donors commit substantial resources each year to interventions working with ordinary citizens, and often youth, with the goal of promoting peace in violence-affected countries. If the violence is between ethnic groups, such interventions often involve bringing members of these groups together. Research has examined if contact with individual members can reduce prejudice toward the whole outgroup with hundreds of studies (Tropp *et al.*, 2012; Paluck *et al.*, 2019). Yet, a disproportionally large number of these studies relies on surveys or takes place in a laboratory. Few have tested intergroup contact in real-world environments, and even fewer in conflict settings.

Those that do, such as Scacco and Warren (2018), Mousa (2020), or Zhou and Lyall (2022), have come away with results suggesting mixed effects in the short term with ambiguous implications for the longer term. In this paper, we examine the impact of youth contact in conflict settings with a randomized controlled trial (RCT) embedded within an existing program in Israel, and a "data fusion" analysis to characterize the effects of multiyear contact exposure. Our study thus contributes substantively to the understanding of contact in conflict settings, and methodologically to the estimation of long-term predictions from cross-sectional data through a combination of experimental and observational data.

We partnered with a civil society organization that has offered an intergroup contact intervention to Jewish-Israeli and Arab-Palestinian youths for 14 years across 20 communities. The contact experience is embedded within a sports program, similarly to other recent intergroup contact

field studies (Mousa, 2020; Lowe, 2021). The program intends to be developmental, working with youth over multiple years. Hence, a fair evaluation ought to account for accumulated exposure over multiple years. Doing so would allow us to go beyond the relatively short-term exposure in recent field studies. In our field research, as in many similar cases, practical and ethical constraints make it difficult to sustain treatment-control differences to evaluate multiyear effects. One of the main contributions of our research is the strategy that we devised to overcome these constraints: using machine-learning techniques, we supplement a modestly sized, one-year experiment ($N = 138$) with a large repeated cross-sectional survey of one-year and multiyear participants ($N = 645$) which was more in line with the monitoring and evaluation strategy of our partner. In doing so, we contribute to the growing body of research that integrates observational and experimental data to estimate treatment effects (Künzel *et al.*, 2019; Colnet *et al.*, 2020; Kallus and Mao, 2020; Li *et al.*, 2020; Rosenman *et al.*, 2020; Imbens *et al.*, 2022; Degtiar and Rose, 2023). The increased sample size also allows us to estimate group-specific effects with reasonable precision. Causal identification is based on the fact that the participants selected for the experiment are drawn from the same population as those who were part of the cross-sectional survey, and assumes that there are no systematic, unobservable difference between the experimental and cross-sectional survey samples. This analytical strategy, we argue, can be particularly useful to researchers who want to maximize learning while working with resource-constrained civil-society organizations.

While our evidence cannot affirm positive program effects in the one-year RCT, the analysis of multiyear exposure suggests that program effects become more pronounced over time in ways that clearly distinguish Jewish-Israeli from Arab-Palestinian participants. For Jewish-Israeli boys in particular, effects are strong and robust to attrition. These results highlight how challenging it is to implement a program that benefits all groups equally.

2. Contact in conflict settings

Intergroup contact theory has been lauded as one of the best-tested tools for reducing intergroup prejudice. This recommendation is often based on a seminal meta-analysis of 515 studies that concluded that intergroup contact works (Pettigrew and Tropp, 2006). Yet, most of these studies are correlational and less than three percent involve groups in conflict. Recognizing this gap, a second meta-analysis of intergroup contact in applied settings assembled a database of an impressive 129 samples from outside of laboratory or survey contexts (Lemmer and Wagner, 2015). Again, the conclusion was that contact works—even in conflict settings. A closer look, however, reveals that only 11 samples evaluating five real-world programs stem from conflict settings. More than a third of the effect sizes are null or negative and none of the studies used a randomized design.

Considering how much scholarly attention intergroup contact theory has received, we know surprisingly little about its validity in real-world programs, especially in conflict settings. Conducting research on intergroup contact in applied conflict settings is particularly difficult which is why, despite great interest in the topic, few scholars have successfully implemented randomized designs there. Recent exceptions are one experiment in Nigeria (Scacco and Warren, 2018) and one in Iraq (Mousa, 2020), which find mixed evidence on intergroup relations with limited to no impact of intergroup contact on outgroup attitudes. Thus, before we can derive policy lessons from intergroup contact theory for applied conflict settings, we need more rigorous tests in applied conflict settings. Our innovative fusion design that combines RCTs with a repeated cross-sectional survey is a tool for conducting such tests. It thus offers technological and analytical advancements to intergroup contact research (Paolini *et al.*, 2021).

Most intergroup contact experiments expose participants for a short time, often one research session, to the outgroup. Yet, intergroup contact is more likely to succeed if repeated: over time individuals develop resources to appraise intergroup interactions as non-stressful (Mendes *et al.*, 2007) and bond with outgroup members (Wright *et al.*, 2005). This is particularly important in

conflict societies where intergroup anxiety is higher, and the effects of repeated contact rarely evaluated. Our fusion design allows us to examine such multiyear contact effects, even when it is difficult to implement a multiyear RCT design.

Israel is an important context for studying intergroup contact, with research in applied programs painting a mixed picture. Non-randomized evaluations of a summer camp (Schroeder and Risen, 2016) and several sport programs (Galily *et al.*, 2013b) showed reduced prejudice effects. A natural experiment found intergroup contact between Jewish-Israeli patients and Arab doctors reduced Jewish patients' anti-Arab prejudice (Weiss, 2021). Other studies illuminate the challenges for implementing intergroup contact programs in the field: the emotionally charged nature of facilitated dialogues (Kahanoff, 2016), hardening of group boundaries in face of security threats and power discrepancy (Hammack, 2006), or even logistical challenges in executing these programs (Litvak-Hirsch *et al.*, 2018).

2.1 Majority versus minority participants and group status

Intergroup contact programs require participation of members from at least two adversarial groups. The program we worked with in Israel brings together youth from a majority (Jewish-Israelis) and a minority (Arab-Palestinians residing within Israel) group. Ideally, both groups should benefit from the program, although evidence suggests that minorities tend to benefit less, and are less likely to change their outgroup attitudes as a function of intergroup contact (Tropp and Pettigrew, 2005). Creating a program that serves both groups equally in the field is complicated, as evidence suggests that members of different status groups have different priorities for intergroup encounters. Many civil-society interventions with a focus on commonalities (e.g., common humanity) reflect the preference of majority groups to be liked and accepted by the minority group (Dovidio *et al.*, 2009). Meanwhile, members of the minority group tend to prefer for societal inequalities and injustices to be more directly addressed (Hässler *et al.*, 2022). This dynamic leads us to expect the program to be less successful in changing the prejudice and ingroup-regulation behavior of Arab-Palestinian participants compared to Jewish-Israeli participants.

2.2 Intergroup contact outcomes

We measure prejudice through several measures of outgroup regard, but also assess ingroup regulation outcomes. In conflict-affected societies where everyday opportunities for intergroup contact are rare (Enos, 2017), civil-society programs are often among the very few modes of intergroup contact. Despite their potential, these programs typically do not reach a significant share of a country's population. In addition to reducing prejudice, it is thus desirable that intergroup contact programs increase participants' motivation to influence their non-participant ingroup peers (Ditlmann *et al.*, 2017).

Influence strategies may include ingroup censuring, ingroup policing, persuasion of outgroup positive intent, and perspective sharing. Ingroup censuring (Ditlmann and Samii, 2016) and policing (Fearon and Laitin, 1996) entail privately or publicly condemning an ingroup members' aggression toward the outgroup. In terms of persuasion, people can convince their peers that outgroup members have benevolent instead of hostile intentions, thus disrupting hostile attribution biases (McGlothlin and Killen, 2006; van Dijk *et al.*, 2019), or share the perspective of the outgroup with their ingroup members (Bruneau and Saxe, 2012).

3. Research hypotheses

Based on the preceding discussion, our core prediction is in keeping with the original contact hypothesis:

H1a: Participating versus not participating in an intergroup contact program within a conflict setting causes an increase in positive outgroup regard.

H1b: The longer youths participate in the program, the stronger the positive effect of intergroup contact on participants' outgroup regard.

Following past research which suggest that the association between intergroup contact and outgroup regard is less conclusive for minority than for majority group members (Tropp and Pettigrew, 2005), we also tested if:

H1c: H1a and H1b are moderated by participants' ethnicity.

As outlined earlier, another way in which contact experiences may positively affect group dynamics is by strengthening individuals' tendency to influence their non-participant peers.

H2a: Participating versus not participating in an intergroup contact program within a conflict setting increases participants' tendency to regulate their peers.

H2b: The longer youths participate in the program, the stronger the positive effect of intergroup contact on participants' tendency to regulate their peers.

We also tested, following the same logic as above, if:

H2c: H2a and H2b are moderated by participants' ethnicity.

Our pre-analysis plan also hypothesized effects on individual self-esteem, as a potential mediator of effects on peer regulation. We find no effects on self-esteem and, given space constraints, point readers to our Appendix (Section 9), which contains all of those results. The Appendix (Table 37) also explains all other deviations from our pre-analysis plan.

4. Intervention

Our study examines the impact of an existing field intervention that combines sports with intergroup contact[1] —a sports league for Arab-Palestinian and Jewish-Israeli youth in Israel of age 4–20. Our research focuses on young adolescents (from 8 to 16 year olds, with the majority of participants being between the age of 10 and 13), a period of intensive socialization into political and social life beyond the family (Lerner and Steinberg, 2009).

The organization—through schools or community organizations—invites youth to join a team with about 12 peers from the same residential community, ethnic group, and gender, leading with high-level sports training. Indeed, participants from both groups report the desire to play sports as their main reason for joining the program (Figure A1). This is echoed in anecdotes the non-governmental organization (NGO) collects in the field:

> At first I joined [the organization] only to play sports and I did not know that it included Jews as well in the practices.
>
> (Interview with an Arab-Palestinian participant, 2018)[2]

[1]The sport activities are sometimes complemented with a peace curriculum, i.e., lessons around recognizing humanity of others. These lessons were rare and sporadic for the cohort we studied.

[2]These quotes were selected from a database gathered by our partner organization.

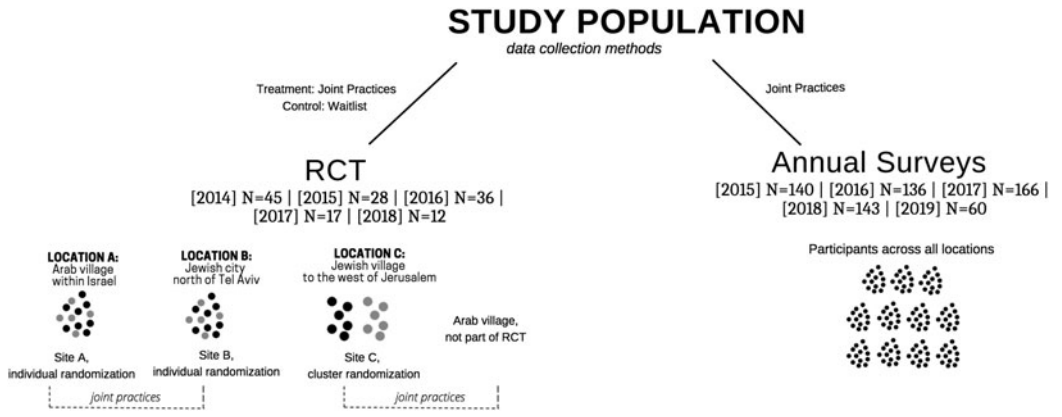


Figure 1. Research design.

The teams practice weekly with their ingroup. After about a month, joint practices between an Arab-Palestinian team and a Jewish-Israeli team begin. These joint practices involve sports drills, team-building exercises, and games between ethnically mixed (single-gender) teams. They take place approximately five to eight times throughout the season (fall to summer). In our RCT sample, each team has their own partner team, but partnering teams are not part of our sample. Even the rather modest contact between Arab-Palestinian and Jewish-Israeli youth in the form of joint practices represents a dramatic departure from the typical experience of youth in Israel, who tend to live in highly segregated communities (Hammack, 2011; Shwed *et al.*, 2018). As described by one participant:

> We can live five minutes from each other, and never meet or talk. [Name], who is a Palestinian participant in the program, she lives five minutes away from me but I would have never had the chance to meet her if it wasn't for the program. (Interview with a 14-year-old Jewish-Israeli participant during an intergroup meet-up session; June 2020)

Our partner organization aims to implement the optimal conditions for intergroup contact (Allport, 1954). Team sport provides a common goal (during joint practices, teams are always ethnically mixed) and requires cooperation. To achieve status equality within the program, coaches and other leadership positions are distributed equally among Jewish-Israelis and Arab-Palestinians, while teams that practice together are matched on age, gender, and athletic skills. Coaches and program leaders encourage and model peaceful intergroup relations, thus satisfying the fourth of Allport's original conditions for optimal contact, i.e., support from relevant authorities.

To ensure ecological validity (i.e., study reflecting the complexity and dynamics of the real world), we integrated our research strategy as much as possible into the existing program. The capacities and resources of an NGO working in an extremely politicized context are already stretched and program implementation depends on the cooperation of diverse actors (teachers, coaches, communities for practice space, etc.) Our efforts not to disrupt this difficult-to-achieve balance sometimes came at a cost to our research design (e.g., smaller RCT sample), yet our newly proposed fusion design aims to overcome this challenge.

5. Research design

Our research design (Figure 1) includes an RCT ($N = 138$) component that allows us to estimate one-year effects, and then a component that fuses the RCT data with data from a general survey

($N = 645$) to estimate how effects accumulated over one, two, and up to three years of program participation.

5.1 RCT design

For the RCT, our implementing partner recruited participants across grades four to six (typically ages 10–13) within three research sites (locations A, B, and C within Figure 1). The three sites for the RCT were not randomly selected. For this reason, in the analysis that fuses the RCT and survey data, we control for covariates that might be out of balance across the sites.

In a manner that is consistent with the way that the program generally operates, teachers at schools and community centers in the RCT sites publicized the program and collected applications. Our partner leads with sports in recruiting which reduces concerns over how selection into our study population affects generalizability. Even if there was outcome-relevant selection into our sample (e.g., participants' parents being particularly open minded) an average effect of the treatment on the treated is still a reasonable estimand. Were this or a similar extracurricular program to be scaled up, youths would still have to opt into it, and hence the effects we estimate are for the subpopulation of youths who would plausibly choose to participate in the program. Once recruited into the RCT, applicants were randomly assigned to the program (treatment group) or put on a waiting list that would permit access in a subsequent season (control group). Youth in the treatment group were placed on the sports teams that then went through the season-long curriculum described earlier.

The teams formed in the RCT sites included youth who were granted access through the RCT random assignment, as well as youth outside the RCT participant pool. The latter came through channels that could not be randomized and were not included in the research (e.g., school programs, younger age cohorts from the same organization that advanced to the next level). As a result, each year only a few slots in the teams were available for randomization. In each location, we filled these slots over five seasons (2014–2018, see Figure 1), with new cohorts of RCT participants. Control group participants were often admitted into the program after one season for ethical reasons, which is why we could not track RCT cohorts beyond one season. Nonetheless, our non-experimental survey data, which we incorporate using a method discussed below, allow us to make predictions about the effects of multiyear exposure under three assumptions described in the next section.

In locations A and B, treatment assigned was at the individual level. In location C, assignment occurred at the 12-person group level, because the available participants were already grouped into teams by a cooperating sports club that could not be split up. Ensuring consistency with this randomization protocol, our analysis below accounts for potential clustering among the group-assigned participants in location C (Gerber and Green, 2012). Table 1 shows compliance rates and attrition. Out of the youth assigned to the control condition, nine did not comply. This happened because coaches or teachers had a strong preference that certain children should be on a team, either because the child would benefit the team a lot (typically due to strong athletic skills) or the child would benefit a lot. For ethical reasons, we did not intervene to prevent such crossovers. We treat those who did not comply with the treatment as non-compliers, and those for whom we were not able to obtain outcome data as attrition. We conducted endline surveys with RCT-treated and control youth at the end of the season, typically about seven months after assignment.

Table 1. Treatment assignment, compliance, and attrition

	Control	Treatment	Missing
Assigned to control	58	9	4
Assigned to treatment	11	60	12

Table 2. RCT sample descriptive statistics

Treatment group assigned						Control group assigned						Comparison
Covariate	Mean	St. Dev	Min	Max	Obs	Covariate	Mean	S.D.	Min	Max	Obs	Std. mean diff.
Age	10.540	0.859	8	12	71	Age	10.730	0.931	8	13	67	0.219
Female	0.507	0.504	0	1	71	Female	0.522	0.503	0	1	67	0.030
Ethnicity: Arab-Palestinian	0.324	0.471	0	1	71	Ethnicity: Arab-Palestinian	0.299	0.461	0	1	67	0.055
Family religiosity						Family religiosity						
Secular	0.155	0.364	0	1	71	Secular	0.149	0.359	0	1	67	0.016
Traditional	0.676	0.471	0	1	71	Traditional	0.716	0.454	0	1	67	0.087
Very religious	0.169	0.377	0	1	71	Very religious	0.134	0.344	0	1	67	0.096

Table 2 shows the demographic composition of our RCT sample. While the distribution of basic demographic attributes is balanced between those assigned to treatment and control groups, the overall sample exhibits important idiosyncrasies (Appendix, Table 1). All Arab-Palestinian participants in our sample are female, largely a consequence of organization's deliberate efforts to improve Arab-Palestinian female engagement in sports. In our analyses of the RCT, as per our pre-analysis plan, we control for gender and location, which is perfectly predictive of ethnicity.

5.2 General survey and fusion design

Each year we also carried out a general cross-sectional survey (Table 3) with participants from the broader population of program sites from which the RCT sites were selected. The general survey functions as a census of participants at sites with programming comparable to the RCT. The aim of the general survey was to assess how program exposure relates to outcomes beyond the first year, and it uses largely the same measures as the RCT surveys. The larger sample size also gave us information on heterogeneity across Arab-Palestinian and Jewish-Israeli youth.

We "fuse" the RCT with observational data to estimate longer-term treatment effects, using a machine-learning method (random forest, Breiman, 2001). Our approach is most similar to that of Athey *et al.* (2020) and Imbens *et al.* (2022), who show how to combine RCT and observational samples to estimate effects that extend beyond the timeline of the RCT. The basic idea is to model the long-term outcomes in the observational data and then use that model to impute long-term outcomes for the RCT subjects. Athey *et al.* (2020) and Imbens *et al.* (2022) assume that the observational data have a panel structure, such that both short- and long-term outcomes are measured for each subject. Our situation is more restrictive in that our observational dataset consists of repeated cross sections. Thus, we cannot impute long-term outcomes for our RCT subjects using a model that relates outcome variables to each other over time. Rather, our imputations rely only on covariates. This is the key difference between our approach and that of Athey

Table 3. Survey sample descriptive statistics

Arab-Palestinian participants						Jewish-Israeli participants					
Covariate	Mean	S.D.	Min	Max	Obs	Covariate	Mean	S.D.	Min	Max	Obs
Age	11.530	1.794	8	16	342	Age	11.930	1.999	9	16	299
Female	0.658	0.475	0	1	342	Female	0.766	0.424	0	1	299
Family religiosity						Family religiosity					
Secular	0.006	0.076	0	1	342	Secular	0.194	0.396	0	1	299
Traditional	0.921	0.270	0	1	342	Traditional	0.736	0.442	0	1	299
Very religious	0.070	0.256	0	1	342	Very religious	0.070	0.256	0	1	299

et al. (2020) and Imbens *et al.* (2022). We use a flexible random forest model as the basis of the imputations. As we describe below, we use a conservative bounding approach to address possibly endogenous attrition in the observational data. Our approach could be helpful for civil-society programs that lack the resources to collect longitudinal data, although we note that the conditions for causal identification are stronger than Athey *et al.* (2020) and Imbens *et al.* (2022) since our observational data do not contain direct information on how individuals' outcomes evolve over time.

The longer-term effect that we target is the effect of continued participation in the program on individual's outcome trajectories as they progress through 10, 11, and 12 years of age. If youths remain in the program, then for each year they get older, they also accumulate another year in the program. In the RCT, we have youths of mixed ages that are either exposed to the program for one year or not at all; in the general survey, we have youths of mixed ages that are all exposed to the program, some for one year, others for two or three years. The fusion combines the information from the general survey and the RCT to make predictions about what would happen to RCT participants if they stayed in the program from the time they are 10 to the time they are 12, as compared to how their outcomes would evolve during those years without being in the program. We focus attention on 10–12 year olds because they are the majority participants in our RCT sample.[3] We use the covariate profile of the ten year olds to define a reference population. The fusion analysis was also specified in our pre-analysis plan.

More formally, let $Y_p(t)$ denote an individual's outcome p periods since the onset of treatment eligibility given assignment to treatment condition t. The difference in outcome trends while under the program as compared to never being in the program is given by the following sequence:

$$E[Y_1(1) \mid \text{Age} = 10] - E[Y_1(0) \mid \text{Age} = 10]$$

$$E[Y_2(1) \mid \text{Age} = 11] - E[Y_2(0) \mid \text{Age} = 11]$$

$$E[Y_3(1) \mid \text{Age} = 12] - E[Y_3(0) \mid \text{Age} = 12],$$

where the expectations ($E[\,\cdot\,]$) are with respect to our reference population (the ten-year-old RCT participants). Age and time in the program are perfectly collinear here, and so time-in-program and age-specific program effects cannot be disentangled. In the Appendix (Section 5), we show that there is no appreciable interaction effect with age in the RCT, which provides indirect evidence that the fusion analysis is picking up on effects from accumulated exposure. In any case, the analysis is meaningful in characterizing how trends would differ depending on whether a person participated in the program from age 10 to 12.

A model of the data-generating process (DGP) and identifying assumptions are needed to bring all of this information together to identify the trend effect. Figure 2 illustrates the DGP that justifies our approach. The variables X and U denote observed and unobserved, respectively, background characteristics of individuals. The variable T is for whether a person is assigned to treatment or control. The variables Y_1 and Y_2 are outcomes in a first period and a subsequent period, respectively, and S is an indicator for whether an individual stays in the program in the subsequent period. We use Y_2 and S to characterize potential attrition biases, as described below.

The first assumption that is implicit under this graph is that the treatment operates the same way regardless of whether the site is an RCT site or a general program site. That is, we assume that an RCT participant's potential outcome under treatment would have been the same if, counterfactually, their site was just part of the general program. All of our control observations come from RCT sites. The second assumption from the graph is that selection into an RCT site

[3]Making similar predictions for outcomes from 11–13 is more difficult because we have less information about 13 year olds.

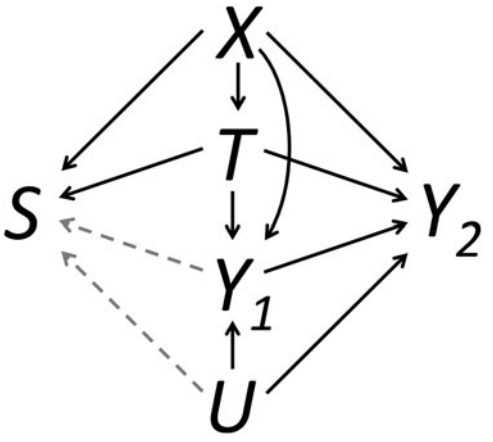


Figure 2. Directed acyclic graph showing the DGP for treatment assignment, treatment effects in year one, selection into programming in year two, and treatment effects in year two.

depends, systematically, only on observed characteristics (X). We justify these assumptions from the fact that RCT sites were selected for incidental reasons: they were sites where new cohorts were already planned on the basis of considerations that had been used to initiate programming more generally. At the same time, given the clustered nature of the selection into an RCT site (even though treatment assignment within RCT sites was mostly at the individual level), we expect some incidental imbalances in background characteristics across RCT versus general program sites. Hence, we control for background characteristics, which we chose based on past research and the suggestions of the implementing organization. These covariates are the program year and individuals' gender, ethnicity, region of residence, religiosity, parents' location of birth, and parents' occupations.

Another key assumption is that, conditional on X, individuals in the different RCT control group cohorts can be used to construct the control trend ($E[Y_1(0)|\text{Age} = 10]$, $E[Y_2(0)|\text{Age} = 11]$, $E[Y_3(0)|\text{Age} = 12]$). Consider the ten year olds that are in the control group in year one. Ideally we would observe their outcomes under control in years two and three to construct their cohort-specific control trend. However, we agreed with our implementation partner that individuals in the control group would have the opportunity to participate in subsequent years. Thus, we assume that, conditional on covariates, we can estimate this cohort-specific control trend by using outcomes from the 11-year-old control group members in year two, and 12-year-old control group members in year three.

Given these assumptions, for year one outcomes, we have that $(Y_1(1),\ Y_1(0)) \perp T|\ X$, in which case,

$$
\begin{aligned}
&E[Y_1(1) \mid \text{Age} = 10] - E[Y_1(0) \mid \text{Age} = 10] \\
&\quad = \int E[Y_1 \mid T = 1, \text{Age} = 10, X = x] dP(x \mid \text{Age} = 10) \\
&\quad\quad - \int E[Y_1 \mid T = 0, \text{Age} = 10, X = x] dP(x \mid \text{Age} = 10),
\end{aligned}
$$

where $E[Y_1|\ T = 1,\ \text{Age} = 10,\ X = x]$ can be estimated from the combined RCT treatment group and general survey data, $E[Y_1|\ T = 0,\ \text{Age} = 10,\ X = x]$ can be estimated from the RCT control group data, and the covariate distribution $P(x|\text{Age} = 10)$ is derived from the RCT treatment and control group data.

Estimating effects beyond one year are complicated by the fact that some individuals drop out of the program each year, and this could occur for reasons that also affect outcomes in subsequent years. Figure 2 illustrates this possibility for outcomes in year two. (The same logic would apply for outcomes in year three.) We observe second-year-treated outcomes conditional on S, and so if we apply the method that we used for year one outcomes for year two outcomes, we would be estimating:

$$E[Y_2(1) \mid \text{Age} = 11, S = 1] - E[Y_2(0) \mid \text{Age} = 11],$$

where the outcomes for the "no drop-out" types $E[Y_2(1)|\text{Age} = 11,\ S = 1]$ could differ from the overall mean $E[Y_2(1)|\text{Age} = 11]$ because S is a "collider" variable (Elwert and Winship, 2014). Now, if drop-out is endogenous only to T and X, then the approach that we used for year one outcomes would also work for year two (and year three) results, because the conditioning on T and X would remove any confounding. In Figure 2, this would imply that the gray-dashed arrows are not active, and our point estimates would be valid. However, if drop-out is endogenous to Y_1 or U as well as T and X (i.e., the gray-dashed arrows are active), then our approach would not identify the relevant effects. Because our general survey data are repeated cross sections (administrative constraints and privacy concerns made a panel survey impossible), we cannot control for any Y_1 variables. Even if we could, there may be unmeasured outcomes or background characteristics that confound the analysis.

From our discussions with program administrators, most of the year-to-year dropout was attributed to exogenous logistical factors, rather than participants deciding to leave because of a bad experience. Nonetheless, to be fully transparent, we use the trimming bounds of Lee (2009) to characterize "worst-case" consequences of endogenous attrition. We focus on an effect that pertains to "no drop-out" types in each period. Ideally, to estimate an effect for such "no drop-out" types, we would use an estimate of $E[Y_2(0)|\text{Age} = 10,\ S = 1]$. However, what we can estimate with our RCT control group data is

$$\begin{aligned} E[Y_2(0) \mid \text{Age} = 11] = {} & E[Y_2(0) \mid \text{Age} = 11, S = 1]P(S = 1 \mid \text{Age} = 11) \\ & + E[Y_2(0) \mid \text{Age} = 11, S = 0]P(S = 0 \mid \text{Age} = 11), \end{aligned}$$

which we can see is based on a mixture of individuals for which S would be one or zero were it that they, counterfactually, were in the general program. Considering that we do not have outcome data for survey participants that dropped out of the program after year one, what we can do is to estimate Lee bounds on $E[Y_2(0)|\text{Age} = 11,\ S = 1]$ by trimming the upper and lower tails of the control group distribution by the retention rate $P(S = 1|\text{Age} = 11)$. We estimate this rate using the general survey data. We also follow Lee (2009) in using covariates to tighten the bounds. The details are explained in the Appendix (Section 12).

5.3 Outcome measures

Corresponding to our analytical framework, we measure the effects of the intervention on outgroup regard and ingroup regulation. We operationalize them with several indicators, each consisting of batteries of items (Table 4). We use youth-appropriate and validated measures of outgroup regard, as well as create new ingroup regulation measures building on the existing literature: tendency for ingroup censuring, ingroup policing, peer persuasion, and perspective sharing. Cross-sectional survey questions are largely identical to RCT questions, although they also included descriptive measures capturing program experience and additional operationalizations of our outcomes. Survey measures were identical for all participants irrespective of their ethnic membership with the exception of the measure of perspective sharing tendencies, since the perspective each group might share differs.

Table 4. RCT and annual surveys measures

Concept	Indicator	Operationalization
Outgroup regard	Social distance	5-items social distance index (willingness to meet outgroup friends, invite an outgroup member to one's house, study in the same school as the outgroup, live in the same neighborhood, play sports or do other activities with the outgroup); 5-point scale
Outgroup regard	Support for peace process	"Do you support the peace process between Israelis and Palestinians?"; 5-point scale
Outgroup regard	Interest in perspective taking	Question about the interest in seeing more images and narratives by/about the outgroup with friends. Image in the Appendix (Figure 6); narrative examples below: *"Shadi (16) is an Arab-Palestinian boy from a City of Bethlehem in the West Bank. Every time Shadi comes back from school, he is afraid that his house is no longer there."* OR *"Noam (15) is a boy from a Jewish family in the suburbs of Tel Aviv. Every time he has to take the public bus, he is afraid for his safety".*
Outgroup regard	Hostile attribution by subject	Asking respondents to interpret an ambiguous image where outgroup member could be doing something harmful or helpful to ingroup member; e.g., *("What is X doing?")*, *("How good/bad the action is?")*, *("How likely it is that the two are friends?")*
Outgroup regard	Hostile attribution by peers	Ambiguous images questions: After being shown illustrations of ten kids, indicating these are ingroup and non-PPI peers, participants asked to report how many kids were frowning versus smiling.
Outgroup regard	Optimism about peace	*"Do you think there will be peace between Israelis and Palestinians in the next few years?"* 5-point scale
Outgroup regard	Ingroup identity esteem	8-item index with 7-point scales, capturing one's esteem for Arab/Jewish group identity
Ingroup regulation	Effort to persuade	Participants asked to attempt to persuade ingroup members of their interpretation of the event in the ambiguous images (effort measure is the number of characters in each participant's explanation)
Ingroup regulation	Ingroup censuring and policing tendency	In a vignette in which an ingroup member commits aggression toward an outgroup member, willingness to censure the ingroup behavior and support efforts to stop the aggression by an ingroup member; 7-point scale (survey 2015–2016)
Ingroup regulation	Perspective sharing tendency	Narrative as in the measure of perspective taking above; image in the Appendix (Figure 6); *"Would it be a good or a bad* idea for you to share this with your friends?"

Given the contentiousness of the political situation and the fact that all participants in our sample are minors, measuring actual behavior on our outcomes of interest is particularly challenging. Our measures thus capture self-reported behavioral intentions as proxies for the actual behavior. We form the outgroup regard index using principal component analysis because we consider the seven different indicators to be different operationalizations of the same underlying construct. Our approach for the ingroup regulation index is different, because we are interested in a series of qualitatively different strategies that participants can employ. As such, the index of Ingroup Regulation is formed as a sum of Z-scores, treating all indicators equally.

6. Data analysis

To test the hypotheses about short-term effects with the RCT data, we estimate intention to treat (ITT) effects using a linear regression specification as per our pre-analysis plan (see Appendix, Section 5). To estimate effects of multiyear exposure with the fusion design, we carry out the following steps:

(1) Use the RCT control group to model control outcome trends $E[Y_p|T=0,\ \text{Age}=a,\ D=0,\ X=x]$ for $(p,\ a)\in\{(1,\ 10),\ (2,\ 11),\ (3,\ 12)\}$ and the covariates (x) listed above (year, gender, ethnicity, region of residence, religiosity, parents' location of birth, and parents' occupations).[4] Modeling these outcome trends improves precision relative to just using the age-specific averages from the control group data, and also allows us to predict control trends for the same fixed reference population that we used for predicting treatment trends.
(2) Use the RCT treatment group and survey data to model outcome trends $E[Y_p|T=1,\ \text{Age}=a,\ D=0,\ X=x]$ for $(p,\ a)\in\{(1,\ 10),\ (2,\ 11),\ (3,\ 12)\}$ and the same covariates, together with an indicator for whether the observation is from the RCT or survey.
(3) Construct the weighted averages (the $\int(\cdot)dP(x\mid \text{Age}=10)$ operation) for the reference population: ten-year-old RCT participants. For each ten-year-old RCT participant, we take their covariates and generate a prediction for their expected outcome under treatment and control at age 10, 11, and 12. We take the average of all of these predictions to obtain the estimated trajectory for "All" participants. We take the averages of predictions for the Arab-Palestinian girls, Jewish-Israeli boys, and Jewish-Israeli girls to get their respective group-specific estimated trends.
(4) Construct the Lee (2009) trimming bounds based on an estimates of $P(S=1|\text{Age}=10,\ X=x)$.
(5) Bootstrap the entire process to obtain the confidence intervals. (The interpretation of the individual bootstrap here is that we are accounting for those who join the program being a sample from a broader population of youth who might participate each year.)

We use a flexible random forest model based on the bartMachine package in R (Kapelner and Bleich, 2016) to estimate the $E[Y_p|T=t,\ \text{Age}=a,\ D=0,\ X=x]$ quantities. Our data are too sparse to do this by using a model-free, matching approach. This particular non-parametric modeling approach has proven highly effective in recent applied work, including in political science (Hill, 2011; Green and Kern, 2012; Kern *et al.*, 2016; Samii *et al.*, 2017). As Athey *et al.* (2019) show, the random forest model can be understood as deriving predictions as kernel-weighted averages, where the kernel weights are optimized so as to capture potentially complex functional relationships between covariates and outcomes without overfitting.

7. Results

In this section, we test and discuss our hypotheses about the program's impact on outgroup regard (H1) and on ingroup regulation (H2) separately, after which we integrate findings in the conclusion. For each of the two outcomes, we test the impact of the program in a one-year RCT (hypotheses: H1a, H2a),[5] after which we proceed by testing the impact of a multiyear exposure to the program (H1b, H2b) across all participants and then subgroups with our fusion

[4] D is a binary indicator of treatment received.
[5] We do not present pre-registered heterogeneous treatment effects from the RCT because of the small RCT sample size.

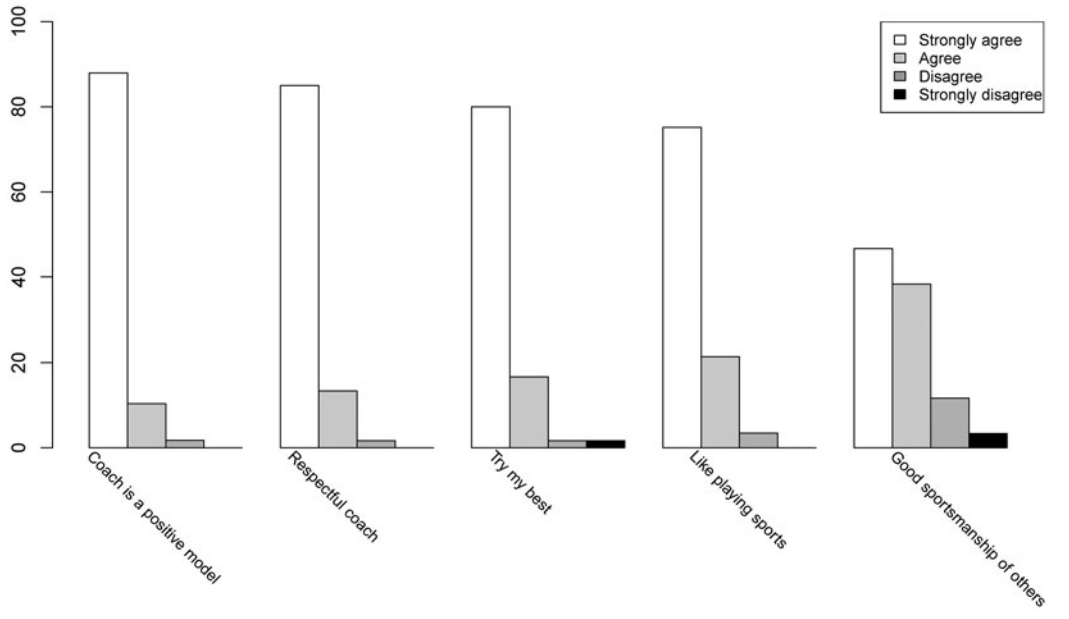


Figure 3. Program experience, as captured through annual surveys. The question about how much participants like playing sports comes from the survey with 435 participants (2015–2017). The other four questions were measured on 60 participants in 2019 and capture participants' agreement with the following statements: (a) coach is a positive model; (b) coach is respectful; (c) I try my best in practices; (d) other players show good sportsmanship.

analysis. We adjust for multiple comparisons controlling for false-discovery rate following Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995), and report both adjusted and unadjusted results. Prior to presenting the results, we describe why participants join the program and how they experience it.

7.1 Descriptive results

Annual surveys show that respondents enjoyed participating in the program, ranking their interest in the sports activity, experiences with coaches, and the integrated practices very highly (Figure 3). The program also appears to fulfill an important condition for optimal intergroup contact: friendship potential. In total, 63 percent of participants made an outgroup friend in the program, a remarkably high number in a deeply segregated context. Friendship formation is significantly positively associated with the number of years in the program across both groups, with the effect being particularly strong for Jewish-Israeli participants. These positive descriptive results give us confidence that the program is implemented well.

7.2 Outgroup regard: RCT results

Hypothesis H1a stipulates that participating in a one-year intergroup contact program within a conflict setting increases outgroup regard. In Table 5, we present the results from the ITT analysis for all outgroup regard outcomes, pooling the data across members of both ethnic groups and controlling for covariates as specified in Section 6. With pooled data, we find no evidence to affirm a one-year effect on outgroup regard, with all adjusted p-values greater than 0.10.

Table 5. ITT effect on the index of outgroup regard (OR) and corresponding indicators

	OR index	Soc. Dist.	Sup. Peace	Persp. Tk.	Host. Att. Self	Host. Att. Peer	Opti. Peace	Ingroup Ident.
Program effect	−0.09	0.34	0.11	−0.25	−0.38	−0.26	−0.06	−0.51
(S.E.)	(0.22)	(0.74)	(0.17)	(0.23)	(0.25)	(0.34)	(0.19)	(0.26)
[p]	[0.67]	[0.65]	[0.51]	[0.27]	[0.13]	[0.45]	[0.74]	[0.06]
[p FDR]	[0.67]	[0.74]	[0.72]	[0.62]	[0.44]	[0.72]	[0.74]	[0.41]
Control mean	0.04	17.99	4.03	0.13	0.27	5.21	3.03	0.24
(Control S.D.)	(1.59)	(5.14)	(1.27)	(1.42)	(1.65)	(2.19)	(1.17)	(1.64)
N	138	138	138	138	138	138	138	138

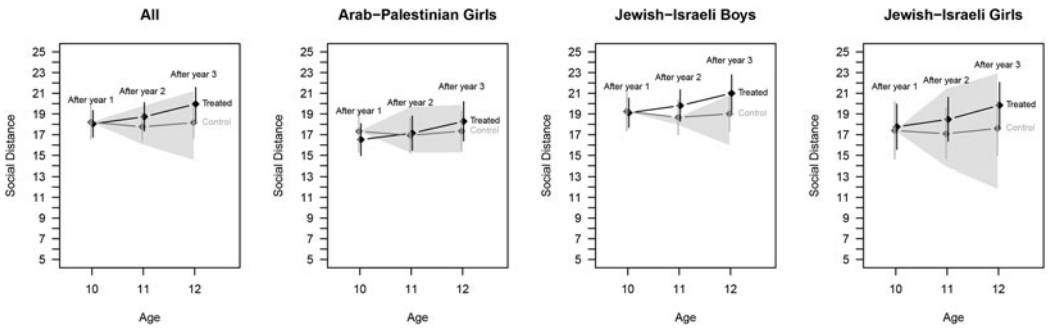


Figure 4. Fusion analyses for social distance within the three subgroups (numerical estimates in the Appendix, Section 5.E). We show point estimates for the treatment and control groups after one, two, and three years following possible program initiation, along with 95 percent bootstrap confidence intervals. The gray-shaded area shows the range of possible values for the control group trend for "no drop-out" types given possibly endogenous drop-out.

7.3 Outgroup regard: fusion results

Hypothesis 1b stipulates that multiyear exposure to the program increases outgroup regard. We used social distance as our main indicator of outgroup regard in the fusion analysis, as it appears in both the RCT and the survey data. Results are presented in Figure 4. The graphs show point estimates for under treatment and control along with 95 percent bootstrap confidence intervals for these estimates. If drop-out is endogenous only to T and X, then the point estimates for years two and three are valid. If drop-out is endogenous also to previous period outcomes or unobservables, we use the trimming bounds to characterize the full range of possible point estimates for the latent "no drop-out" types in the control group in each period. The gray-shaded area shows the trimming bounds. The width of the bars depends on the drop-out rates after each year, which determines how much of outcome variable distribution will be trimmed to form the bounds. We can assess whether estimated effects are robust to program attrition by checking whether the point estimate for the treatment group is outside the gray-shaded area.

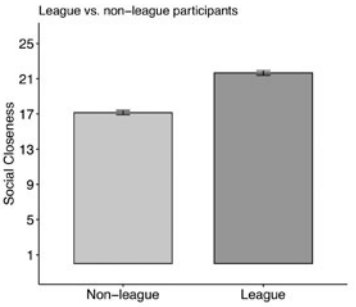
The figures show that the fusion analysis replicates the null effects for social distance that we observe in the RCT after the first year of the program. The fusion point estimates show a positive, though imprecisely estimated effect for participants who took part in the program for two (β = 0.98, S.E. = 0.95, p = 0.29) and three years (β = 1.8, S.E. = 1.09, p = 0.1). The positive effects were strongest for Jewish-Israeli boys who were part of the program for three years (β = 1.99, S.E. = 1.2, p = 0.08).

7.4 Outgroup regard: discussion

With our RCT results, we cannot affirm the hypothesis that the program increases outgroup regard in participants' first year. The fusion analyses suggest modest but positive effects after

Table 6. (a) Descriptive statistics and (b) social distance scores

	League	Non-League
N	176	469
Age (mean, SE)	13.44 (0.12)	11.07 (0.07)
Female (N, %)	169 (96%)	289 (62%)
Family religiosity (N, %)		
secular	11 (6.25%)	34 (7.25%)
traditional	138 (78.41%)	399 (85.07%)
religious	26 (14.77%)	36 (7.68%)



second and third years of participation for Jewish participants, with results that cannot be explained away by worst-case endogenous drop-out for boys. To further explore if the intensity of contact experience mattered, we also analyzed the results for a subgroup of survey respondents who participated in an interethnic league throughout the year (Table 6). These tend to be the most talented players, with the most intensive program experience. Figure (b) shows that these players indeed report significantly more positive social distance scores ($t = 12.44$, $p < 0.01$) compared to the rest of participants. These descriptive results corroborate the conclusion from the fusion results: multiyear, repeated exposure to the outgroup is needed for intergroup contact to have a positive, prejudice-reducing impact in conflict-ridden societies.

7.5 Ingroup peer regulation: RCT results

Hypothesis H2a stipulates that participating in a one-year intergroup contact program within a conflict setting increases one's engagement in ingroup regulation. In Table 7, we present the ingroup regulation results from the ITT analysis pooling across ethnic groups and controlling for covariates as specified within Section 6. With pooled data, we find no evidence of a one-year effect on inclination to regulate ingroup members for aggressions toward the outgroup.

7.6 Ingroup peer regulation: fusion results

We use the fusion analysis to test hypothesis 2b about multiyear exposure and hypothesis 2c about moderation by ethnicity. We find some significant results for perspective sharing tendencies, but not for ingroup censuring nor ingroup policing. Figure 5 presents the perspective sharing results (for null results for ingroup censuring and ingroup policing, see Appendix 10). Unlike for social distance, we did not assess perspective sharing with precisely the same instruments in the

Table 7. ITT effect of treatment on ingroup regulation (IR) index and corresponding indicators

	IR index	Effort to persuade	Ingroup censure	Ingroup police	Perspective sharing tendency
Program effect	−0.002	0.05	0.53	−0.39	−0.06
(S.E.)	(0.18)	(0.09)	(0.57)	(0.26)	(0.17)
[p]	[0.99]	[0.54]	[0.35]	[0.14]	[0.71]
[p FDR]	[0.99]	[0.71]	[0.71]	[0.55]	[0.71]
Control mean	−0.04	4.32	11.81	6.09	0.03
(Control S.D.)	(0.99)	(0.56)	(3.86)	(1.79)	(1.14)
N	138	138	138	138	138

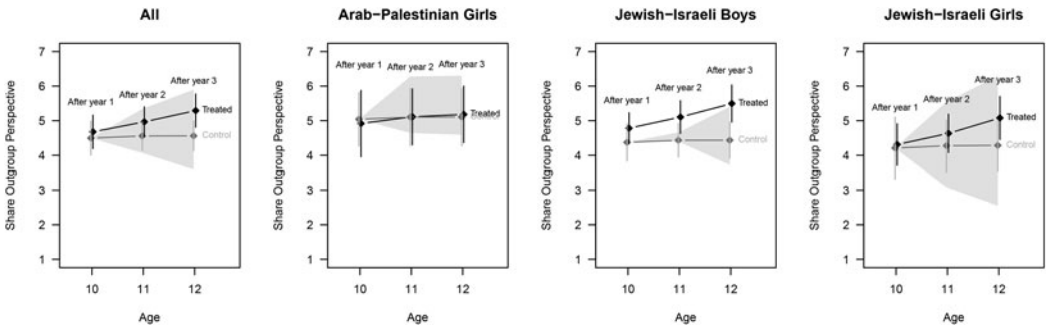


Figure 5. Fusion analyses for the tendency to share outgroup perspective within the three subgroups (numerical estimates in the Appendix, Section 5.F). We show point estimates for the treatment and control groups after one, two, and three years following possible program initiation, along with 95 percent bootstrap confidence intervals. The gray-shaded area shows the range of possible values for the control group trend for "no drop-out" types given possibly endogenous drop-out.

RCT and in the survey. Instead, for the fusion we selected a survey question that asks how often participants defended the outgroup's perspective in discussions with ingroup members. As for the social distance outcome, in year one the results mostly replicate the null effect from the RCT. The fusion point estimates show a positive effect for participants who took part in the program for two ($\beta = 0.41$, S.E. = 0.31, p = 0.19) and three years ($\beta = 0.73$, S.E. = 0.32, p = 0.02). The positive effects were strongest for Jewish-Israeli boys who were part of the program for three years ($\beta = 1.06$, S.E. = 0.36, $p < 0.01$).

7.7 Ingroup peer regulation: discussion

The RCT cannot affirm our hypothesis that the program increases ingroup regulation in participants' first year. But the fusion analyses suggest modest but positive effects for perspective sharing after second and third years of participation for Jewish participants, with results that cannot be explained away by worst-case endogenous drop-out for boys. As for social distance, we again explored results for league participants. Figure 6 shows that these players indeed report significantly more perspective sharing than their non-league peers. Again, this descriptive evidence corroborates the argument that the intensity of exposure to the outgroup matters.

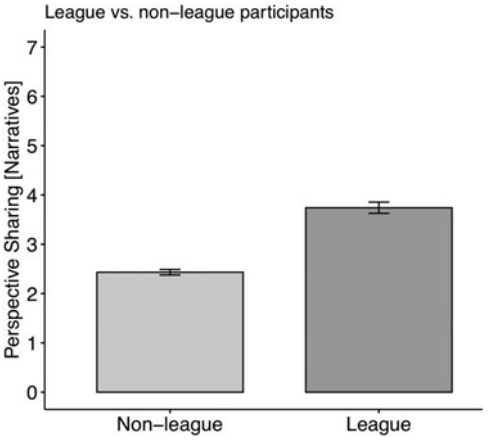


Figure 6. Narrative perspective sharing.

8. General discussion and conclusion

We set out to study how combining sports and intergroup contact between youths might contribute to peace in the context of the ongoing violent conflict in Israel. We formulated six hypotheses about the impact of one-year and multiyear intergroup contact on outgroup regard and ingroup regulation, and of the heterogeneity by ethnic group. We find no significant effects of one-year contact. The analysis of multiyear contact suggests that the effects of the program on outgroup regard and perspective sharing become more pronounced over time in a way that clearly distinguishes Jewish-Israeli from Arab-Palestinian participants. To our surprise, the fusion results suggest heterogeneity beyond ethnicity—by gender. These results raise interesting questions about how the effects of contact interventions may depend on one's understanding of one's own ethnicity, as filtered through the experience of one's other group memberships (Ghavami *et al.*, 2020). While an association between intergroup contact and prejudice reduction appears to have been well-established in the survey and lab-based literature, our field-experimental evidence is quite mixed, in a manner that is similar to other field-experimental work in conflict settings by Scacco and Warren (2018) and Mousa (2020). Designing effective interventions in conflict settings on the basis of the logic of contact theory is neither trivial nor certain to generate consistently strong effects. Research in field settings is the best way to learn more about this challenge.

Two key insights emerge from our research for the design of contact interventions. First, for intergroup contact in a conflict setting to have positive benefits, one year (with an average of eight sessions of playing in ethnically mixed teams) may not be enough. The results of our fusion analysis—which rely on critical assumptions that we believe are plausible in our context but may not always be met—suggest modest yet positive effects associated with two to three years of program exposure. Positive results for participants who took part in the more intensive league program provide further suggestive evidence of the importance of multiyear exposure. This insight is consistent with recent scholarship proposing that people have to receive a certain dosage of intergroup contact for it to yield stronger positive effects. The findings on length of exposure raise two questions that ought to be the focus of further research. First, what dosage of contact is minimally needed and how may that differ for different groups of participants? This question is crucial for practitioners who want to set up intergroup contact programs in conflict settings: if they fail to reach the threshold, they may do more harm than good. Second, what exactly happens in high-intensity contact that ultimately brings about the positive effects? MacInnis and Page-Gould (2015) propose several processes; the one most consistent with our data is that more intense exposure to the outgroup has greater potential for bonding and friendship formation. Future research should more systematically test the processes that occur during high-intensity intergroup contact over time.

The second key insight from our research is that even a well-designed and well-received intergroup contact program may not achieve positive results for all groups. Initially insignificant program effects start intensifying more clearly over time for Jewish-Israeli participants, and in particular, Jewish-Israeli boys. In contrast, we never observe any impact of the program on Arab-Palestinian participants (in our sample, all girls). This may not be surprising in light of recent scholarship questioning the benefits of joint social activities for members of minority groups (Hässler *et al.*, 2022). Programs that want to serve multiple communities may have to structure their program differently (e.g., including dialogues about inequalities, Saguy *et al.*, 2009) or offer alternative benefits to the minority group (e.g., career opportunities as our partner organization does). Our partner organization believes that these conversations happen over time after trust has been built—results from league participants are consistent with that hypothesis which future research should test.

Finally, we illustrated a strategy for overcoming limitations of many RCTs: small samples and short follow-up times. We combined the RCT with a cross-sectional survey and machine-learning methods, and used a highly agnostic bounds approach to address attrition. This analysis, contingent upon several critical assumptions, provides a template for those seeking to capitalize on opportunities to study innovative interventions in the field. It allows us to find that intergroup

contact within a conflict setting is most successful when it offers a multiyear, intensive experience that takes group asymmetries into account. This demonstrates that—even after 70 years of research on the topic—the design, implementation, and evaluation of intergroup contact interventions in conflict-ridden society remain challenging.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2024.8. To obtain replication material for this article, https://doi.org/10.7910/DVN/M3DXXJ

References

Allport GW (1954) *The Nature of Prejudice*. Reading MA: Addison-Wesley.

Athey S, Tibshirani J and Wager S (2019) Generalized random forests. *Annals of Statistics* **47**, 1148–1178.

Athey S, Chetty R and Imbens G (2020) Combining experimental and observational data to estimate treatment effects on long-term outcomes. Preprint arXiv:2006.09676.

Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300.

Breiman L (2001) Random forests. *Machine Learning* **45**, 5–32.

Bruneau EG and Saxe R (2012) The power of being heard: the benefits of "perspective-giving" in the context of intergroup conflict. *Journal of Experimental Social Psychology* **48**, 855–866.

Colnet B, Mayer I, Chen G, Dieng A, Li R, Varoquaux G, Vert J-P, Josse J and Yang S (2020) Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science* **39**, 165–191.

Degtiar I and Rose S (2023) A review of generalizability and transportability. *Annual Review of Statistics and Its Application* **10**, 501–524.

Ditlmann R and Samii C (2016) Can intergroup contact affect ingroup dynamics? Insights from a field study with Jewish and Arab-Palestinian youth in Israel. *Peace and Conflict: Journal of Peace Psychology* **22**, 380.

Ditlmann R, Samii C and Zeitzoff T (2017) Addressing violent intergroup conflict from the bottom up?. *Social Issues and Policy Review* **11**, 38–77.

Dovidio JF, Gaertner SL and Saguy T (2009) Commonality and the complexity of "we": social attitudes and social change. *Personality and Social Psychology Review* **13**, 3–20.

Elwert F and Winship C (2014) Endogenous selection bias: the problem of conditioning on a collider variable. *Annual Review of Sociology* **40**, 31–53.

Enos RD (2017) *The Space Between Us: Social Geography and Politics*. Cambridge, UK: Cambridge University Press.

Fearon JD and Laitin DD (1996) Explaining interethnic cooperation. *American Political Science Review* **90**, 715–735.

Galily Y, Leitner MJ and Shimion P (2013b) The effects of three Israeli sports programs on attitudes of Arabs and Jews toward one another. *Journal of Aggression, Conflict and Peace Research* **5**, 243–258.

Gerber AS and Green DP (2012) *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: WW Norton.

Ghavami N, Kogachi K and Graham S (2020) How racial/ethnic diversity in urban schools shapes intergroup relations and well-being: unpacking intersectionality and multiple identities perspectives. *Frontiers in Psychology* **11**, 3133.

Green DP and Kern HL (2012) Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly* **76**, 491–511.

Hammack PL (2006) Identity, conflict, and coexistence: life stories of Israeli and Palestinian adolescents. *Journal of Adolescent Research* **21**, 323–369.

Hammack PL (2011) *Narrative and the Politics of Identity: The Cultural Psychology of Israeli and Palestinian Youth*. New York, NY: Oxford University Press.

Hässler T, Ullrich J, Sebben S, Shnabel N, Bernardino M, Valdenegro D, Van Laar C, González R, Visintin EP, Tropp LR, Ditlmann RK, Abrams D, Aydin AL, Pereira A, Selvanathan HP, von Zimmermann J, Lantos NA, Sainz M, Glenz A, Kende A, Oberpfalzerová H, Bilewicz M, Branković M, Noor M, Pasek MH, Wright SC, Žeželj I, Kuzawinska O, Maloku E, Otten S, Gul P, Bareket O, Corkalo Biruski D, Mugnol-Ugarte L, Osin E, Baiocco R, Cook JE, Dawood M, Droogendyk L, Loyo AH, Jelić M, Kelmendi K and Pistella J (2022) Need satisfaction in intergroup contact: a multinational study of pathways toward social change. *Journal of Personality and Social Psychology* **122**, 634.

Hill JL (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 217–240.

Imbens G, Kallus N, Mao X and Wang Y (2022) Long-term causal inference under persistent confounding via data combination. Preprint arXiv:2202.07234.

Kahanoff M (2016) *Jews and Arabs in Israel Encountering Their Identities: Transformations in Dialogue*. Lanham, MD: Lexington Books.

Kallus N and Mao X (2020) On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. Preprint arXiv:2003.12408.

Kapelner A and Bleich J (2016) bartMachine: machine learning with Bayesian additive regression trees. *Journal of Statistical Software* **70**, 1–40.

Kern HL, Stuart EA, Hill J and Green DP (2016) Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness* **9**, 103–127.

Künzel SR, Sekhon JS, Bickel PJ and Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* **116**, 4156–4165.

Lee DS (2009) Training, wages, and sample selection: estimating sharp bounds on treatment effects. *The Review of Economic Studies* **76**, 1071–1102.

Lemmer G and Wagner U (2015) Can we really reduce ethnic prejudice outside the lab? A meta-analysis of direct and indirect contact interventions. *European Journal of Social Psychology* **45**, 152–168.

Lerner RM and Steinberg L (2009) *Handbook of Adolescent Psychology, Volume 1: Individual Bases of Adolescent Development*. Hoboken, NJ: John Wiley & Sons.

Li H, Miao W, Cai Z, Liu X, Zhang T, Xue F and Geng Z (2020) Causal data fusion methods using summary-level statistics for a continuous outcome. *Statistics in Medicine* **39**, 1054–1067.

Litvak-Hirsch T, Galily Y and Leitner M (2018) Evaluating conflict mitigation and health improvement through soccer: a two-year study of Mifalot's "United Soccer for Peace" programme. In Parnell D and Pringle A (eds), *Football and Health Improvement: an Emergent Field*. London, UK: Routledge, pp. 47–62.

Lowe M (2021) Types of contact: a field experiment on collaborative and adversarial caste integration. *American Economic Review* **111**, 1807–1844.

MacInnis CC and Page-Gould E (2015) How can intergroup interaction be bad if intergroup contact is good? Exploring and reconciling an apparent paradox in the science of intergroup relations. *Perspectives on Psychological Science* **10**, 307–327.

McGlothlin H and Killen M (2006) Intergroup attitudes of European American children attending ethnically homogeneous schools. *Child Development* **77**, 1375–1386.

Mendes WB, Blascovich J, Hunter SB, Lickel B and Jost JT (2007) Threatened by the unexpected: physiological responses during social interactions with expectancy-violating partners. *Journal of Personality and Social Psychology* **92**, 698.

Mousa S (2020) Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq. *Science* **369**, 866–870.

Paluck EL, Green SA and Green DP (2019) The contact hypothesis re-evaluated. *Behavioural Public Policy* **3**, 129–158.

Paolini S, White FA, Tropp LR, Turner RN, Page-Gould E, Barlow FK and Gómez Á (2021) Intergroup contact research in the 21st century: Lessons learned and forward progress if we remain open. *Journal of Social Issues* **77**, 11–37.

Pettigrew TF and Tropp LR (2006) A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology* **90**, 751.

Rosenman ET, Basse G, Owen AB and Baiocchi M (2020) Combining observational and experimental datasets using shrinkage estimators. *Biometrics* **79**, 2961–2973.

Saguy T, Pratto F, Dovidio JF and Nadler A (2009) Talking about power: group power and the desired content of intergroup interactions.

Samii C, Paler L and Daly SZ (2017) Retrospective causal inference with machine learning ensembles: an application to anti-recidivism policies in Colombia. *Political Analysis* **24**, 434–456.

Scacco A and Warren SS (2018) Can social contact reduce prejudice and discrimination? Evidence from a field experiment in Nigeria. *American Political Science Review* **112**, 654–677.

Schroeder J and Risen JL (2016) Befriending the enemy: outgroup friendship longitudinally predicts intergroup attitudes in a coexistence program for Israelis and Palestinians. *Group Processes & Intergroup Relations* **19**, 72–93.

Shwed U, Kalish Y and Shavit Y (2018) Multicultural or assimilationist education: contact theory and social identity theory in Israeli Arab–Jewish integrated schools. *European Sociological Review* **34**, 645–658.

Tropp LR and Pettigrew TF (2005) Relationships between intergroup contact and prejudice among minority and majority status groups. *Psychological Science* **16**, 951–957.

Tropp LR, Hawi DR, Van Laar C and Levin S (2012) Cross-ethnic friendships, perceived discrimination, and their effects on ethnic activism over time: a longitudinal investigation of three ethnic minority groups. *British Journal of Social Psychology* **51**, 257–272.

van Dijk A, Thomaes S, Poorthuis AM and Orobio de Castro B (2019) Can self-persuasion reduce hostile attribution bias in young children?. *Journal of Abnormal Child Psychology* **47**, 989–1000.

Weiss CM (2021) Diversity in health care institutions reduces Israeli patients' prejudice toward Arabs. *Proceedings of the National Academy of Sciences* **118**, e2022634118.

Wright SC, Brody SM and Aron A (2005) Intergroup contact: still our best hope for improving intergroup relations. In Crandall CS and Schaller M (eds), *Social Psychology of Prejudice: Historical and Contemporary Issues*. Seattle, WA: Lewinian Press, pp. 115–142.

Zhou Y-Y and Lyall J (2022) Prolonged contact does not reshape locals' attitudes toward migrants in wartime settings: experimental evidence from Afghanistan. *Available at SSRN 3679746*.

Cite this article: Asimovic N, Ditlmann RK and Samii C (2024) Estimating the effect of intergroup contact over years: evidence from a youth program in Israel. *Political Science Research and Methods* 1–19. https://doi.org/10.1017/psrm.2024.8