Has Intergroup Contact Delivered?

Prepared for the Annual Review of Economics.

Matt Lowe*

Abstract

Intergroup contact is arguably the prejudice-reduction intervention with the most empirical support. However, recent meta-analyses of experimental contact interventions find signs of publication and reporting biases. In an effort to avoid such bias, I carry out a meta-analysis of 34 pre-registered contact experiments, considering only treatment effects on pre-registered primary outcomes. I find limited positive effects of intergroup contact of around one-twentieth of a standard deviation. Contact is more effective at changing behavior and attitudes towards people met than toward the outgroup as a whole. I conclude with suggestions for how contact researchers might make progress on this problem of generalization.

Keywords: intergroup contact, discrimination, prejudice, social interactions, meta-analysis, pre-registration.

*matt.lowe@ubc.ca, University of British Columbia. I thank Diego Delic and especially Catalina Garcia Valenzuela for outstanding research assistance. I gratefully acknowledge financial support from the CIFAR Azrieli Global Scholars Program. August 2024.

1 Introduction

Prejudice and discrimination are global concerns, and they feature at the core of key global challenges – whether the politics of inflamed anti-immigrant sentiment (Hangartner et al. 2019), climate change-induced conflict between farmers and pastoralists (McGuirk and Nunn 2024), or the persistence of racial inequality (Derenoncourt et al. 2023). While these challenges create huge demand for reliable prejudice-reduction interventions, and while such demand has existed for decades, it is remarkable how weak the evidence base is for real-world interventions. In a magisterial review of 418 prejudice-reduction experiments, Paluck et al. (2021) note that most experiments evaluate only light-touch interventions, and that the evidence base is beset by signs of publication bias. They conclude that "much research effort is theoretically and empirically ill-suited to provide actionable, evidence-based recommendations for reducing prejudice."

Intergroup contact – interpersonal contact between groups under favorable conditions – is arguably the most-studied and most-supported class of prejudice-reduction intervention. For many, the jury is in – intergroup contact is a proven intervention. For example, Pettigrew (2021) writes that "intergroup contact theory is now one of the best-supported theories in social psychology." For others, the jury is out – in particular, while field experiments tend to find that contact reduces prejudice, their effects display signs of p-hacking and file drawer problems (Paluck et al. 2019; Clochard 2024).

In this review, I take advantage of the boom in contact-based field experiments over the past five to ten years, noting that we now have enough pre-registered experiments to permit a meta-analysis that aims to purge publication bias-related concerns.

In the next section, I review past meta-analytic findings on the effects of contact interventions. These meta-analyses find positive effects of roughly 0.2 to 0.4σ, though with signs of p-hacking and file drawer problems. In Section 3, I describe my approach to meta-analysis, which focusses on 34 pre-registered contact experiments, considering only the effects of contact on pre-registered primary outcomes. The red flags of p-hacking are not present in my eligible set of experiments and effects, suggesting that the focus on pre-registered primary outcomes is effective at purging the sample of bias.

In Section 4 I report the results of the meta-analysis. I find the effects of intergroup contact on prejudice and intergroup relations to be roughly one-twentieth of a standard deviation, and the effects of bundled contact interventions (e.g. including training) to be closer to one-tenth of a standard deviation. I show that the weak positive effects of intergroup contact reflect a lack of *generalized* attitude and behavior change, and that they are not due to the contact lacking favorable conditions (e.g. common goals). In Section 5 I discuss two avenues for future work –

one on the generalization problem, and one on the general equilibrium effects of contact.

2 What Do We Know about the Effects of Intergroup Contact?

The contact hypothesis conjectures that interpersonal contact between groups can reduce prejudice when certain conditions are met. This formulation of the idea goes back to Allport (1954), who wrote in *The Nature of Prejudice* that:

> Prejudice (unless deeply rooted in the character structure of the individual) may be reduced by equal status contact between majority and minority groups in the pursuit of common goals. The effect is greatly enhanced if this contact is sanctioned by institutional supports (i.e., by law, custom or local atmosphere), and provided it is of a sort that leads to the perception of common interests and common humanity between members of the two groups.

Thomas Pettigrew, a graduate research assistant of Allport's at the time, refined the theory. Pettigrew re-stated the necessary conditions for contact to be effective as (i) equal status of the groups within the situation, (ii) the support of authorities, law or custom, (iii) common goals, and (iv) intergroup cooperation (Pettigrew 2021).

The theorizing of Allport and Pettigrew came out of an attempt to unify the results of observational studies, some of which can be thought of as quasi-experiments. Systematic meta-analyses came later, with the most influential being Pettigrew and Tropp (2006). In a paper that went on to be cited over 12,000 times, Pettigrew and Tropp undertook the herculean effort of meta-analyzing 515 studies that tested the effects of in-person intergroup contact. The meta-analysis has three primary findings relevant for the current review. First, intergroup contact typically reduces prejudice, with an average effect of Cohen's $d = 0.43$. Second, the effects of contact generalize far beyond the outgroup members directly met and the situation of the contact. In fact, effect sizes are similar whether the outcome involves the outgroup members directly met, the outgroup as a whole, or even outgroups that were not involved in the intervention. Third, the four necessary conditions for contact to be effective appear not to be necessary at all – even contact that does not satisfy the conditions has beneficial effects, though there is some evidence that studies that satisfy the conditions have larger effect sizes.

The rightly-influential meta-analysis of Pettigrew and Tropp (2006) presents a glowing view of contact as a prejudice-reduction intervention: it works, it works even for outgroups you

don't target, and it works even when the contact does not satisfy desirable conditions. In addition, while most of the underlying studies lack clean identification, Pettigrew and Tropp (2006) present some evidence that even the higher quality studies – including experiments and quasi-experiments – consistently find that contact is effective.

In the decade that followed, intergroup contact experiments became sufficiently common to allow a meta-analysis including only experiments. Paluck et al. (2019) undertook this task, assembling 27 intergroup contact experiments with delayed outcome measures. These experiments corroborate the first key finding of Pettigrew and Tropp (2006) – contact typically reduces prejudice, with a near-identical average effect of 0.39σ. However, Paluck et al. (2019) note signs of p-hacking and file drawer problems. First, studies with larger standard errors have more positive effect sizes, consistent with the under-reporting of null effects. In fact, the predicted effect size for a study with infinite precision is roughly zero. Second, the mean effect size is only 0.016 when restricting only to the three studies with pre-analysis plans.

Clochard (2024) extends the dataset of Paluck et al. (2019), reaching a sample of 44 papers. Like Paluck et al. (2019), he finds that contact typically reduces prejudice, although with a smaller mean effect size of 0.22σ. Again, he finds signs of p-hacking and file drawer problems: effect sizes and standard errors are strongly positively correlated, and most relevant for the current meta-analysis, the 13 pre-registered studies in his sample have smaller effects of about 0.13σ.

Summarizing, existing meta-analyses conclude that contact typically reduces prejudice, but raise concerns that these effects may be over-estimated due to reporting biases. I take these concerns seriously by taking an approach to meta-analysis that is designed to minimize such biases – an approach that focuses on tracking the effects on primary outcomes laid out in pre-registrations.

3 Meta-Analysis Approach

My goal is to characterize the effects of contact interventions on prejudice and intergroup relations while avoiding the selection problems that favor statistically significant results – including publication bias, outcome switching, and selective reporting. I describe my methodological approach in this section. Readers interested only in the results can skip to Section 4.

I take three main steps. First, I identify the universe of randomized controlled trials with an intergroup contact treatment that are pre-registered with the two most popular registry sites. Second, I record standardized treatment effects on outcomes that were pre-specified as primary outcomes. Third, I use frequentist and Bayesian meta-analysis techniques to summarize these

effects.

The first two steps are designed to mitigate reporting biases. By attempting to track the results from all pre-registrations, we avoid selection at the study-level in which studies come to our attention. By focussing on primary pre-specified outcomes, we avoid selection at the outcome-level in which studies emphasize outcomes that are more strongly affected. That said, this approach faces issues if pre-registered experiments with null results are less likely to be written up, and if write-ups selectively drop primary pre-specified outcomes that were less affected. I give evidence below suggesting that these two channels of selective reporting are unlikely to substantially affect my conclusions.

Identifying Experiments. In the first step, I use keyword searches to identify intergroup contact-related experiments pre-registered in either the *AEA RCT Registry* or in the *EGAP Registry* by the end of December 2023.[1] Among these 1,663 pre-registrations, I narrow down to the 46 experiments that satisfy the following three eligibility criteria:

1. The study must be pre-registered prior to the analysis of outcomes
2. The study must randomize contact between pre-existing, well-defined groups
3. The study must include at least one outcome that can be used to measure the effects of contact on prejudice, or more broadly, intergroup relations

The first criterion ensures that pre-registration of a study is indeed a *pre*-registration, and not posted after researchers have estimated treatment effects. The second criterion rules out natural experiments (e.g. Weiss 2021) and experiments that create variation in social interaction along dimensions that would not be thought of as well-defined groups (e.g. test scores). The third criterion mainly rules out studies in which researchers are interested in the effects of group-mixing on the performance of the group (e.g. team productivity), without looking at effects on the subsequent beliefs and behaviors of group members.

In other ways, these criteria are fairly loose. Unlike the meta-analyses discussed above, I allow for intergroup contact online, by not requiring that the contact be in-person (although I exclude imagined and vicarious contact). Like the meta-analyses above, I allow for the intergroup contact to be part of a bundled intervention (e.g. contact bundled with skills training, as in Zhou and Lyall (2023)). I include studies with contact of any length, and with contact that

[1] Specifically, I search for all registry entries that include at least one of the following words, in any part of the registry text: contact, intergroup, discrimination, integration, and prejudice. The total number of entries up to December 2023 are 8,142 for the AEA registry and 2,733 for the EGAP registry. The keyword search narrows these down to 1,250 AEA studies and 413 EGAP studies.

need not satisfy desirable conditions (like that of common goals). Though I show below how the effects of contact differ according to these different features.

Of the 46 pre-registered experiments, I was able to find results for 34 experiments across 32 papers through web searches and contacting authors for preliminary drafts. The large number of completed pre-registered contact experiments demonstrates both the maturity of this line of research, and its rapid expansion over just a few years. In particular, Paluck et al. (2019) reviewed 27 intergroup contact experiments, without pre-registration as an eligibility criteria. Only three of these experiments are included in our sample, meaning that in only a few years, we have reached a substantial sample of pre-registered experiments with almost no overlap with Paluck et al. (2019). In addition, while Clochard (2024) updates the analysis of Paluck et al. (2019) to include 44 papers with contact interventions, only 14 overlap with our eligible sample of 32 papers.

Even with a focus on tracking the results of pre-registered experiments, there may be a selection problem if the 34 experiments with results are not representative of the full set of 46 pre-registered experiments. In particular, one might worry that academics write up their results faster when their results are statistically significant. To explore this concern, I contacted study authors to ask about the status of their experiments. Of the 12 pre-registered experiments without results, five halted due to plausibly exogenous logistical reasons (e.g. recruitment problems, an ongoing conflict), and three have not yet collected endline data. Of the remaining four, three collected endline data within the past 18 months. This leaves little scope for selection into the sample of 34 experiments based on results. However, to the extent that some selection is possible, and that null results are less likely to be written up, we could consider positive treatment effects estimated below to be upper bounds.

Characteristics of Experiments. The 34 eligible experiments cover 20 countries and 35,391 participants (Table 1). Thirty-one experiments involve in-person contact, with the remaining three studying the effects of online contact (without video) with out-partisans.

Different experiments permit different types of comparisons (far-right columns, Table 1). Twenty-two experiments have what I call "clean" exogenous variation in intergroup contact – meaning that the intergroup contact treatment is not bundled with other substantive components. An example of clean variation would be the random assignment of individuals to mixed versus homogeneous sports teams (Mousa 2020; Lowe 2021). Nineteen experiments allow us to compare individuals assigned to a bundled outgroup contact intervention with those in a pure control group. For example, Zhou and Lyall (2023) study the effects of a vocational skills-training program in Afghanistan in which locals interacted with migrants. Finally, six experiments also

permit a comparison of ingroup interaction with a pure control group. For example, Scacco and Warren (2018) assign some Christian and Muslim men in Nigeria to homogeneous-class vocational training and some to a pure control group (among other treatments). While I discuss the effects of such ingroup contact below, I focus primarily on the effects of clean and bundled outgroup contact.

Most contact interventions satisfy most or all of the four conditions emphasized by Allport (1954), and 27 of 34 have delayed outcome measurement (Table A1). However, contact interventions vary vastly in how intensive they are: ranging from just a few minutes of interaction on a doorstep (Kalla and Broockman 2020), to over one thousand hours of contact in the Norwegian military's boot camp (Finseraas and Kotsadam 2017; Dahl et al. 2021).

Recording Treatment Effects. For each of the 34 contact experiments with results, I record standardized treatment effects and standard errors by taking the following steps:[2]

1. I identify the primary outcomes related to prejudice and intergroup relations pre-specified in the pre-registration (and using pre-analysis plans where available).

2. I find the corresponding outcomes reported in the paper. On average, each experiment reports results for 92% of pre-specified primary outcomes, suggesting little scope for selective reporting.

3. I code each outcome as either *generalized* to the outgroup or not. An example of a generalized outcome would be a feeling thermometer score for how a local feels about immigrants in general. An example of a non-generalized outcome would be the number of outgroup friends a participant has, given that this number could include people met as part of the contact intervention. This is an important distinction given that the contact hypothesis posits that certain types of interpersonal contact reduce general prejudice toward the outgroup, and not just that certain types of interpersonal contact lead to new intergroup social connections.

4. I record the point estimate and standard error of the effect of contact on each of these outcomes. I code the sign of the point estimate such that a more positive point estimate means participants are becoming less prejudiced towards their ingroup (or more inclusive towards the outgroup). For choosing the specification and comparison to use:

[2]See Online Appendix A.1 for additional details, and Online Appendix A.3 for a detailed description of paper-by-paper judgment calls taken when coding.

- I choose the treatment effect for the pooled sample (e.g. both the majority and minority group, or all endline timepoints) where available, otherwise I record each group's or timepoint's treatment effect separately.
- I record the effects of the following types of comparisons: (i) clean comparison of high versus low or no intergroup contact, holding constant social interaction (e.g. comparing the effects of being assigned to a mixed versus a homogeneous sports team), (ii) outgroup contact versus pure control group (a "bundled" comparison), and (iii) ingroup contact versus pure control group. Comparison (i) is the central comparison of interest, as it allows a test of the contact hypothesis without obvious confounding factors.
- Where there is a choice between specifications, I choose the one that was pre-specified.
- If there are multiple types of contact, I record the effects of the type pre-specified to be more effective (e.g. for Greene et al. (2024) I record only the effects of equal status contact).

5. I record standardized effect sizes and standard errors when available in the paper or in posted replication files. Otherwise, I request the control group standard deviation from authors, and use this to standardize effects and standard errors reported in the paper. In the latter case, the standardized effect is Glass's delta.

These steps result in 191 standardized treatment effects from the 34 experiments. Most experiments have multiple treatment effects recorded, through having multiple primary pre-specified outcomes, multiple timepoints or groups for a given outcome, or multiple eligible comparisons. To answer different questions below, I keep different sets of these treatment effects before aggregating – for example, in one exercise, I keep the effects of clean variation in contact, and restrict to generalized outcomes.

After deciding on the set of treatment effects to keep, the final step before analysis is to collapse each experiment's set of treatment effects and standard errors to only one effect and standard error. This aggregation should balance two concerns: experiments that report more effect sizes should not be mechanically given more weight, and yet the extra information from having effects on two related outcomes rather than one should deliver a corresponding reduction in standard errors. Here I follow Borenstein et al. (2021). I collapse standardized effect sizes to the experiment-level by taking the simple mean. I collapse standardized standard errors by also taking into account the typical positive correlation between effect sizes on related outcomes, or the same outcome at different timepoints. I calibrate this correlation to deliver the

same reduction in standard errors that I find in data from Lowe (2021) and Ghosh et al. (2024) when estimating effects on an index outcome as opposed to individual components (see Online Appendix A.2 for further details). Nevertheless, the core conclusions are not sensitive to the correlation assumed (see Figures A10 and A11).

Checks for p-Hacking. Experiment-level treatment effects are not correlated with standard errors, whether focussing on the clean or bundled effects of contact (Panels (a) and (b), Figure 1). This speaks against the possibility of under-reporting of statistically insignificant results, in which case we would expect a positive correlation between effect sizes and standard errors, as found in Paluck et al. (2019). The weak correlation remains when using effect-level observations (Panels (c) and (d)), in contrast to Clochard (2024). In addition, there is little evidence of heaping of t-statistics just above significance thresholds (Panels (a) to (c), Figure A1), unlike the effects studied in Clochard (2024) (Panel (d)), where there was no requirement for effects to be pre-registered as primary outcomes. Such heaping can be a sign of p-hacking (Brodeur et al. 2016). The lack of heaping then speaks in favor of my approach to study and effect selection.

Meta-Analysis Approach. I take experiment-level standardized effect sizes ($\hat{\tau}_k$) to be unbiased estimates of experiment-specific underlying effects (τ_k), such that

$$\hat{\tau}_k = \tau_k + \varepsilon_k \tag{1}$$

for each experiment k. Our interest is in estimating the average treatment effect of contact across settings, $\tau = E[\tau_k]$, as well as measures of the heterogeneity in τ_k.

With experiment-level effect sizes and standard errors in hand, I take two approaches to aggregation. Following a frequentist approach, I estimate a random effects model with restricted maximum likelihood. This approach estimates the pooled effect size ($\hat{\tau}$) as the weighted sum of the effect size from each experiment, with the weights equal to the inverse of the estimated variance of each effect – meaning that greater weight is given to experiments with more precise estimates.

Second, I follow Meager (2019) and Lund et al. (2024) and estimate a Bayesian Hierarchical Model by applying the simple evidence aggregation model from Rubin (1981):

$$\hat{\tau}_k \sim N(\tau_k, \hat{\text{se}}_k^2) \quad \forall\, k \tag{2}$$

$$\tau_k \sim N(\tau, \sigma_\tau^2) \quad \forall\, k \tag{3}$$

where $\hat{\text{se}}_k$ denotes the standard error of standardized effect size k. The Rubin model involves a hierarchical structure in which each experiment has its own unobserved treatment effect parameter τ_k, for which we have an unbiased estimate, $\hat{\tau}_k$, given randomization. Beyond this, each unobserved treatment effect parameter is drawn from a common distribution with mean τ and variance ${\sigma_\tau}^2$. The Bayesian approach aims to jointly estimate the average effect and the heterogeneity in effects, properly isolating effect size variation driven by true heterogeneity as opposed to sampling variation.[3] Unlike the frequentist approach, Bayesian estimation requires priors. I use weakly informative priors, following the default in R's `baggr` package:

$$\tau \sim N(0, 10 \times \max\{\hat{\tau}_k\}_{k=1}^K) \tag{4}$$

$$\sigma_\tau \sim \text{U}(0, 10\tilde{\sigma}) \tag{5}$$

where $\tilde{\sigma}$ is the standard deviation of $\{\hat{\tau}_k\}_{k=1}^K$.

Like the frequentist approach, the Bayesian approach delivers an estimate of τ, the average effect of intergroup contact. In addition, I characterize heterogeneity by reporting the Bayesian estimate of σ_τ, the standard deviation of the distribution of possible effect size draws. Finally, I also use the model estimates to deliver 95% posterior predictive intervals. These intervals ask the policy-relevant question: what τ_k should we expect if we are to run an intergroup contact intervention in a new setting? The 95% interval is the central range of τ_k we expect to happen 95% of the time.

4 Meta-Analysis Results

Overview. The core meta-analysis results are summarized in Figure 2.[4] I estimate an average effect of clean contact of roughly 0.03 to 0.05σ, with 95% confidence intervals always including zero (Panel (a), left side). This average effect is roughly one-tenth the size of the 0.39σ

[3] For a description of the advantages of Bayesian estimation of hierarchical models over frequentist estimation, see Meager (2019).

[4] See Figures A2 to A9 for frequentist forest plots. The results are similar when assuming a within-experiment effect size correlation of zero or 0.8 prior to collapsing effects to the experiment-level (Figures A10 and A11).

estimated by Paluck et al. (2019), and one-third the size of the 0.13σ estimated by Clochard (2024) when restricting to pre-registered studies. The estimated effect is similar when restricting only to generalized outcomes, to Allport-optimal in-person studies, and to interventions that last at least four hours.

I estimate larger effects close to 0.1σ of bundled contact interventions (Panel (b), left side). The bundled effect is smaller and becomes statistically insignificant when restricting to Allport-optimal in-person studies and longer interventions.

Beyond mean effects, I find evidence of moderate heterogeneity of the effects of contact across sites (right side of Figure 2), with an estimated standard deviation of roughly 0.1σ. It follows that while the estimated mean effect of clean interventions is small, the predictive interval for the effects in future sites has a moderate range – with the 50% interval roughly spanning 0 to 0.1σ, and the 95% interval roughly spanning -0.15 to 0.25σ. The spans for bundled interventions are somewhat larger. A rough summary of the posterior predictive intervals would be that 75% of future clean and bundled contact interventions are predicted to have positive effects.

The Generalization Problem. I estimate an average effect of roughly one-twentieth of a standard deviation for clean interventions, including all 81 treatment effects from 22 experiments (top row of Figure 2). For a binary outcome, this would amount to a three percentage point increase relative to a base of 50%.

Of the 81 treatment effects, 68 are generalized while 13 effectively include the specific people met during the intervention. The estimated effect size drops from 0.042 to 0.028 when keeping only the generalized effects. If I repeat the frequentist meta-analysis keeping only the 13 non-generalized effects across six experiments, I estimate an average effect of 0.16σ (95% confidence interval -0.04 to 0.36). Similarly, effect size-level regressions suggest that treatment effects on non-generalized outcomes are substantially larger than those on generalized outcomes, even when including experiment fixed effects (Table 2). Where Pettigrew and Tropp (2006) concluded that contact interventions had generalized effects far beyond what we would predict (even to outgroups not involved in the contact), it appears that pre-registered contact interventions have a generalization problem.

Mousa (2020) and Lowe (2021) are two studies that reflect the general finding in Table 2. Mousa (2020) randomly assigned Iraqi Christians to all-Christian or mixed Muslim-Christian football teams, while Lowe (2021) randomly assigned Indians to homogeneous or mixed-caste cricket teams. Mousa (2020) finds the largest positive effect on the least generalized outcome: Christians assigned to mixed teams were 49 percentage points more likely to train with Muslims six months after the intervention ended. Lowe (2021) finds large positive effects on a similar

outcome: participants assigned to teams with only other-caste players selected roughly 50% more other-caste players to be in their team for a future match.

These two studies suggest that randomly assigned contact can stick – the contact is positive enough that people form close friendships and want to keep interacting. But neither study gives strong evidence that these positive effects spill over to the outgroup as a whole. Mousa (2020) finds no detectable effects on "off-the-field" outcomes – whether participants visit a mixed social event or a restaurant in Mosul, or donate to a mixed NGO. Lowe (2021) presents some evidence of generalization – treated players trade and become friends with other-caste participants that were not on their team – but these effects are weaker, and do not amount to generalization beyond the social network of the village.

It appears that the weak effects of clean contact documented in Figure 2 are not due to a failure of the first step – whether participants have positive interactions and become friends with outgroup members – but instead the step of generalization. How can we make progress on the problem of (non-)generalization? I return to this question in Section 5.

The Allport Conditions. Allport (1954) posited four necessary conditions for the effects of contact to be positive, while Pettigrew and Tropp (2006) found empirically that such conditions are facilitating, but not necessary. In my data, restricting to the 16 in-person contact experiments that satisfy the four conditions does not meaningfully increase the estimated effect size (Panel (a), Figure 2). However, several studies report more compelling evidence on the Allport conditions, by designing experiments that explicitly randomize the presence of a condition.

In Lowe (2021), I explore the role of common goals and intergroup cooperation. As well as randomly assigning Indians to cricket teams – creating variation in common-goal, or collaborative, contact – I randomly assigned the teams to opponents – creating variation in opposing-goal, or adversarial contact.[5] I find that collaborative contact has strong positive effects on intergroup behaviors, while adversarial contact has imprecisely estimated negative effects. These results support the intuitive idea that common goals should make contact more effective.

Beyond common goals, I find a less intuitive result when studying the role of intergroup cooperation. I randomly assigned the cricket teams to receive either individual-level or team-level performance-related pay, with the former providing incentives for intra-team competition, and the latter providing incentives for intra-team cooperation. The effects of collaborative contact are similar for both types of incentive. The core finding is then not only that common goals make contact more effective, but also that once groups have common goals, the introduction of

[5]Note that the effects of adversarial contact are not included in the meta-analysis summarized in Figure 2 as I did not pre-specify these effects in the pre-registration.

intra-group competition at the margin does not undo the positive effects of contact.

Ghosh (2023) provides another angle on the intergroup cooperation condition in a fascinating study of Hindu-Muslim contact within a factory in West Bengal, India. He randomly assigned Hindus to homogeneous or mixed production teams, and explores how the effects of contact depend on the nature of the production function. In particular, some teams work on high-dependency tasks – assembly line-type tasks that require a high degree of continuous coordination between workers. Other teams work on low-dependency tasks – tasks that require little coordination. As one might expect, mixing reduces productivity in high-dependency, but not low-dependency, tasks. But remarkably, mixing tends to improve the attitudes of Hindus only in high-dependency tasks – i.e. it is precisely the tasks in which mixing creates friction and productivity losses that see mixing improve intergroup relations.

For an exploration of the equal status condition, Greene et al. (2024) randomly assigned participants in Mexico to collaborate with out-partisans on a ten-minute task (since there is no condition with interaction with co-partisans, I classify this as a bundled intervention). The task involved answering trivia questions and discussing whether citizens value friendship or professional success more highly. The authors varied status by randomly informing some participants that their answers would count equally for pair-level rewards, and telling others that the answers of only one of the two would count. While both types of contact have immediate positive effects on tolerant behavior towards the out-partisans, only the effects of equal status persists. This finding then suggests an intriguing interpretation of a facilitating condition: it could be that conditions of contact matter more for the dynamics of effects than for the initial level.

Null Effects Despite Optimal Conditions. While the studies above provide some support for three Allport conditions, Figure 2 nevertheless shows that pre-registered studies tend to have small effects even when Allport-optimized and long. Which experiments drive this?

Elwert et al. (2023b) and Elwert et al. (2023a) study the effects of inter-ethnic and cross-gender interaction by randomizing the seating chart of public schools in Hungary. This created roughly 180 hours of intergroup contact before outcomes were measured. A key advantage of their approach is that this form of contact intervention is a type that can be easily scaled to millions of children, through public schooling systems, and at little cost. Indeed, this low-cost scalability would suggest that even a 0.05σ effect would pass a cost-benefit test. Nevertheless, despite high-powered estimates, given a sample of over two thousand students, neither study can reject the null of no effects – on outgroup discrimination and friendships (Elwert et al. 2023b) and on beliefs about the other gender's ability and preferences for mixed teams (Elwert et al. 2023a).

Two experiments in Norway had an even more intensive contact intervention in which soldiers were randomized to rooms for an eight-week boot camp period (Finseraas and Kotsadam 2017; Dahl et al. 2021). This military setting is arguably as intensive and Allport-optimized as we could hope – as Finseraas and Kotsadam (2017) write, "Soldiers of private rank have equal social status within the army, they share the common goals of the unit, they need to cooperate to solve their tasks and contact takes place in the context of an explicit, enforcing authority." Despite this, both papers find limited effects. Finseraas and Kotsadam (2017) find that contact with ethnic minorities improves views of the work ethic of immigrants, but does not affect views of welfare dualism or of whether immigrants make Norway a better place to live. Dahl et al. (2021) finds that men randomly assigned to live with women have more egalitarian attitudes later, but no effects persist at a six-month follow-up (and only these long-term effects enter the meta-analysis as the short-term effects were not pre-registered).

Collectively, these studies pose the question: would additional "conditions" ensure more consistently positive effects of contact? While this question is a little ill-formed (particularly given a history in psychology of adding more and more conditions to the original four, see Pettigrew 2021, p128), I discuss one idea, that of general equilibrium contact interventions, in Section 5.

Negative Effects Despite Optimal Conditions. Negative effects of non-optimal intergroup contact are well-known (for example, see Enos 2014; Hangartner et al. 2019; Lowe 2021). However, the posterior predictive intervals in Figure 2 suggest that even contact interventions that satisfy the Allport conditions will have negative effects roughly 25% of the time. Can even well-structured intergroup contact backfire?

Mousa et al. (2024) and Ghosh et al. (2024) both find rare evidence of negative effects of collaborative contact. Mousa et al. (2024) randomly assigned Lebanese and Syrian youths to mixed or homogeneous classes for a 12-week psycho-social support program. Two weeks after the classes ended, participants assigned to mixed classes held less prejudicial attitudes, but were five percentage points less likely to attend an event that emphasized outgroup culture. The negative effect is driven by the dominant group, the Lebanese. Ghosh et al. (2024) report a similar result – campers that have additional outgroup contact have 0.22σ lower willingness to play with an outgroup stranger. The effect is again driven by the dominant group, Hindus in this case. Mousa et al. (2024) suggest that the finding may be due to a saturation effect – outgroup contact during the intervention may crowd out subsequent interest in intergroup interaction.

In the case of Ghosh et al. (2024), an alternative possibility is that the effects of intergroup contact are non-linear, with positive effects on the extensive margin (which we cannot measure,

as there are no homogeneous teams in Ghosh et al. (2024)), and negative effects on the intensive margin, since our variation compares Hindus in teams that are 20% Muslim with those in teams that are 50% Muslim. Bazzi et al. (2019) and Anderberg et al. (2024) present natural experimental evidence related to this point. Bazzi et al. (2019) use a population resettlement program in Indonesia to explore how the size of ethnic groups affects integration, measured by national language use at home, among other outcomes. Integration is greatest in fractionalized communities with many small groups, and is lower in communities with few large groups (the parallel to 50:50 Hindu:Muslim teams). Anderberg et al. (2024) report similar findings from the quasi-random assignment of students to classrooms in Germany – the ingroup bias of native Germans is an inverse U-shaped function of the share of immigrant peers, with the peak at roughly 50%. Future experimental work might build on this work to establish what group sizes are optimal for contact to be effective.

Bundled Comparisons. The evidence for the full set of bundled contact interventions is more positive than that for clean interventions (Figure 2). This difference was not obvious ex ante. In particular, if the effects of *ingroup* contact are negative (as found in Scacco and Warren (2018)), and the clean effects of outgroup contact relative to ingroup contact are somewhat positive, the bundled effect is less positive than the clean effect. That said, the effect of bundled interventions is larger for the first two sets of estimates in Figure 2, but smaller for the two more restrictive sets.

To explore the relationship between bundling and effects more systematically, I report results from effect size-level regressions in Table A2. Bundled treatment effects are not statistically different to clean treatment effects, whether or not experiment or experiment-outcome-group fixed effects are included. There is then no general evidence that bundled interventions are more effective than their clean counterparts. But since the aspects bundled with contact are quite heterogeneous, let us consider lessons we can learn from particular examples.

Bundled Examples. In Ghosh et al. (2024) we explore the effects of mixed Hindu-Muslim youth camps on intergroup relations in West Bengal, India. We randomly assigned 412 adolescent boys to a pure control group, or one of two two-week long youth camps. We randomly assigned campers to teams of either five Hindus and five Muslims, or two Muslims and eight Hindus. In this experiment the clean contact estimate compares campers assigned to teams with more versus fewer outgroup teammates. The bundled estimate compares campers with the control group.

The clean effects of contact are negative for Hindus, positive for Muslims (although impre-

cise), and null overall. Despite this, the bundled effect is positive and substantive for both groups – 0.15σ on an omnibus index for Hindus and 0.1σ for Muslims – and remarkably, positive effects on intergroup friendships fully persist one year later. Why does the bundle outperform the additional contact? We find some evidence of positive effects of inclusive nationalism-oriented lectures during the camps, while we see little role for collective rituals, like synchronized singing and dancing. More generally, the results from this experiment demonstrate that an integrative intervention need not work primarily through exposure to more outgroup members.

In contrast, Zhou and Lyall (2023) find no effects of an integrated vocational skills-training program on the behaviors and attitudes of locals towards immigrants, despite the program involving over three hundred hours of intergroup contact. These null effects are precisely estimated, and persisted until at least eight months after the program ended.

Effects of Ingroup Contact. Given that clean contact interventions typically use ingroup-only contact as a comparison group, a separate question concerns the typical effects of ingroup contact. The answer to this question can help us better interpret the effects of outgroup contact. In particular, when outgroup contact is effective, is it because the outgroup contact is crowding out harmful ingroup contact? Or is it because outgroup contact is in and of itself beneficial?

Six of the 34 eligible experiments have both ingroup-only contact and pure control groups (far-right column, Table 1). Applying the frequentist analysis to these six, I do not find general evidence for the negative finding of Scacco and Warren (2018) – the estimated average effect is 0.02σ, or 0.04σ when keeping only generalized outcomes, though the 95% confidence interval is large in each case (-0.08 to 0.12, and -0.09 to 0.17). The imprecision shows that we learn relatively little about the effects of ingroup contact from the eligible set of experiments.

Scacco and Warren (2018) provide the most interesting case of ingroup contact backfiring. They assigned Christian and Muslim men in Nigeria to a 16-week vocational training program or a control group. Those assigned to the training were assigned to heterogeneous or homogeneous classes, while those in heterogeneous classes were assigned to an ingroup or an outgroup study partner. Four to six weeks after the program ended, participants assigned to homogeneous classes discriminate substantially more against the outgroup in a dictator game than those assigned to the control group.

What mechanisms might lead to negative effects of ingroup contact? Scacco and Warren (2018) note that participants assigned to homogeneous classes give more money to *both* outgroup and ingroup members, with a bigger increase for ingroup members leading to a bigger ingroup-outgroup gap. The authors conjecture that ingroup bonding leads to greater generosity to the ingroup, without changing generosity to the outgroup. Lowe (2021) reports a related find-

ing: men assigned to homogeneous-caste cricket teams are more trusting of others than control participants: they send more money in a trust game to their partner, regardless of the partner's caste group (i.e. in this case, the ingroup-outgroup gap in trust is not affected). Holding constant intergroup contact, it is also the case that participants assigned to teams with pre-existing friends send more in the trust game later.

Taking the two papers together, we might conjecture that bonding with ingroup members and friends increases generosity and trust. Though why these effects extended more to the ingroup in Scacco and Warren (2018) than in Lowe (2021) is more of a puzzle.

Of course, we might also expect the effects of ingroup contact on prejudice to depend on (i) the nature of the conversations between ingroup members, and (ii) whether prejudice is socially approved or sanctioned by the ingroup. Outside of the set of meta-analyzed experiments, Webb (2024) provides evidence on these points. Webb uses a field experiment to study discrimination against transgender women in Chennai, India. Participants in the control group are highly discriminatory – they are 32% less likely to hire a transgender worker than a non-transgender worker for a free grocery delivery. Participants randomly assigned to discuss hiring decisions with two known neighbors later do not discriminate at all – i.e. remarkably, discussion with ingroup members eradicates discrimination.

Why is the ingroup contact here so effective? Webb argues that when the discussions cover the topic of hiring trans workers, pro-trans statements are much more common than anti-trans statements. Part of this is explained by the fact that pro-trans participants are more vocal in the discussion. Moral arguments made during these discussions are persuasive – a treatment in which participants listen to the discussions, but do not participate, is similarly effective at reducing discriminatory behavior.

One interpretation of Webb (2024) is that it provides a possibility result: a particular type of communication with ingroup members can dramatically reduce discrimination. In addition, his experiment demonstrates a potentially large role for *structured* communication, as opposed to the typical unstructured communication of intergroup contact experiments.

Structured Communication vs. Contact. A series of creative experiments by Broockman and Kalla corroborates the effectiveness of structured communication, and two of their papers enter the meta-analysis here (Broockman and Kalla 2016; Kalla and Broockman 2020). Broockman and Kalla (2016) study a door-to-door canvassing intervention in Florida, USA, randomizing whether the canvasser initiated a discussion about transgender people, including a perspective-taking component, or a placebo discussion about recycling. In addition, the authors cross-randomized whether the canvasser was transgender themselves. The transgender-related

discussions increase tolerance by 0.3 to 0.4σ, with effects persisting at least three months later, and with similar effects for trans and non-trans canvassers. Kalla and Broockman (2020) report results from three experiments that replicate these positive effects, although with effect sizes roughly half as large. One of these experiments additionally shows that the perspective-taking component is crucial: a conversation involving arguments alone has no effect.

This set of experiments permit us to horse-race the effects of a short perspective-taking conversation with that of intergroup contact (the latter being the effects that enter the meta-analysis). Through a reanalysis of the publicly posted data, I find a +0.1σ effect of transgender contact at the first follow-up of Broockman and Kalla (2016), but otherwise the effects are null or even negative in the case of interaction with an immigrant canvasser in Kalla and Broockman (2020) (-0.05σ, SE = 0.02). A short perspective-taking conversation then clearly dominates a short interaction with an outgroup member. More recently, Mousa et al. (2024) report a related set of results – they find that empathy education is more effective than contact at improving prejudicial attitudes and behaviors.

Experiments with structured communication pose a theory of change that contrasts with that of the contact hypothesis. As discussed in Broockman and Kalla (2016), these experiments place System-2, active, effortful processing center stage, with participants challenged to explicitly think through their behaviors and attitudes regarding the outgroup. In comparison, the contact hypothesis posits a more passive process in which participants learn about the outgroup through osmosis. But the evidence in Figure 2 suggests that this passive process, even when playing out over hundreds of hours, leads to limited prejudice reduction.

5 Paths Forward

I have argued that pre-registered randomized contact interventions have had limited effects on their primary outcomes – in the region of one-twentieth of a standard deviation. In this section I consider two avenues for future work in response.

The Generalization Problem. A central question concerns when, and why, participants generalize from their experiences with outgroup members to the outgroup as a whole. Careful design of outcome measurement is crucial for understanding generalization. Perhaps surprisingly, relatively few contact experiments measure outcomes that characterize a participant's feelings towards the other participants they met as part of the intervention. Even fewer measure comparable outcomes at different levels of generalization. Scacco and Warren (2018) provide a helpful model for such measurement: their participants play a dictator game multiple times,

sometimes with a classmate, and sometimes with a stranger, with the order randomized. More generally, any behavioral game played with a named partner can cover multiple levels of generalization: my teammate, a study participant not on my team, a non-participant from the same village, an outgroup member from the same country, an outgroup member from a different country, etc. With this kind of measurement, researchers can trace generalization as a function of relatedness.

Otherwise, future work might take a more lab-experimental approach to understanding learning about outgroups, following the style of Conlon et al. (2022) – an experiment that sheds light on barriers to social learning. A crucial feature in that experiment is that the underlying signal structure is known, allowing the authors to compare the learning of participants with a Bayesian benchmark. An adapted version could be used to study learning about outgroups, first characterizing how much generalization we should even expect from a rational automaton, and then characterizing whether people under- or over-generalize relative to this benchmark (as in Augenblick et al. (2021)). This would answer the reframed question: is the lack of generalization because participants are being too rational (e.g. properly discounting noisy signals), or not rational enough (e.g. treating an outgroup member as an exception to the rule)?

With careful measurement in hand, we can explore what types of contact maximize generalization. For example, one hypothesis would be that *broad* contact – short interactions with many different outgroup members – may be important for generalization. Negative stereotypes about the outgroup may require many counterexamples to be corrected. In contrast, *deep* contact – intensive interactions with one outgroup member, as in random-roommate studies – may hinder generalization if the one outgroup member is rationalized as an exception to the rule. Alternatively, some psychologists argue that friendship potential is a key condition for effects of contact to be positive (Pettigrew 1998, p76), and friendship potential is more likely with deep contact. Consistent with this, Corno et al. (2022) find in South Africa that white students randomly assigned to live with a Black roommate are less racially biased, as measured by an implicit association test, and they have more positive attitudes towards Black people. Building on their work, we are testing the broad versus deep hypothesis in an ongoing experiment, by randomizing Hindus to work with the same Muslim for six days, or with a different Muslim each day (Chakraborty et al. 2024).

Psychologists have of course theorized about when and why generalization should occur. They consider the salience of existing group boundaries to be central to the process of generalization (Dovidio et al. 2003), though two opposing views coexist. One view is that generalization is maximized when group boundaries are made less salient during contact, making interactions personalized, as opposed to category-based, undermining the basis for category use

in future interactions (Brewer and Miller 1984). The opposing view is that salient group boundaries promote generalization, allowing participants to draw a direct link between the observed behavior of outgroup members and their group as a whole (Hewstone and Brown 1986, p16-20), and particularly when outgroup members are perceived as typical of their group (Brown et al. 1999). Making this more concrete, Hewstone and Brown (1986) write that "the more cues that indicate the group membership of a target, the greater should be the generalization."

The potential positive role of category salience is particularly interesting given that modern contact experiments often intentionally avoid category salience, in an attempt to reduce experimenter demand effects (e.g. see Lowe 2021, p1815). Future work might instead explicitly manipulate category salience and the perceived typicality of outgroup members, taking inspiration from the wealth of creative lab experiments in psychology (e.g. Wilder 1984; Van Oudenhoven et al. 1996; Gaertner et al. 1989). One naturalistic example of a high-category salience intervention is Model United Nations – an educational simulation in which students from different countries act as representatives of those countries, working together to solve a problem. The Model UN format could be easily re-applied to group boundaries other than those related to nationality.

The General Equilibrium Effects of Contact. Contact experiments typically randomize treatment at the individual-level. As a result, estimated effects may be underestimates if there are positive spillover effects on control participants, who are embedded in the same social networks as treated participants. Furthermore, to the extent that prejudicial behaviors are influenced by social norms that operate at the level of larger groups, like villages, individual-level interventions may fail to meaningfully change behavior, since the behavioral change of treated individuals may be constrained by entrenched norms. Only a handful of existing studies speak to these issues of spillovers and community-level norm change.

Grady et al. (2023) partnered with Mercy Corps to randomize a bundled contact intervention at the community-level, in the context of farmer-pastoralist conflict in Nigeria. Ten communities received the 18-month-long intervention: with mixed-group committees formed to spend 27,000 USD on community projects (e.g. boreholes), public forums held to discuss the drivers of conflict, and mediation provided to community leaders. Five communities were randomized to control. Post-intervention, treated communities report improved intergroup attitudes and greater physical security, and they are more likely to have pastoralists visiting the markets in the farmers' communities. Since most community members did not participate directly in the intervention's activities, the authors consider these treatment effects to be primarily driven by general norm change that affects the community as a whole. They also argue that the positive

effects may be due to the visibility of their intervention to community members. Consistent with this, psychologists have written extensively about the "extended contact effect," which posits that mere knowledge that an ingroup member has a close outgroup friendship can lead to improved attitudes (Wright et al. 1997; Zhou et al. 2019).

Mousa et al. (2024) explore spillovers more cleanly by surveying the parents of the Lebanese and Syrian adolescent participants of their 12-week psycho-social program. They find mostly null results, though one might speculate that vertical spillovers from child-to-parent may be weaker than horizontal spillovers from child-to-child.

Future experiments would ideally combine the approaches of Grady et al. (2023) and Mousa et al. (2024), by randomizing an intervention at the community-level, and then in treated communities, randomizing participation in the intervention at the individual-level. Such a design would allow a decomposition of the overall effect of the intervention into that driven by participation and that driven by spillovers to the broader community.

6 Conclusion

Judged by pre-registered field experiments, intergroup contact interventions have disappointed, delivering small positive effects, and limited generalization. My hope is that this review sparks two effects on future work on intergroup contact.

First, policymakers would do well to revise somewhat downward their hopes for "passive" contact-type interventions, and consider a move in the direction of interventions that engage active processing, perhaps in tandem with intergroup contact. To the extent that intergroup contact can be incorporated in programs at low cost, it likely should be – since the evidence from pre-registered experiments still suggests that the effects tend to be positive, albeit small.

Second, researchers would do well to move beyond the Allport conditions, recognising that even Allport-optimal contact interventions do not appear to consistently deliver positive effects. Future research might nevertheless aim to understand what types of contact are most effective, with a focus on understanding when and why contact leads to generalization. On this question there is a rich literature in psychology – theory and lab experiments – ripe for re-application within naturalistic field experiments.

References

Abril, V, E Norza, SM Perez-Vincent, S Tobón, and M Weintraub, "Building trust in state actors: A multi-site experiment with the Colombian National Police. InterAmerican Development Bank (Technical Note No: IDB-TN-2790)," 2023.

Adamu, Sewareg, David A. Dow, Fitsum Heilu, Mesele Mengsteab, Jeremy Springman, and Juan F. Tellez, "The Effect of Social Ties on Engagement Cohesion: Evidence from Ethiopian University Students," Technical Report, Working Paper 2024.

Allport, Gordon, *The Nature of Prejudice*, Garden City, NJ: Anchor, 1954.

Anderberg, Dan, Gordon Dahl, Cristina Felfe, Helmut Rainer, and Thomas Siedler, "Diversity and Discrimination in the Classroom," Technical Report, National Bureau of Economic Research 2024.

Asimovic, Nejla, Ruth K Ditlmann, and Cyrus Samii, "Estimating the effect of intergroup contact over years: Evidence from a youth program in Israel," *Political Science Research and Methods*, 2024, *12* (3), 475–493.

Augenblick, Ned, Eben Lazarus, and Michael Thaler, "Overinference from weak signals and underinference from strong signals," *arXiv preprint arXiv:2109.09871*, 2021.

Barros, Henrique, "The Power of Dialogue: Forced Displacement and Social Integration amid an Islamist Insurgency in Mozambique," Technical Report, Working Paper 2024.

Baseler, Travis, Thomas Ginn, Robert Hakiza, Helidah Ogude-Chambert, and Olivia Woldemikael, "Can Redistribution Change Policy Views?: Aid and Attitudes Toward Refugees in Uganda," Technical Report, Center for Global Development 2023.

Bazzi, Samuel, Arya Gaduh, Alexander Rothenberg, and Maisy Wong, "Unity in Diversity? How Intergroup Contact Can Foster Nation Building," *American Economic Review*, November 2019, *109* (11), 3978–4025.

Bezabih, Mintewab, Sosina Bezu, Tigabu Getahun, Ivar Kolstad, Päivi Lujala, and Arne Wiig, "Inter-group interaction and attitudes to migrants," *The Journal of Politics*, 2024, *0* (ja), null.

Borenstein, Michael, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein, *Introduction to meta-analysis*, John Wiley & Sons, 2021.

Brewer, Marilynn B and N Miller, "Beyond the contact hypothesis: Theoretical perspectives on desegregation," *Groups in contact: The psychology of desegregation*, 1984, *281*.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, "Star wars: The empirics strike back," *American Economic Journal: Applied Economics*, 2016, *8* (1), 1–32.

Broockman, David and Joshua Kalla, "Durably Reducing Transphobia: A Field Experiment on Door-to-Door Canvassing," *Science*, 2016, *352* (6282), 220–224.

Brown, Rupert, James Vivian, and Miles Hewstone, "Changing attitudes through intergroup contact: The effects of group membership salience," *European Journal of Social Psychology*, 1999, *29* (5-6), 741–764.

Burlacu, Sergiu, Davide Azzolini, and Federico Podestà, "Beyond Sight: Exploring the Impact of a Multifaceted Intervention on Knowledge, Attitudes and Behaviors towards Persons with Visual Impairment," Technical Report, Research Institute for the Evaluation of Public Policies (IRVAPP), Bruno … 2024.

Chakraborty, Anujit, Arkadev Ghosh, Matt Lowe, and Gareth Nellis, "Learning about Outgroups: Comparing Deep versus Broad Connections," 2024. AEA RCT Registry #13126, March 15.

Chaudhry, Zain and Karrar Hussain, "The Economic Effects of Inter-sectarian Contact," Technical Report, Working Paper 2024.

Clochard, Gwen-Jiro, "Improving the Perception of the Police by the Youth," Technical Report, Working Paper 2022.

_ , "Contact Interventions: A Meta-Analysis," *Center for Research in Economics and Statistics, mimeo*, 2024.

_ **, Guillaume Hollard, and Omar Sene**, "Low-Cost Contact Interventions Can Increase Inter-Ethnic Trust when Previous Contacts Were Scarce: Evidence from Senegal," Technical Report, Working Paper 2023.

Conlon, John J, Malavika Mani, Gautam Rao, Matthew W Ridley, and Frank Schilbach, "Not learning from others," Technical Report, National Bureau of Economic Research 2022.

Corno, Lucia, Eliana La Ferrara, and Justine Burns, "Interaction, Stereotypes, and Performance: Evidence from South Africa," *American Economic Review*, 2022, *112* (12), 3848–3875.

Dahl, Gordon, Andreas Kotsadam, and Dan-Olof Rooth, "Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams," *Quarterly Journal of Economics*, 2021, *136* (2), 987–1030.

Derenoncourt, Ellora, Chi Hyun Kim, Moritz Kuhn, and Moritz Schularick, "Wealth of Two Nations: The U.S. Racial Wealth Gap, 1860–2020*," *The Quarterly Journal of Economics*, 09 2023, *139* (2), 693–750.

Dovidio, John F, Samuel L Gaertner, and Kerry Kawakami, "Intergroup contact: The past, present, and the future," *Group processes & intergroup relations*, 2003, *6* (1), 5–21.

Elwert, Felix, Tamas Keller, and Andreas Kotsadam, "Effects of deskmate gender on confidence, attitudes toward mixed gender teams, and prejudice - Evidence from a large scale field experiment in Hungarian schools," Technical Report, Working Paper 2023.

_ , Tamás Keller, and Andreas Kotsadam, "Rearranging the Desk Chairs: A large randomized field experiment on the effects of close contact on interethnic relations," *American Journal of Sociology*, 2023, *128* (6), 1809–1840.

Enos, Ryan, "Causal Effect of Intergroup Contact on Exclusionary Attitudes," *Proceedings of the National Academy of Sciences*, 2014, *111* (10), 3699–3704.

Finseraas, Henning and Andreas Kotsadam, "Does personal contact with ethnic minorities affect anti-immigrant sentiments? Evidence from a field experiment," *European Journal of Political Research*, 2017, *56* (3), 703–722.

Freddi, Eleonora, Jan Potters, and Sigrid Suetens, "The effect of brief cooperative contact with ethnic minorities on discrimination," *Journal of Economic Behavior Organization*, 2024, *222*, 64–76.

Friedman, Willa, Guthrie Gray-Lobe, and Michael Kremer, "Worker Assignment and National Unity: Are All Stable Matches Socially Stable?," Technical Report, Working Paper 2024.

Gaertner, Samuel L, Jeffrey Mann, Audrey Murrell, and John F Dovidio, "Reducing intergroup bias: The benefits of recategorization.," *Journal of personality and social psychology*, 1989, *57* (2), 239.

Ghosh, Arkadev, "Religious Divisions and Production Technology: Experimental Evidence from India," *Duke University, mimeo*, 2023.

_ , Prerna Kundu, Matt Lowe, and Gareth Nellis, "Creating Cohesive Communities: A Youth Camp Experiment in India," Technical Report, Working Paper 2024.

Grady, Christopher, Rebecca Wolfe, Danjuma Dawop, and Lisa Inks, "How Contact Can Promote Societal Change Amid Conflict: An Intergroup Contact Field Experiment in Nigeria," *Proceedings of the National Academy of Sciences*, 2023, *120* (43), e2304882120.

Greene, Kenneth, Erin Rossiter, Enrique Seira, and Alberto Simpser, "Interacting as equals: How contact can promote tolerance among opposing partisans," Technical Report, Working Paper 2024.

Hangartner, Dominik, Elias Dinas, Moritz Marbach, Konstantinos Matakos, and Dimitrios Xefteris, "Does Exposure to the Refugee Crisis Make Natives More Hostile?," *American Political Science Review*, May 2019, *113* (2), 442–455.

Hewstone, Miles and Rupert J. Brown, "Contact is not enough: An intergroup perspective on the 'Contact Hypothesis'," in Miles Hewstone and Rupert Brown, eds., *Contact and conflict in intergroup encounters*, Oxford: Basil Blackwell, 1986, pp. 1–44.

Kalla, Joshua L. and David E. Broockman, "Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments," *American Political Science Review*, 2020, *114* (2), 410–425.

Loiacono, Francesco and Mariajose Silva-Vargas, "Can work contact improve social cohesion between refugees and locals? Evidence from an experiment in Uganda," Technical Report, Working Paper 2023.

Lowe, Matt, "Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration," *American Economic Review*, 2021, *111* (6), 1807–1844.

Lund, Crick, Kate Orkin, Marc Witte, John H Walker, Thandi Davies, Johannes Haushofer, Sarah Murray, Judy Bass, Laura Murray, Wietse Tol et al., "The Effects of Mental Health Interventions on Labor Market Outcomes in Low-and Middle-Income Countries," Technical Report, National Bureau of Economic Research 2024.

McGuirk, Eoin F and Nathan Nunn, "Transhumant pastoralism, climate change, and conflict in Africa," *Review of Economic Studies*, 2024, p. rdae027.

Meager, Rachael, "Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments," *American Economic Journal: Applied Economics*, 2019, *11* (1), 57–91.

Mousa, Salma, "Building Social Cohesion Between Christians and Muslims Through Soccer in Post-ISIS Iraq," *Science*, 2020, *369* (6505), 866–870.

_ , Lennard Naumann, and Alexandra Scacco, "Intergroup Contact, Empathy Training, and Refugee-Native Integration in Lebanon," Technical Report, Working paper 2024.

Oudenhoven, Jan Pieter Van, Jan Tjeerd Groenewoud, and Miles Hewstone, "Cooperation, ethnic salience and generalization of interethnic attitudes," *European Journal of Social Psychology*, 1996, *26* (4), 649–661.

Paler, Laura, Leslie Marshall, and Sami Atallah, "How Cross-Cutting Discussion Shapes Support for Ethnic Politics: Evidence from an Experiment in Lebanon," *Quarterly Journal of Political Science*, 2020, *15* (1), 33–71.

Paluck, Elizabeth Levy, Seth A Green, and Donald P Green, "The contact hypothesis reevaluated," *Behavioural Public Policy*, 2019, *3* (2), 129–158.

Paluck, Elizabeth, Roni Porat, Chelsey Clark, and Donald Green, "Prejudice Reduction: Progress and Challenges," *Annual Review of Psychology*, 2021, *72*, 533–560.

Pettigrew, TF, "Contextual social psychology: Reanalyzing prejudice, voting, and intergroup contact. American Psychological Association," 2021.

Pettigrew, Thomas F., "Intergroup contact theory," *Annual review of psychology*, 1998, *49* (1), 65–85.

_ **and Linda R. Tropp**, "A meta-analytic test of intergroup contact theory.," *Journal of Personality and Social Psychology*, 2006, *90* (5), 751.

Porat, Roni, Sarit Larry, John-Henry Pezzuto, and Devorah Manekin, "The Costs of Collaboration: Evidence From Two Field Experiments in Jerusalem," Technical Report, Working paper 2024.

Rossiter, Erin, "The similar and distinct effects of political and non-political conversation on affective polarization," Technical Report, Working paper 2023.

Rossiter, Erin L and Taylor N Carlson, "Cross-partisan conversation reduced affective polarization for republicans and democrats even after the contentious 2020 election," *The Journal of Politics*, 2024, *86* (4).

Rubin, Donald B, "Estimation in parallel randomized experiments," *Journal of Educational Statistics*, 1981, *6* (4), 377–401.

Scacco, Alexandra and Shana Warren, "Can Social Contact Reduce Prejudice and Discrimination? Evidence from a Field Experiment in Nigeria," *American Political Science Review*, 2018, *112* (3), 654–677.

Webb, Duncan, "Silence to Solidarity: How Communication About a Minority Affects Discrimination," Technical Report 2024.

Weiss, Chagai M, "Diversity in health care institutions reduces Israeli patients' prejudice toward Arabs," *Proceedings of the National Academy of Sciences*, 2021, *118* (14), e2022634118.

Wilder, David A, "Intergroup contact: The typical member and the exception to the rule," *Journal of Experimental Social Psychology*, 1984, *20* (2), 177–194.

Wright, Stephen C, Arthur Aron, Tracy McLaughlin-Volpe, and Stacy A Ropp, "The extended contact effect: Knowledge of cross-group friendships and prejudice.," *Journal of Personality and Social psychology*, 1997, *73* (1), 73.

Zhou, Shelly, Elizabeth Page-Gould, Arthur Aron, Anne Moyer, and Miles Hewstone, "The extended contact hypothesis: A meta-analysis on 20 years of research," *Personality and Social Psychology Review*, 2019, *23* (2), 132–160.

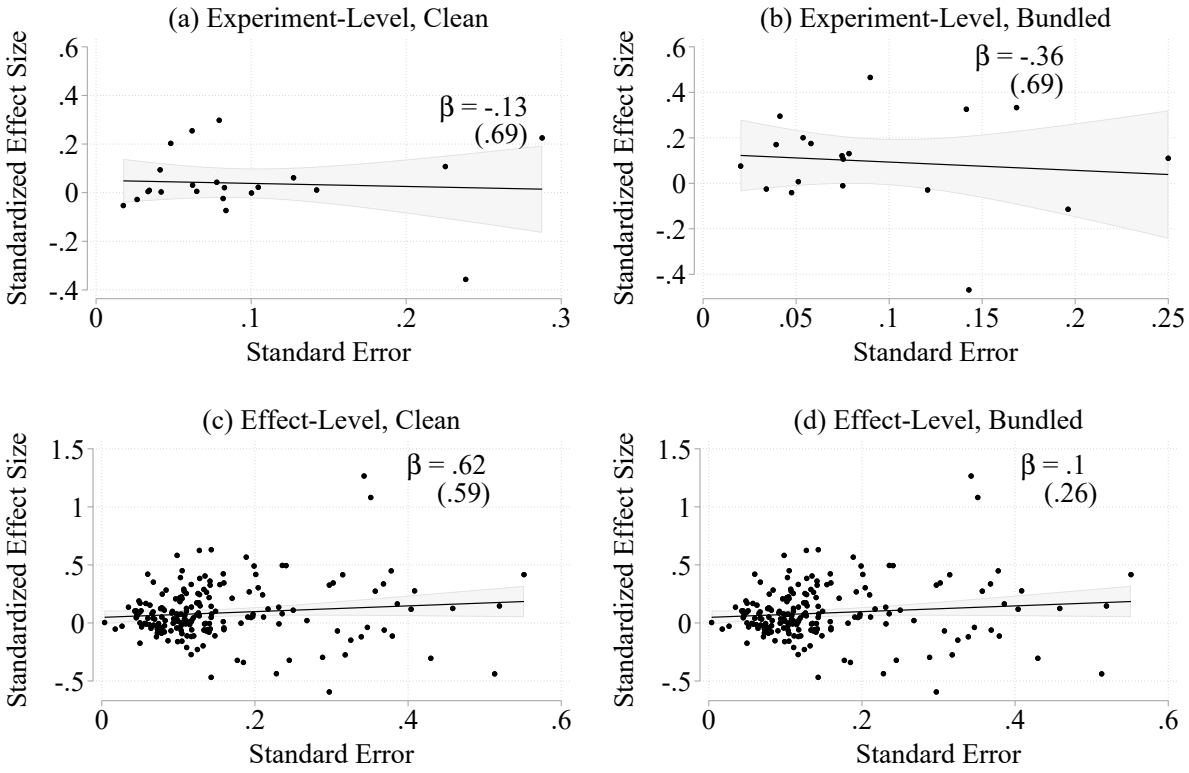
Zhou, Yang-Yang and Jason Lyall, "Prolonged contact does not reshape locals' attitudes toward migrants in wartime settings," *American Journal of Political Science*, 2023.

Table 1: Characteristics of Eligible Experiments

Paper	Pre-reg Year	Groups	Country	N	Type	Comparison Type		
						Clean	Outgroup-Ctrl	Ingroup-Ctrl
AEA Registry								
Finseraas and Kotsadam (2017)	2014	Ethnicity	Norway	577	In-person	✓		
Lowe (2021)	2016	Caste	India	1,261	In-person	✓	✓	✓
Elwert et al. (2023b)	2018	Ethnicity	Hungary	2,395	In-person	✓		
Elwert et al. (2023a)	2018	Gender	Hungary	2,776	In-person	✓		
Mousa (2020)	2018	Religion	Iraq	459	In-person	✓		
Friedman et al. (2024)	2018	Ethnicity	Kenya	2,251	In-person		✓	
Freddi et al. (2024)	2019	Ethnicity	Netherlands	114	In-person	✓		
Ghosh (2023)	2019	Religion	India	546	In-person	✓		
Baseler et al. (2023)	2020	Refugees	Uganda	1,406	In-person	✓	✓	✓
Bezabih et al. (2024)	2020	Immigrants	Ethiopia	600	In-person	✓	✓	✓
Dahl et al. (2021)	2020	Gender	Norway	781	In-person	✓		
Clochard (2022)	2021	Students-Police	France	366	In-person	✓	✓	✓
Loiacono and Silva-Vargas (2023)	2021	Refugees	Uganda	650	In-person		✓	
Greene et al. (2024)	2021	Partisans	Mexico	2,454	Online		✓	
Clochard et al. (2023)	2022	Ethnicity	Senegal	895	In-person	✓	✓	✓
Abril et al. (2023)	2022	Citizens-Police	Colombia	4,220	In-person		✓	
Burlacu et al. (2024)	2022	Disability	Italy	344	In-person		✓	
Barros (2024)	2022	IDPs	Mozambique	913	In-person		✓	
Chaudhry and Hussain (2024)	2022	Sect	Pakistan	302	In-person	✓		
Ghosh et al. (2024)	2022	Religion	India	412	In-person	✓	✓	
EGAP Registry								
Scacco and Warren (2018)	2015	Religion	Nigeria	138	In-person	✓	✓	✓
Broockman and Kalla (2016)	2015	Transgender	USA	501	In-person	✓		
Grady et al. (2023)	2015	Farmers-Pastoralist	Nigeria	1,539	In-person		✓	
Zhou and Lyall (2023)	2015	Immigrants	Afghanistan	1,276	In-person		✓	
Kalla and Broockman (2020)	2016	Immigrants	USA	1,578	In-person	✓		
	2016	Transgender	USA	1,044	In-person	✓		
Paler et al. (2020)	2016	Ethnicity/Class	Lebanon	720	In-person	✓		
Asimovic et al. (2024)	2020	Ethnicity	Israel	850	In-person		✓	
Rossiter (2023)	2020	Partisans	USA	740	Online		✓	
Rossiter and Carlson (2024)	2021	Partisans	USA	578	Online		✓	
Porat et al. (2024)	2022	Ethnicity	Israel	420	In-person	✓		
	2022	Ethnicity	Israel	430	In-person	✓		
Adamu et al. (2024)	2022	Ethnicity	Ethiopia	968	In-person		✓	
Mousa et al. (2024)	2023	Refugees	Lebanon	887	In-person	✓		

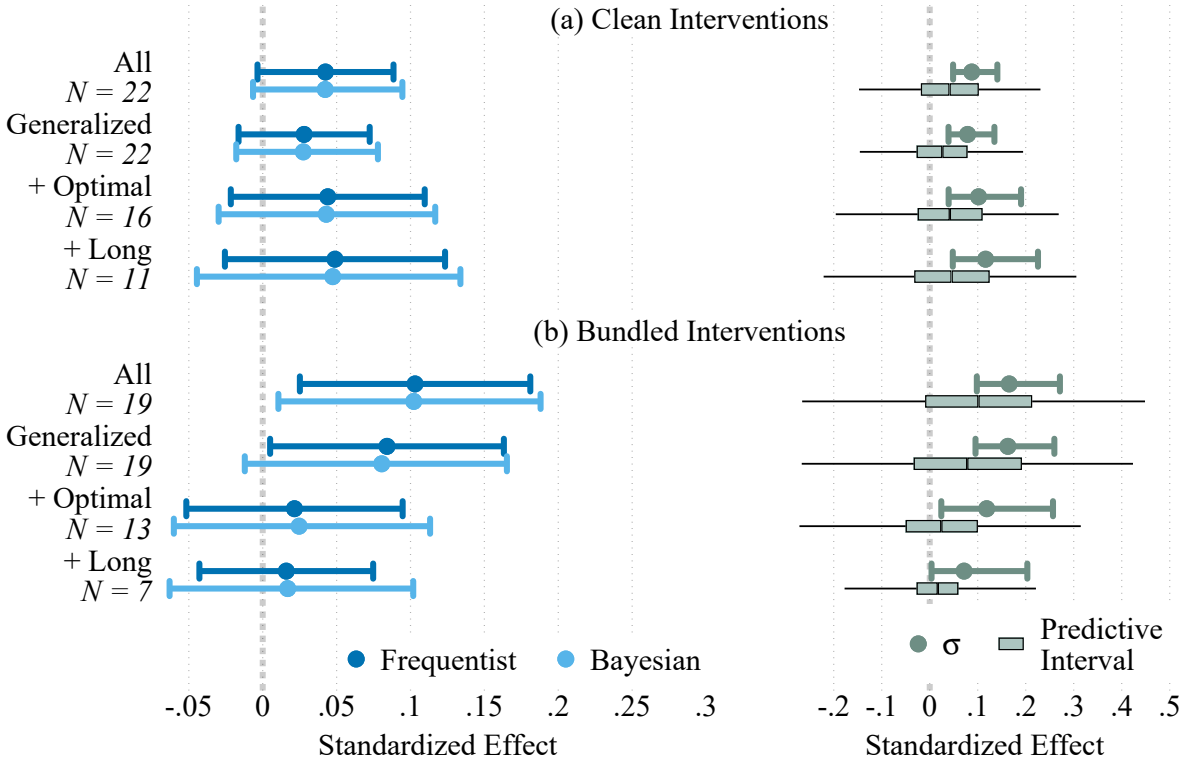
Notes: The papers are sorted by the date of the initial pre-registration. Clean denotes whether the experiment has clean exogenous variation in intergroup contact (high versus low or some versus none), without contact being bundled with other factors. An example of Clean variation would be random assignment to mixed-gender versus single-gender teams. Outgroup-Ctrl denotes whether the experiment can compare those with outgroup contact with a pure control group. Ingroup-Ctrl is the equivalent for ingroup contact.

Figure 1: Correlating Effect Sizes With Standard Errors



Notes: The figure plots standardized effect sizes against standard errors. Panels (a) and (b) use experiment-level observations (after aggregating multiple effect sizes for each experiment). Panels (c) and (d) use effect-level observations. Panels (a) and (c) use only clean comparisons of high versus low or no intergroup contact (e.g. assignment to mixed-gender versus single-gender teams). Panels (b) and (d) use only bundled comparisons of outgroup contact relative to a pure control group. Each panel includes a linear fit, 95% confidence interval, and the estimated slope and standard error from a regression with robust standard errors.

Figure 2: Effects of Intergroup Contact Interventions



Notes: The figure shows the output of frequentist and Bayesian meta-analysis of sets of experiment-level effect sizes and standard errors. Panel (a) considers only clean comparisons of high with low or no intergroup contact. Panel (b) considers only comparisons of bundled outgroup contact relative to a pure control group. Within these two groups, I show estimates using all outcomes and interventions (All), and only outcomes that are generalized to the outgroup (Generalized), only generalized outcomes with in-person contact interventions that satisfy all four Allport conditions (+ Optimal), and the same, but adding the restriction that the contact lasts at least four hours (+ Long). N denotes the number of experiments included in each sample. Point estimates and 95% confidence intervals for the average effect of contact (τ) are denoted in blue, while the equivalent for the standard deviation of effects across settings (σ) is denoted in green. Box plots describe posterior predictive intervals for the effect of contact in a new setting (τ_{K+1}). The whiskers span the 95% interval, the box spans the 50% interval, while the vertical line denotes the mean.

Table 2: The Generalization Problem

	Effect Size			
	(1)	(2)	(3)	(4)
Generalized Outcome	-0.15	-0.08*	-0.14**	-0.14**
	(0.09)	(0.04)	(0.06)	(0.06)
Observations	81	81	81	81
Non-Generalized Mean	0.20	0.20	0.20	0.20
Experiment FE	No	No	Yes	Yes
Experiment-Group FE	No	No	No	Yes
Weighted	No	Yes	Yes	Yes

Notes: The unit of observation is the treatment effect. The sample includes only the effects of clean contact interventions. Observations are weighted by the inverse of the standardized standard error in columns 2 to 4. Standard errors are clustered at the experiment-level in columns 1 and 2, and otherwise are robust. Experiment-by-group fixed effects are experiment fixed effects interacted with dummies for the group the effect was estimated for (e.g. a group could be immigrants). *** p<0.01, ** p<0.05, * p<0.1.

Online Appendix

Table A1: Allport Conditions and Other Characteristics of Eligible Experiments

Paper	Common Goals	Intergroup Cooperation	Equal Status	Authority Support	Hours	Measurement Timepoints
AEA Registry						
Finseraas and Kotsadam (2017)	✓	✓	✓	✓	1,344	0
Lowe (2021)	✓	✓	✓	✓	7.5	14
Elwert et al. (2023b)	✓	✓	✓	✓	178.5	30
Elwert et al. (2023a)	✓	✓	✓	✓	420	90
Mousa (2020)	✓	✓	✓	✓	28	75/82/120/180
Friedman et al. (2024)	✓	✓		✓	1,040	240
Freddi et al. (2024)	✓	✓	✓	✓	0.25	26
Ghosh (2023)	✓	✓	✓	✓	768	30
Baseler et al. (2023)	✓	✓		✓	1	0/300
Bezabih et al. (2024)	✓	✓	✓	✓	0.25	0
Dahl et al. (2021)	✓	✓	✓	✓	1,344	300
Clochard (2022)	✓	✓	✓	✓	0.17	0
Loiacono and Silva-Vargas (2023)	✓	✓	✓	✓	42	30/210
Greene et al. (2024)	✓	✓	✓	✓	0.17	0/21
Clochard et al. (2023)	✓	✓	✓	✓	0.17	0/30
Abril et al. (2023)	✓	✓		✓	0.25	0/15
Burlacu et al. (2024)	✓	✓	✓	✓	0.83	18
Barros (2024)	✓	✓	✓	✓	3	2.5/75
Chaudhry and Hussain (2024)			✓	✓	2	30
Ghosh et al. (2024)	✓	✓	✓	✓	48	42/380
EGAP Registry						
Scacco and Warren (2018)	✓	✓	✓	✓	64	30
Broockman and Kalla (2016)		✓	✓	✓	0.17	3/21/42/90
Grady et al. (2023)	✓	✓	✓	✓	0.25	0
Zhou and Lyall (2023)	✓	✓	✓	✓	540	38/195
Kalla and Broockman (2020)		✓	✓	✓	0.08	4
		✓	✓	✓	0.17	7
Paler et al. (2020)			✓	✓	1	0
Asimovic et al. (2024)	✓	✓	✓	✓	13	0
Rossiter (2023)			✓	✓	0.13	0
Rossiter and Carlson (2024)					0.13	0/3
Porat et al. (2024)	✓	✓	✓	✓	2	10
	✓	✓	✓	✓	4	21
Adamu et al. (2024)	✓	✓	✓	✓	8	120
Mousa et al. (2024)	✓	✓	✓	✓	30	0/14/21

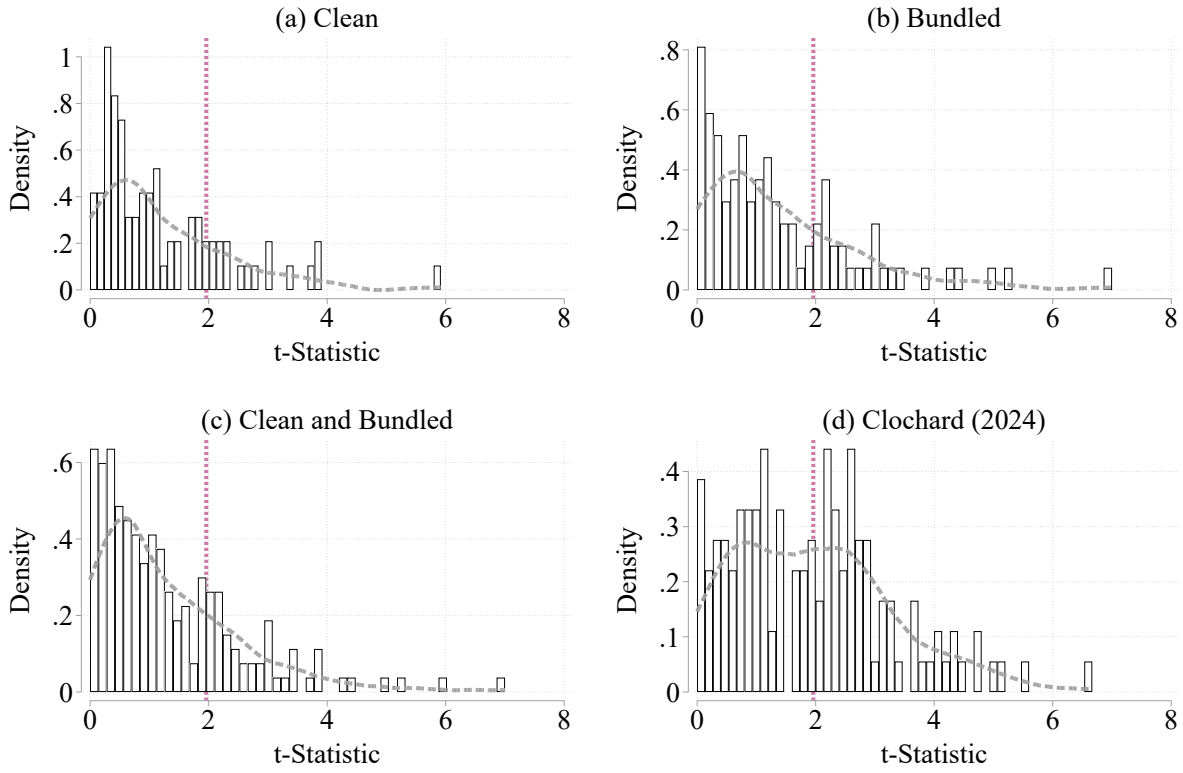
Notes: The papers are sorted by the date of the initial pre-registration. The table shows whether the intergroup contact in each experiment satisfies the four conditions of Allport (1954), using the authors' assessment where possible. Hours denotes an estimate of the number of hours of interaction with the outgroup in each experiment. Measurement Timepoints denotes the best estimate of the number of days that had elapsed after the contact intervention ended when outcomes were measured, taking the median where available.

Table A2: The Effects of Bundled vs. Clean Interventions

	Effect Size			
	(1)	(2)	(3)	(4)
Bundled Outgroup Contact vs. Control	0.03	-0.01	0.01	-0.02
	(0.04)	(0.02)	(0.04)	(0.06)
Observations	178	178	178	178
Clean Effect Mean	0.08	0.08	0.08	0.08
Experiment FE	No	No	Yes	Yes
Experiment-Outcome-Group FE	No	No	No	Yes
Weighted	No	Yes	Yes	Yes

Notes: The unit of observation is the treatment effect. The sample includes only the effects of clean contact and bundled outgroup interventions. Observations are weighted by the inverse of the standardized standard error in columns 2 to 4. Standard errors are clustered at the experiment-level in columns 1 and 2, and otherwise are robust. Experiment-by-outcome-by-group fixed effects are experiment fixed effects interacted with the outcome variable, and the group the effect was estimated for (e.g. a group could be immigrants). *** p<0.01, ** p<0.05, * p<0.1.

Figure A1: Little Evidence of Heaping of t-Statistics



Notes: The figure uses effect-level data to plot histograms and kernel densities of different sets of effects. Panel (a) includes only the clean effects of high versus low or no intergroup contact (e.g. assignment to mixed-gender versus single-gender teams). Panel (b) includes only the effects of bundled outgroup contact relative to a pure control group. Panel (c) includes both sets of effects. For comparison, Panel (d) includes the effects studied in Clochard (2024), in which effects were not required to be pre-registered as primary outcomes (data available here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TRZUBI). In Panel (d) I do not show three outlying observations with a t-statistic above eight. The vertical dashed line is at 1.96, the critical z-score value for statistical significance at the 5% level.

Figure A2: Random Effects Forest Plot: Clean, All

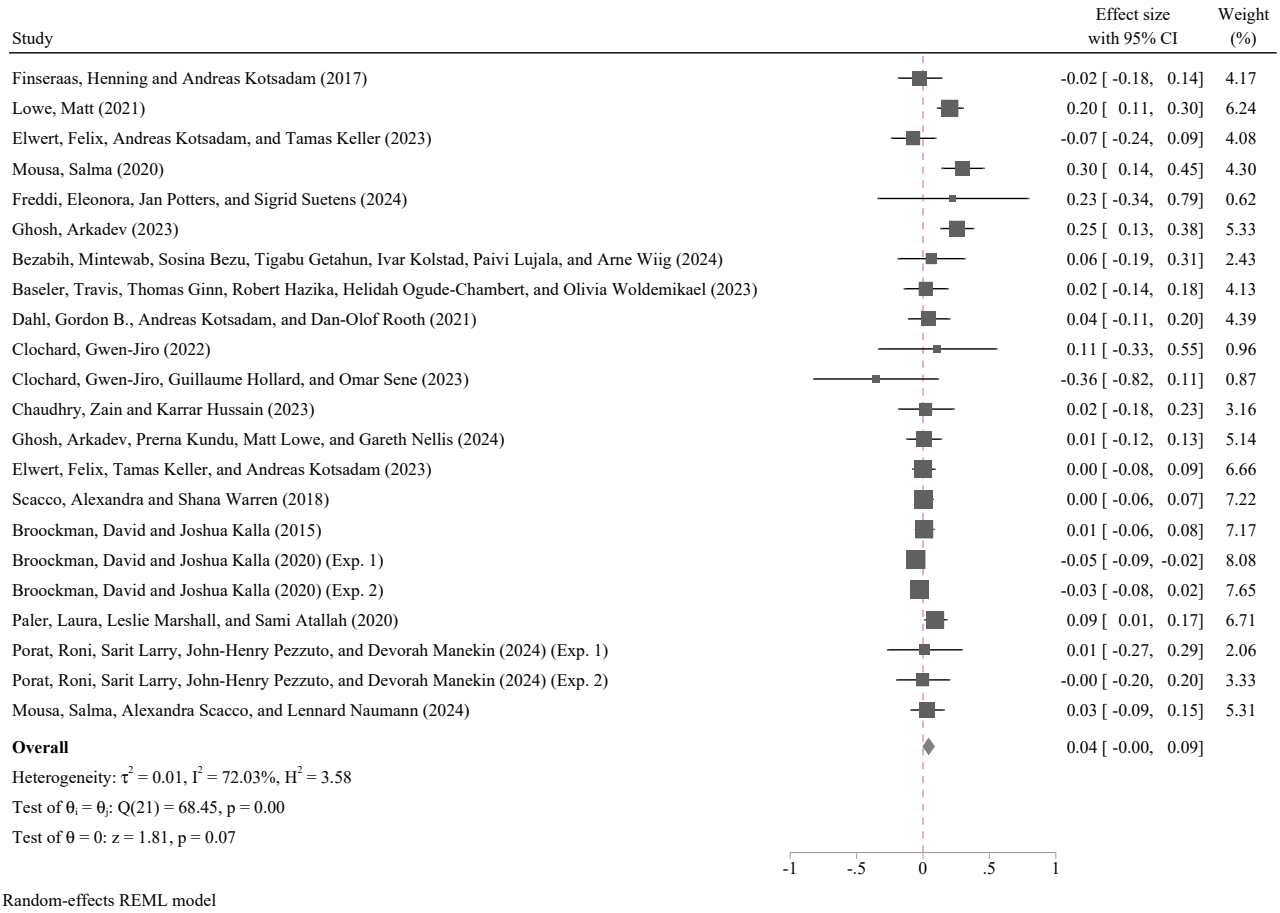


Figure A3: Random Effects Forest Plot: Clean, Generalized

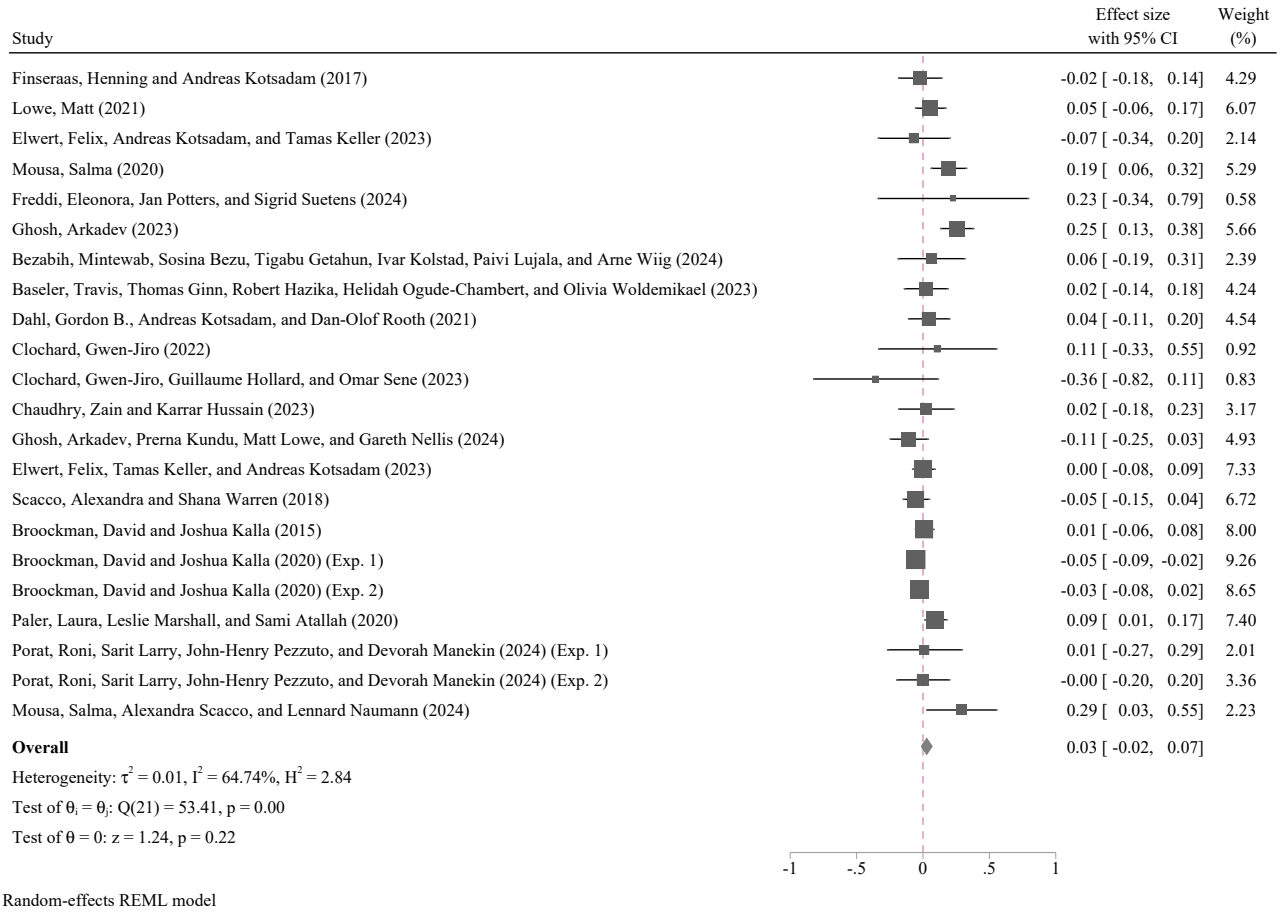


Figure A4: Random Effects Forest Plot: Clean, +Optimal

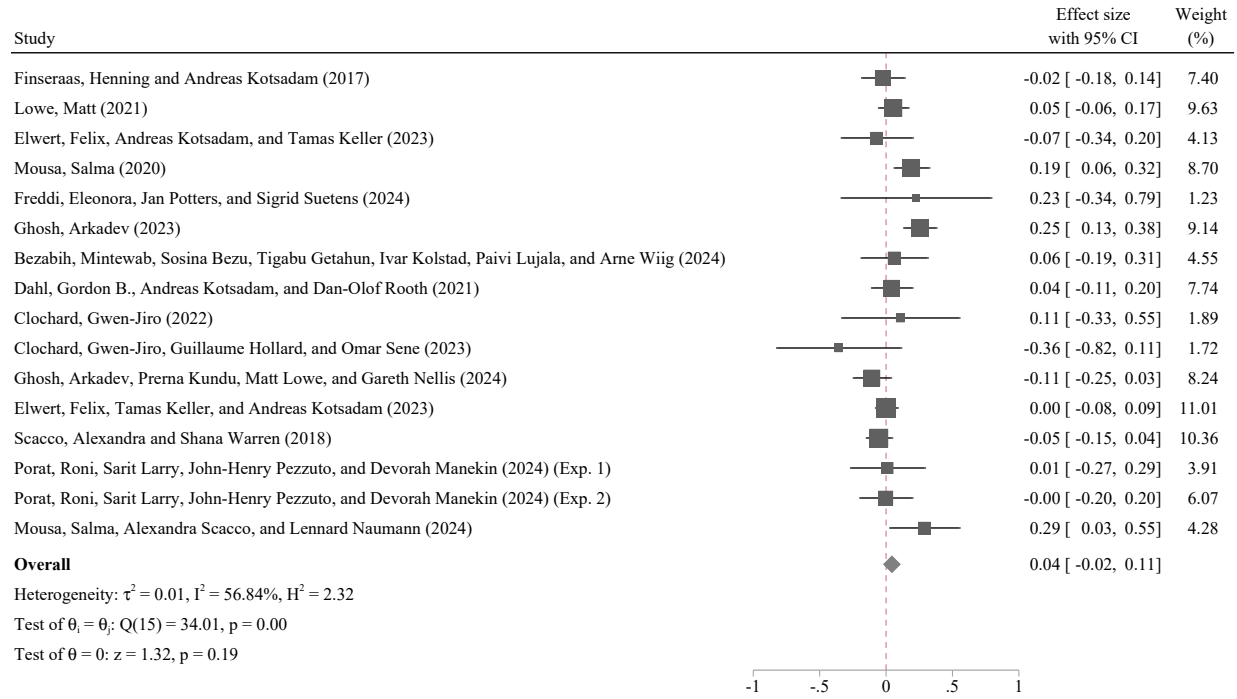
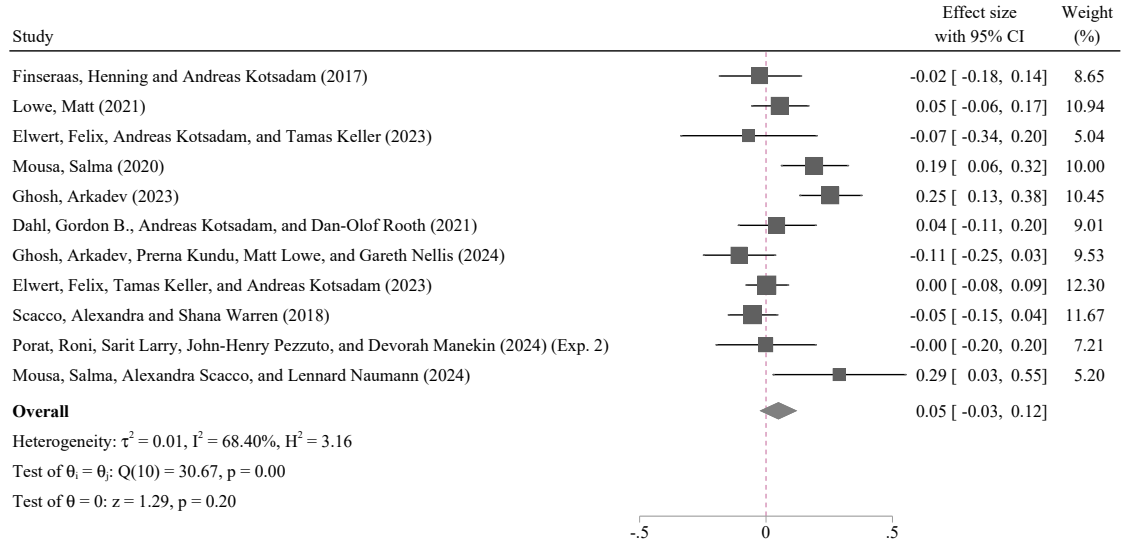
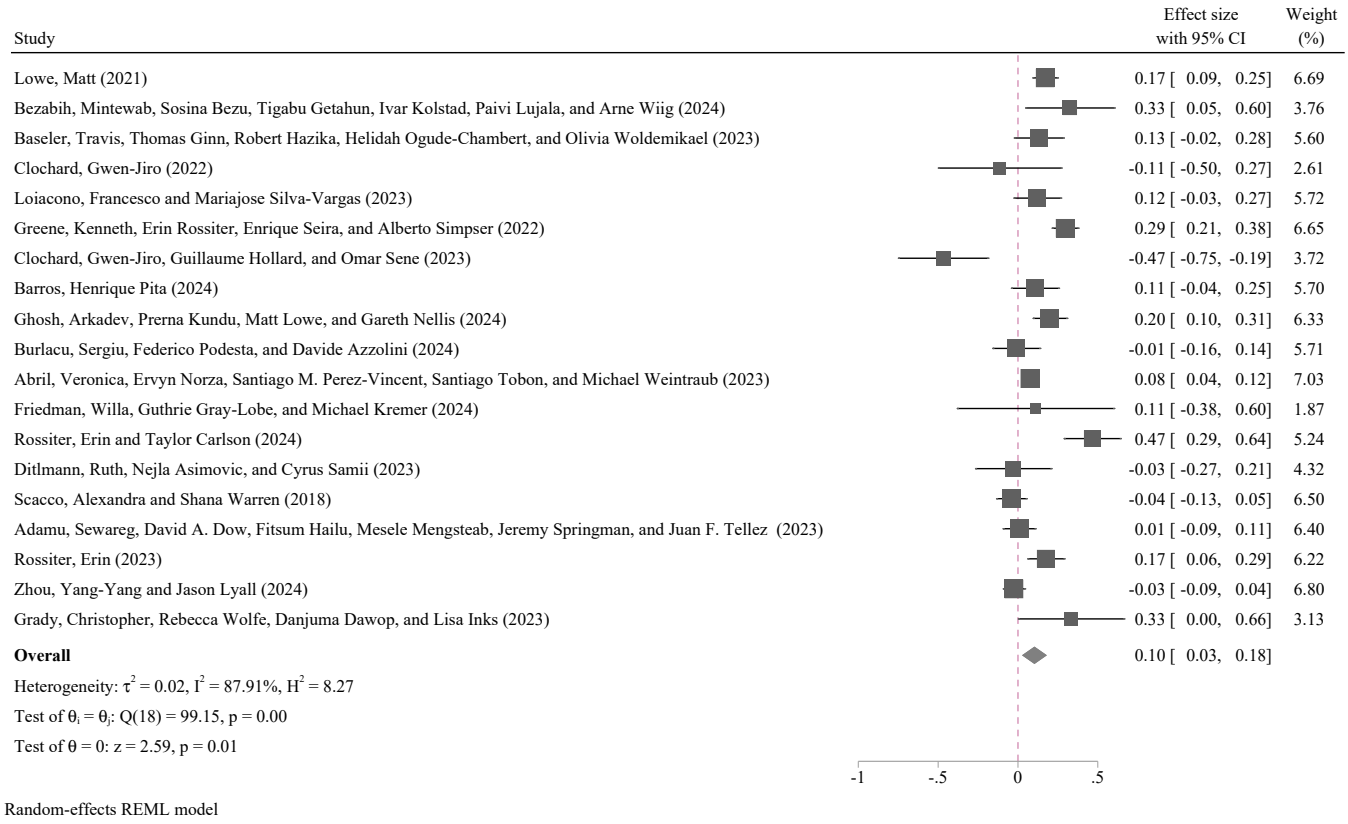


Figure A5: Random Effects Forest Plot: Clean, +Long



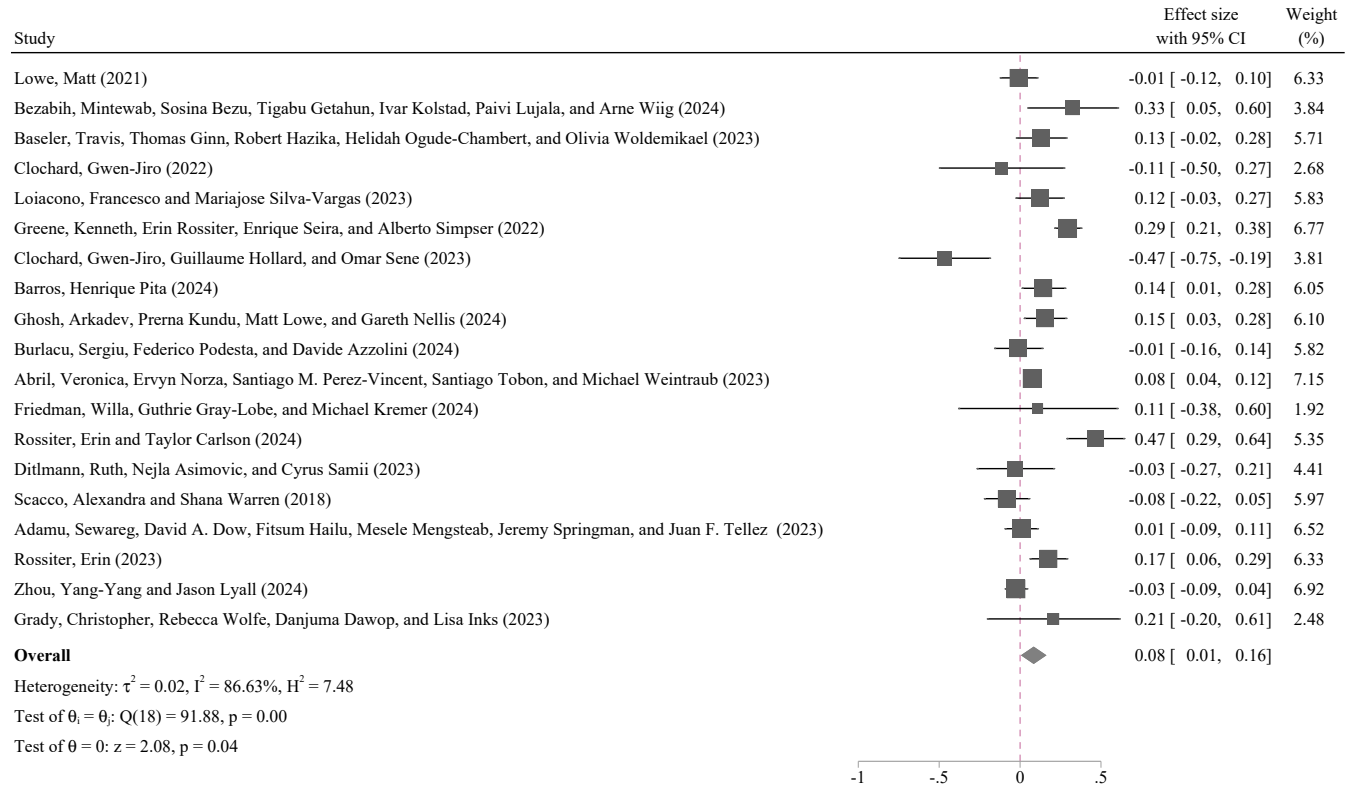
Random-effects REML model

Figure A6: Random Effects Forest Plot: Bundled, All



Random-effects REML model

Figure A7: Random Effects Forest Plot: Bundled, Generalized



Random-effects REML model

Figure A8: Random Effects Forest Plot: Bundled, +Optimal

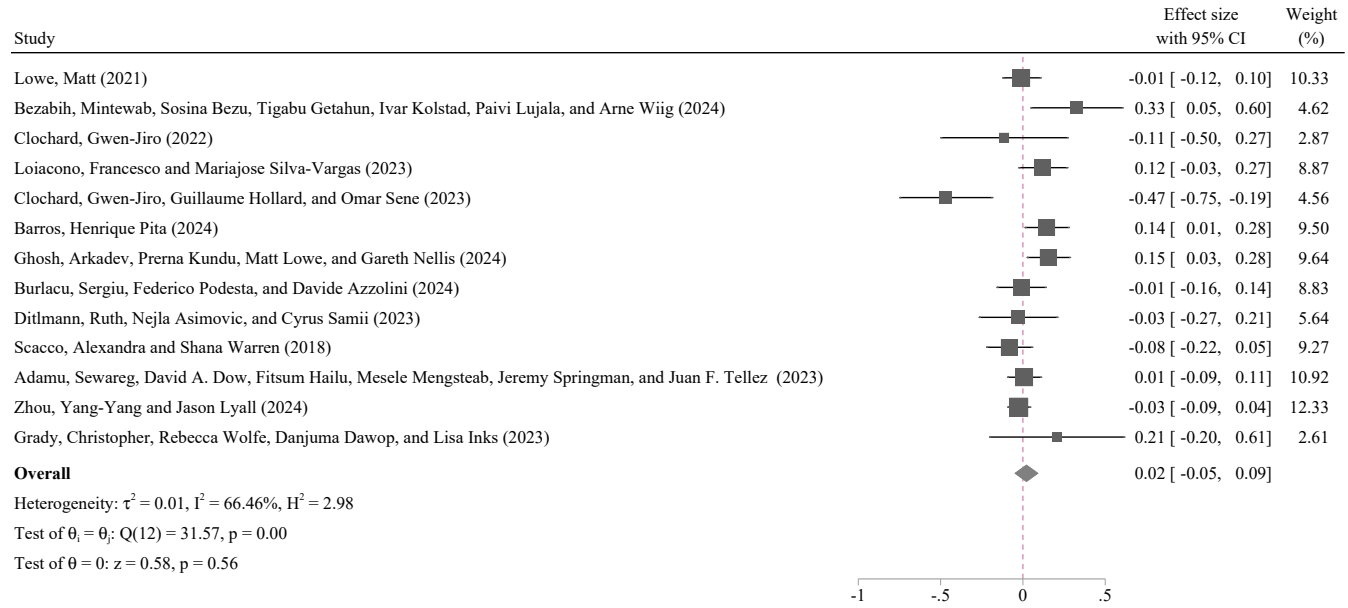
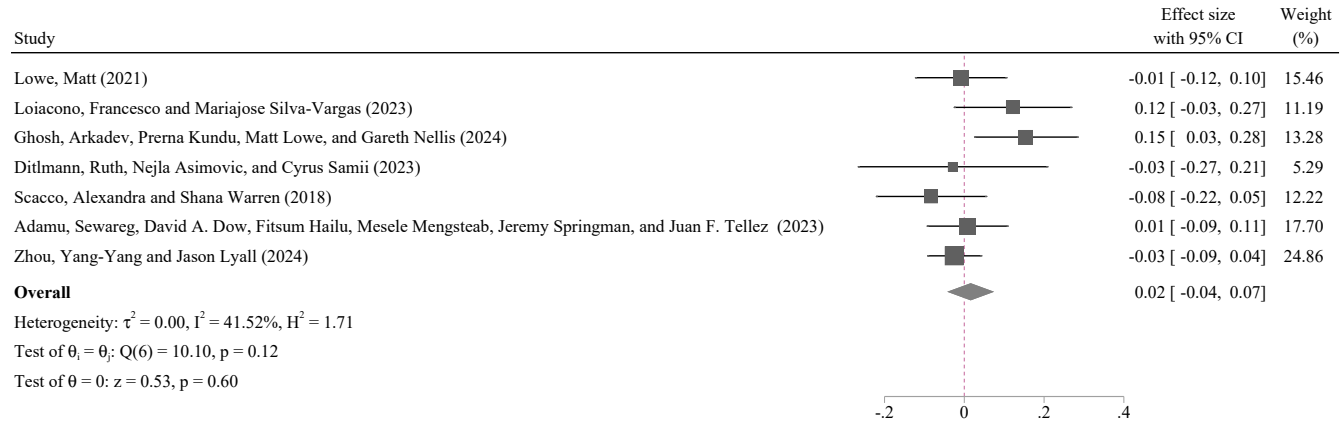
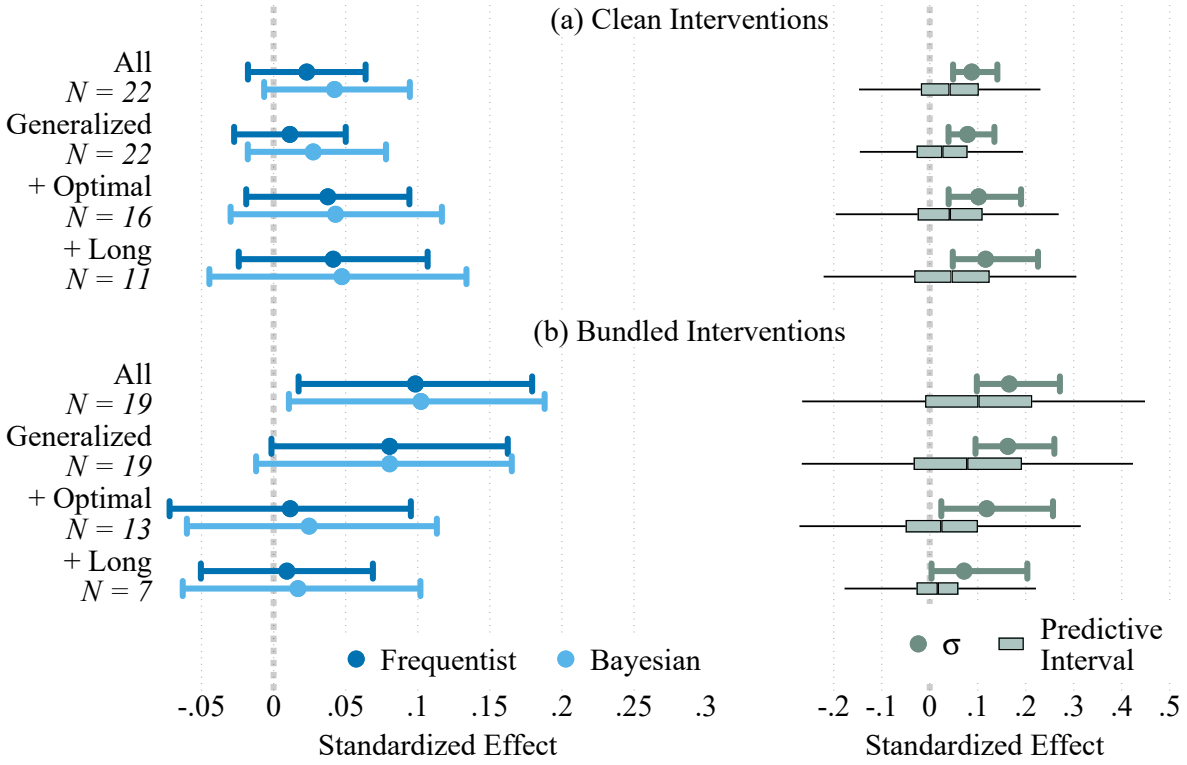


Figure A9: Random Effects Forest Plot: Bundled, +Long



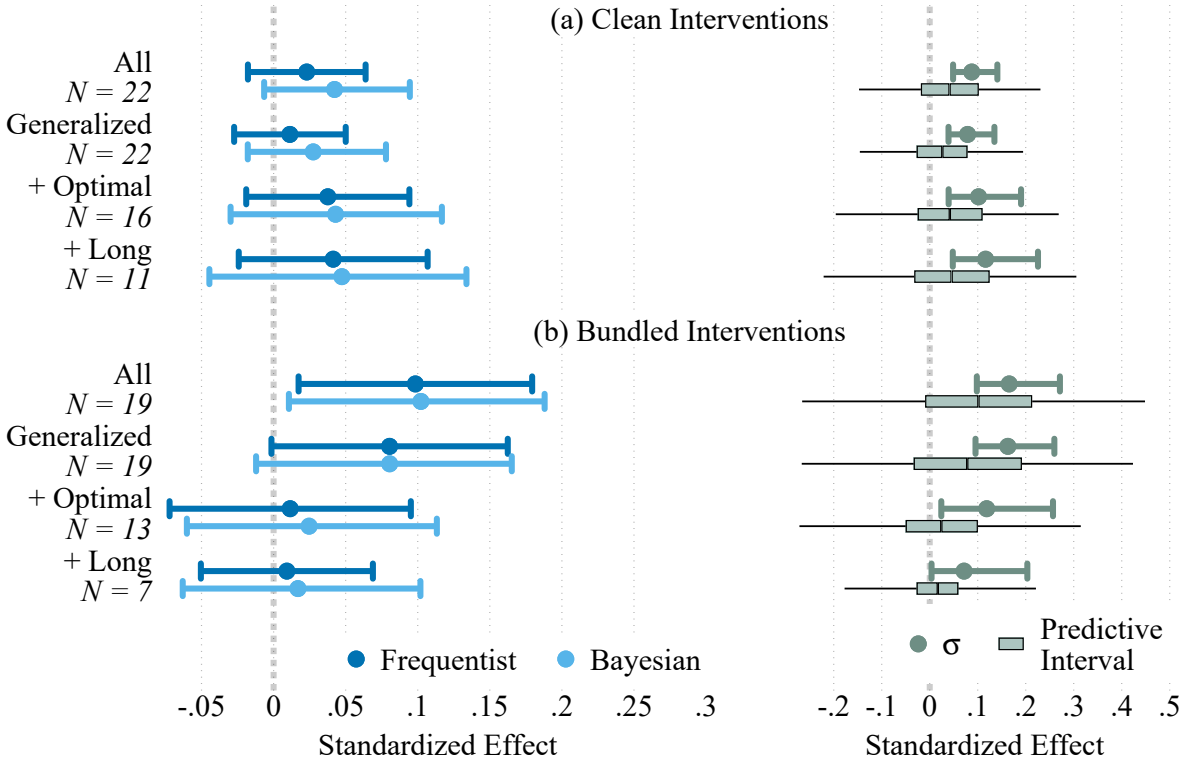
Random-effects REML model

Figure A10: Effects of Intergroup Contact Interventions (ρ=0)



Notes: The figure replicates Figure 2, but with effect sizes collapsed to experiment-level assuming a correlation between an experiment's effect sizes of zero (rather than 0.12).

Figure A11: Effects of Intergroup Contact Interventions (ρ=0.8)



Notes: The figure replicates Figure 2, but with effect sizes collapsed to experiment-level assuming a correlation between an experiment's effect sizes of 0.8 (rather than 0.12).

A Additional Meta-Analysis Details

A.1 Secondary Coding Rules

Selecting and Coding Outcomes. The core rule I follow is to identify primary outcomes related to prejudice and intergroup relations as specified in pre-registrations. If outcomes are pre-specified but with no distinction as to which are 'primary', I include all relevant outcomes. If intergroup relations outcomes are only included as 'secondary' outcomes, I include these outcomes (as with Zhou and Lyall (2023)).

In rare cases there are intergroup relations-related outcomes for which there is no clear directional prediction or for which it is not clear which direction is more inclusive/good, I exclude the outcome from the analysis.

Where an index outcome includes only pre-registered components, I record the treatment effect on the index outcome as opposed to the separate effects on each component, where possible. For coding whether an index outcome is generalized to the outgroup: if it has at least one component that is generalized and one that is not, I code the index as generalized if the majority of the components are generalized.

Selecting Specifications. When effects are reported with and without controls, and the pre-registration does not specify which was to be preferred, I use the specification without controls (given that the experimenter has additional degrees of freedom when deciding which controls to include).

When a paper reports both ITT and IV effects and the pre-registration does not specify which is to be preferred, I opt for the ITT effects, to avoid relying on an exclusion restriction.

A.2 Collapsing Effects to Experiment-Level

My approach results in a dataset of 191 treatment effects and standard errors (all standardized) across 34 experiments. I collapse these estimates to experiment-level effects and standard errors before carrying out the meta-analysis.

First, I keep only the relevant treatment effects for a given analysis. For example, in one analysis I keep only the effects of clean variation in contact, leaving me with 81 treatment effects across 22 experiments.

Second, I collapse this data to the experiment-by-endpoint level, where an endpoint is a given outcome measured at a given follow-up (e.g. immigration attitudes measured during a two-week follow-up survey). There are two cases of an experiment having more than one treatment effect for a given endpoint: it could be that the treatment effect was recorded in the paper separately for two subgroups (e.g. for locals and for immigrants), and not reported for the full group (otherwise I take the pooled estimate). Or it could be that the experiment has two separate clean contact comparisons, with no ex ante reason to pick one over the other (e.g. Scacco and Warren (2018) assigns participants to heterogeneous versus homogeneous classrooms, and within heterogeneous classrooms to outgroup or ingroup partners).

To collapse the effect size, I take the simple mean by experiment-endpoint. To collapse the standard error, I use the following formula (see chapter 24, Borenstein et al. (2021)):

$$SE_{ep} = \sqrt{\left(\frac{1}{N}\right)^2 \Sigma_{i=1}^{N} SE_{ep,i}^2} \tag{6}$$

where e denotes the experiment, p denotes the endpoint, N is the number of treatment effects to collapse for experiment-endpoint ep, and $SE_{ep,i}$ is the standard error for treatment effect i for experiment-endpoint ep. This formula assumes independence between the treatment effects we are collapsing, which is appropriate given that these treatment effects are for different populations. To take an example, if a given experiment-endpoint has an effect estimated for locals and immigrants, with a standard error of 0.1 in each case, the collapsed standard error is $\sqrt{0.25 \times 2 \times 0.1^2} = 0.071$. The collapsing in this case gives roughly a 30% reduction in uncertainty, which is essentially the same reduction we would have had the authors ran the pooled regression including locals and immigrants together.

In the final step, I collapse to the experiment-level. The typical case is that we have multiple endpoints measured for the same set of people. I again collapse the treatment effect by taking the simple mean. But to collapse the standard error, we now have to adjust for the fact that the effect sizes estimated for the different endpoints are not independent, since they are estimated on the same people (and one would typically expect a positive correlation between effect sizes when the outcomes are different but related facets of prejudice and intergroup relations). The formula for the collapsed standard error is now (Borenstein et al. 2021; Lund et al. 2024):

$$SE_e = \sqrt{\left(\frac{1}{M}\right)^2 (\Sigma_{i=1}^{M} SE_{ep}^2 + \Sigma_{i \neq j} \rho_{ij} SE_{ei} SE_{ej})} \tag{7}$$

where M is the number of endpoints for experiment e, and ρ_{ij} is the correlation coefficient between effect sizes for endpoints i and j. If the correlation is zero, the effects are independent, and the formula becomes identical to the case above. If the correlation is one, two effects give no more information than one. Applying the example above, with two endpoints with a standard error of 0.1, we will calculate $SE_e = 0.1$. The correlation is typically unknown, but should be expected to lie between zero and one – effects on two related measures of prejudice should give more information than just one effect, but not as much as if the two effects were estimated for two non-overlapping populations.

To calibrate a plausible measure of the correlation coefficient, I use data from the two experiments of mine that enter the meta-analysis (Lowe 2021; Ghosh et al. 2024). I then find the ρ_{ij} that delivers the same standard error reduction as achieved by running a regression of an index variable (a simple average of standardized components) on treatment. I estimate the correlation to be 0.2 for Lowe (2021) and 0.03 for Ghosh et al. (2024). I take the average of the two, giving me $\rho_{ij} = 0.12$ (to two decimal places). I use this calibrated correlation coefficient for all experiments and endpoints.

While this calibration is somewhat arbitrary, it dominates setting a correlation of zero (in which we would certainly overestimate precision) or one (in which we would certainly underestimate precision). In addition, the fact that the calibrated ρ is close to zero suggests that my approach is more likely to overestimate precision rather than underestimate precision. Marginally increasing the confidence intervals on the estimates of the average effects of contact in Figure 2 would not substantively change the findings.

A.3 Paper-by-paper Coding Description and Judgment Calls

This section describes the process by which my research assistant Catalina Garcia Valenzuela and I identified and recorded effect sizes and standard errors for each of the 34 eligible experiments. We also give details on other coding decisions – e.g. on assessing which Allport conditions are met by each experiment, recording the duration of contact, recording whether specific findings were reported in the abstract of the paper, etc.

A.3.1 AEA Registry

A.3.1.1 Finseraas and Kotsadam (2017) Does personal contact with ethnic minorities affect anti-immigrant sentiments? Evidence from a field experiment

Pre-registration link.

Outcomes. The AEA pre-registration says "Attitudes toward welfare dualism" for Primary Outcomes (end points), and "See the full pdf document." for Primary Outcomes (explanation). The latter refers to a pre-analysis plan that was posted Sep 12, 2014. The PAP says "The pre-analysis plan is archived before the second wave of data is collected." – take this to mean that the primary outcomes were pre-specified prior to analysis (although it is not 100% clear without a definition of "wave").

PAP mentions three main outcome variables of interest, but only the first is about welfare dualism, the other two are framed as about mechanisms. So I conclude that there is just one pre-specified primary outcome, based on this question:

"Do you agree or disagree with the following statements: Refugees and immigrants should not have the same rights to social assistance as Norwegians. 1= Strongly agree 2= Agree 3= Neither agree nor disagree 4= Disagree 5= Strongly disagree."

This variable is recoded into *"Immigrants same rights"*: 4 and 5=1, 1 to 3=0.

Treatment. They randomized soldiers within platoons into rooms, such that the treatment group is soldiers with an ethnic Norwegian background who were randomized into a room with at least one soldier with an ethnic minority background, and the control group consists of soldiers who did not share a room with an ethnic minority soldier (page 706; definition of minority below).

Allport Conditions. All conditions satisfied; from page 705: "We conducted an explicit test of contact theory and its relevance for support for welfare dualism. We ran this test as a field experiment in the military, which provides an institutional context where the specified conditions for contact to improve tolerance are fulfilled. Soldiers of private rank have equal

social status within the army, they share the common goals of the unit, they need to cooperate to solve their tasks and contact takes place in the context of an explicit, enforcing authority. Moreover, the army explicitly promotes views of unity and equality among soldiers of the same rank. Thus, contact theory should operate in this context."

Duration of Contact. From pg706: "The assigned room is where they live for the eight weeks of the recruit period." Given that it is a fully immersive experience, we will consider the full length of contact.

Days Since Contact Ending and Measurement. From pg707: "We surveyed the soldiers for the second time at the end of the recruit period so we have pre- and post-treatment data on the outcomes." Code this as zero days.

Reported in Abstract. "The study finds (...), but small and insignificant effects on support for welfare dualism".

Specification. From the PAP: "The main independent variable is a dummy variable which equals one if there is at least one person with at least one parent born in a non-Western country." This is followed in the published paper (see pg708). PAP says that they will control for the outcome measured at baseline (pg10), platoon FE, and they say they will "only include control variables for which treatment and control differ and results with and without these controls will be presented." Since with and without control specifications are pre-registered with no priority of one over the other, we choose the results from the specification without controls (following instruction 2b).

This means we pick Table 2, column 1, Panel A (pg713): b=0.038, se=0.085. SD deviation of this outcome is 1.1 (reported on pg713). However: on second glance, the variable used in Table 2 is categorical, rather than being recoded to the binary variable (and the PAP is fairly clear that the primary outcome is the recoded variable, not the categorical variable). To find the effects on the recoded variable, we need to turn to Table A5, column 1, Panel B (pg9 of the online appendix). Here we have b=-0.012, se=0.041, N=534. Standard errors are clustered at the room level, and there's no information on the number of clusters, but "rooms vary in occupancy between 3 and 12 persons, but 73 per cent of our sample lived in six-person rooms" (page 707), so we estimate the number of clusters is 826 (total N) over 6 (138 rooms).

SDs. Standard deviations obtained from authors.

A.3.1.2 Lowe (2021) Types of Contact: A Field Experiment on Collaborative and Adversarial Caste Integration

Pre-registration link.

Outcomes. Primary outcomes pre-specified before intervention start date. They are:

- Individual and Collective (Team) Voting on Field Trip Participants (for individual-level caste preference and its translation to group decisions)
- Caste Implicit Association Tests (for subconscious caste biases)
- Trading Exercise (for caste bias in willingness to interact/cooperate for economic surplus)
- Future Team Choice (for caste bias in willingness to interact/network formation)

- Trust Games (for generalized caste bias).

Caste IATs were dropped from the paper (justification in deviations from pre-registration section); doesn't make sense to estimate outgroup exposure effects on collective team voting (because outgroup exposure at individual-level, but collective voting at team-level). Leaves us with the following specific outcome measures:

- Individual voting on field trip participants (effects on caste bias)
  - Table 4, column 5, -0.13 (0.07) (reverse-coded in google sheet), N=9180 (clusters/individuals = 751)
- Cross-caste trading
  - Table 5, column 2, 0.06 (0.04), N=1,510 (clusters/teams = 160)
- Number of other-castes chosen for future team
  - With stakes: Table 3, column 1, 0.71 (0.12), N=768 (clusters/team = 160)
  - There is also a no stakes version (Table 3, column 2, 0.45 (0.11)), which is nearly identical. We keep the stakes version only given that this is in some sense "more incentivized", and given that this is the outcome that is reported also for the additional bundled comparisons.
- Amount given in trust game (effects on caste bias; played with real stakes)
  - Table 6, column 2, -0.26 (1.48) (reverse-coded in google sheet), N=2253 (clusters/individuals = 751)

Treatment. Participants are randomly assigned to homogeneous-caste teams or mixed-caste teams. We record the effect of collaborative contact, as was pre-registered.

Allport Conditions. The intervention had no equal status: "Together the data suggest that different castes do not enjoy equal status on each team, but rather reflect the status hierarchy of the caste system itself" (page 1841). From page 1817: "Surveyors paid players on Individual Pay teams according to individual performance (giving on-team inequality) while players on Team Pay teams were paid based on team performance (giving on-team equality). The variation in incentives allows a test of an additional Allport (1954) condition: that of intergroup cooperation. Team Pay increases intergroup cooperation on each team, by making own pay depend positively on the performance of teammates. In contrast, Individual Pay increases competition, giving incentives to "jockey for position" to ensure enough play-time to make money." However, even with individual pay, it is essentially still a cooperative task, so will still infer as cooperation with a common goal. Intervention has support from authorities (inferred).

Duration of Contact. Each team plays 8 matches (page 1817), about 5-10 hours total.

Days Since Contact Ending and Measurement. Endline 1 at 1 week and Endline 2 at 3 weeks, we use 14 days for all outcomes.

Reported in Abstract. "Collaborative contact increases cross-caste friendships and efficiency in trade, and reduces own-caste favoritism." Then in Table 1 we see that voting outcome's concept is caste favoritism, and both trading and trust outcomes are about efficiency.

Specification. Pre-registration doesn't give details on exact specification. So we follow the instructions: (i) highest collaborative contact versus no collaborative contact (i.e. the "beta" coefficient as per the paper's specification), (ii) strata fixed effects but no controls. We code the effects of collaborative, but not adversarial, contact (following rule 2c).

SDs. I take SDs from my own analysis data. Control SD is SD for those in homogeneous-caste teams.

<u>Additional Bundled Effects:</u> Outgroup contact bundled versus pure control and Ingroup contact bundled versus pure control

Treatment. The experiment also has the bundled treatment of participating in the league versus control group (backup players). Participants in the control group played as backup players but a protocol was followed such that only high-priority players (based on random priority number) played frequently (page 1816). "The low-priority backups are close to a pure control group given that they played on average only 1.6 matches each, compared with 6.1 matches for league players" (page 1838). Then, we obtain two additional comparisons: the comparison between low-priority back ups and those assigned to (1) mixed-caste teams (outgroup contact) and (2) homogeneous-caste teams (ingroup contact).

Reported in Abstract. "League participation reduces intergroup differences, suggesting that the positive aspects of intergroup contact more than offset the negative aspects in this setting." We consider the outcome as reported in the abstract if the effect for the comparison is statistically significant.

Specification. This comparison uses the full sample and regresses the outcome on separate indicators for homogeneous teams, mixed teams, and high priority back ups, such that each gives the effect relative to low-priority back ups. Standard errors are clustered at the team level for teams, and to the participant level otherwise.

Results are in Figure 5 plus in-text point estimates. Results are not available for all outcomes, but since this is not the main focus of the paper they will not be considered as "Not in the paper". Matt obtained point estimates, standard errors, number of clusters, and sample size.

For sample size and N clusters relevant treatment arms, from page 1808 we know: there were 1261 boys in the total sample, 800 assigned to play, and 65% of that assigned to mixed teams (520), and only 3 of the boys that didn't play are high priority back ups. So, for the comparison between mixed teams and back ups, the sample size for the relevant treatment arms is 520 + 458. For the comparison between homogeneous teams and back ups, it is 280 + 458. Since the regressions with the full sample don't have N = 1261, I rescale. The number of clusters is 104/56 mixed/homogeneous teams, plus 458 low priority back ups.

SDs. There's no SD for the pure control group, so we use the SD from the previous comparison.

A.3.1.3 Elwert et al. (2023b) Rearranging the desk chairs: a large randomized field experiment on the effects of close contact on interethnic relations

Pre-registration link.

Outcomes. The AEA pre-registration says "Lend to Classmate and Roma friend" for Primary Outcomes (end points). The pre-registration is from March 2018 and the intervention enddate is May 2018, so we can assume that the primary outcomes were pre-registered before analysis, plus they reassure us that they haven't received endline data yet in the PAP (page 5). The explanations for these outcomes are, respectively, "Lend to Classmate is based on a survey experiment where students were presented with a scenario where they could lend money to a classmate" and "Roma friend captures whether an individual has a Roma friend among his or her best friends."

For Roma friend, it is marked as *Includes specific people met*, because the question doesn't seem to exclude the deskmate and because if they have a Roma friend it is most likely a classmate. For lending, the question explicitly excludes deskmates, so counts as *Generalized to outgroup*.

Treatment. They randomized the seating chart in classrooms. The sample is non-Roma students; the first treatment variable, *RomaDeskmate*, equals 1 if a student is randomized to sit next to a Roma deskmate at the beginning of the fall semester and 0 otherwise (page 1823; use for *Roma friend* outcome). The other relevant treatment (used for *Lend to classmate outcome*) is *RomaDeskmate x RomaVignette*, which equals 1 if a student sitting next to a Roma deskmate is asked to lend money to a Roma classmate and 0 otherwise (page 1824).

Allport Conditions. All conditions satisfied; from page 1818: "We argue that deskmate exposures to ethnic others, even more so than grade-level exposures, best fit the scope conditions of contact theory by Allport (1954)"

Duration of Contact. From page 1834: "for the duration of one semester (five months)". The intervention is only for 3 subjects, that account for 7-10 hours per week, depending on the grade level (page 1822). Our best guess is 21 weeks*8.5 hours = 178.5 hours.

Days Since Contact Ending and Measurement. For days between contact and measurement, the paper and appendix are not explicit about it. For the intervention they encourage adherence until January 2018. And then "Data collection will conclude in April of 2018. The research team will receive outcomes data in May, 2018" (page 5 PAP). Authors clarified over email that the intervention ended at the end of January 2018, and that endline collection began in February 2018 and ended in April 2018. We then assume the median measurement timepoint of March 2018, so 30 days.

Reported in Abstract. "The authors find that neither manipulating the closeness of interethnic exposure between students within classrooms nor variation in ethnic composition across grade levels affects antiminority discrimination." We consider both outcomes to fall under "antiminority discrimination".

Specification. PAP says that for *Roma Friend* they will control for a set of individual level control variables either measured at baseline or reflecting stable characteristics, with class fixed effects, and that they will present results with and without the baseline controls but the main specification is without controls (page 8).

For *Lend to Classmate* the specification includes treated dummy, whether the questions

asked if they would lend money to a Roma classmate or not, the interaction between the two, controls, and class fixed effects. They don't mention in the PAP that it will be done with and without controls, or which would be their preferred specification, but they do in the paper, specifying that without controls is preferred (page 1829).

Results are obtained from Table 3 in page 1831. We use the results without controls following rule 2b.

SDs. SDs are obtained from Table 1 in page 1826.

A.3.1.4 Elwert et al. (2023a) Effects of deskmate gender on confidence, attitudes toward mixed gender teams, and prejudice - Evidence from a large scale field experiment in Hungarian schools

Pre-registration link.

Outcomes. Paper states that PAP was registered before any data was received (page 5). 3 main outcomes according to pre-registration, their explanations are in the PAP (obtained from the paper's appendix) and align with the primary outcomes in the paper (page 6).

1. Prejudice: believes the other sex to be inferior than what it actually is (i.e. in mathematics). This outcome is tested for boys only.

2. Preference for mixed teams: This outcome is tested for boys only.

3. Confidence: Categorical variable based on comparison between self assessment of math ability and their position in the baseline class grade distribution. This outcome is tested for girls only.

Treatment. Treatment corresponds to being randomly assigned to a deskmate that is of the other sex (same sex is control group; page 5).

Allport Conditions. From page 3: "Those in contact should have equal status in the particular context, share common goals, be in a cooperative context, and the contact should take place under some form of authority and have a high degree of friendship potential. It is an open question how gender peer exposure plays out in more common and less streamlined conditions." While the paragraph above is ambiguous on whether the authors believe or not that all conditions are met, we can infer that all conditions are met given the nature of classroom work and the author's description in their other 2023 paper.

Duration of contact. Intervention starts at the beginning of the fall semester 2017 and is encouraged until the end of the semester in January 2018 (page 4). Based on the author's other deskmate paper (see "Rearranging the desk chairs" above), I'll assume 5 months. Students sit together for around 20 hours per week (page 14). So our best guess is 21 weeks*20 hours = 420 hours.

Days Since Contact Ending and Measurement. For the intervention they encourage adherence until January 2018, and data collection will conclude in April 2018 (page 6 PAP). Will assume 90 days.

Reported in Abstract. "We find that boys (wrongly) perceive boys to be better at mathematics, are generally overconfident, and are less likely than girls to think that mixed gender teams

perform better. Girls on the other hand are more likely to be underconfident. In a large pre-registered field experiment, where we randomly assign deskmates in Hungarian schools, we test whether such attitudes change with mixing sexes at the deskmate level. We find that deskmate composition does not affect any of these outcomes." The underlined concepts correspond to the outcomes (1)-(3).

Specification. Their preferred specification is without controls. They include class fixed effects and use robust standard errors in all estimations (page 11). Results are obtained from columns 1-3-5 in Table 4 (page 12).

SDs. Standard deviations for the full sample are obtained from Table 1 (page 10) from the Boys and Girls columns, respectively. SDs for the control are not available.

A.3.1.5 Mousa (2020) Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq

Pre-registration link. We use the available data.

The paper under EGAP corresponds to the first wave of this AEA-registered paper. This paper uses the data from that first wave (experiment ran between March and June 2017), plus a second wave, adding new outcomes, as detailed below. Most likely the data from the first wave had already been analyzed by the time of pre-registration, but we will include outcomes that were measured in this first wave nonetheless, since it seems that data from the second wave had not been analyzed at the time of pre-registration.

Outcomes. There are 2 versions of the AEA pre-registration, but they are 2 minutes apart and the only change is a link to an EGAP registration. The PAP under AEA ("PAP-update") refers to an original PAP ("PAP") under the EGAP registration 20170603AA, which was the PAP for the pilot experiment. "The scale-up includes a larger sample size, the addition of a comparison group, and an expanded range of outcomes. The experimental protocol in the original PAP holds unless otherwise noted." Thus, I'm using both PAPs to find the pre-registered outcomes.

The study was done in two waves. The first wave is from March 2017 to May 2017 (page 30 supplementary materials). And the second wave is from September 2018 to November 2018. The first PAP (the one in the EGAP registration) is from June 2017, however, the EGAP registration specifies that the registration is prior to realization of outcomes'. The updated PAP under the AEA registration is from November 2018, no explicit mention to whether there has been analysis of data for the second wave, but I'm assuming the data has not been analysed yet. However, most likely the data for the first wave had already been analyzed, since the EGAP paper is from December 2018.

The behavioural outcomes in the original PAP are:

- Volunteer to join mixed team next season (Register for mixed team in the future in the paper)
- Attend Ramadan BBQ (Attend mixed dinner event in the paper)
- Bringing family to BBQ (conditional on BBQ attendance). Dropped since it is conditional and can't interpret causally.

- Continuing to play soccer with Muslims (Train with Muslims at least once a week in the paper).

I categorised *Register for mixed team in the future* as *Generalized to outgroup* given that this question is about general willingness to be assigned to a mixed team roster in future (as opposed to, for example, reporting that you would like to continue with the same team). I categorize *Train with Muslims at least once a week* as *Includes specific people met*, given that it could potentially include the same players (although in practice, the paper reports that it is not only this: "The training outcome does not merely capture the inertia of continuing to play with teammates: 15% of treated teams recruited Muslim players from other teams in the league or from the neighborhood.")

The updated PAP adds 4 new behavioral outcomes:

- Players' households patronizing Muslim (Christian)– owned businesses, and the money spent conditional on attending (*Patronize restaurant in Mosul* in the paper or *Visit Mosul* in supplementary materials, ends up being operationalized as dummy for attending the restaurant).
- Female household members signing up for a mixed community program. I could not find it in the paper or any explanation for why it's missing.
- Player's vote for the best new player (*Vote for Muslim player to receive sportsmanship prize* in the paper)
- A survey item asking respondents to choose between donating to a charity that benefits only Christians, only Muslims, or both communities ("We will donate $1 to a charity that you choose. Which charity should we donate to on your behalf?") (*Donate $1 survey compensation to church versus neutral nongovernmental organization* in the paper).

Coded *Vote for Muslim player to receive sportsmanship prize* outcome as *Generalized to outgroup* given that it excludes teammates (page 2).

The attitudinal outcomes in the original PAP are 3 indices, but then they get updated to 5 different ones in the updated PAP:

- Coexistence, combines different outcome types:
  - Believe that Sunni Arabs are welcoming toward Christians (explicit evaluations).
  - Proud or very proud to be Iraqi (political and cultural attitudes).
  - Agree that I share a lot in common with Sunni Arabs (explicit evaluations).
  - Feel comfortable going to "different" neighborhoods in my town (indirect measure).
- Prospects for Peace (called National Unity in paper), includes:
  - Believe that Iraq would be a better society if Iraqis treated each other as Iraqis first.
  - Believe that dividing Iraq into ethnic and religious groups is arbitrary.

- Tolerance, combines different outcome types:
  - Believe that most Sunni Arabs disapproved of ISIS (explicit evaluations).
  - Describe current friendship group as mixed (behavioral measures).
- Muslims as Neighbors. In the paper it has one more item than specified in the PAP, but it seems to be a super minor difference. It asks them how comfortable they would be with neighbors from different religions.
- Absolve Muslim Civilians of Blame (called Blaming Muslims in the paper): Believe that Shi'ite Shabak or Sunni Arabs are responsible for Christian suffering.

Coexistence and *Tolerance* get dropped from the paper because they yielded a low Cronbach's alpha of 0.2 or below (this method was not pre-registered). 2 out of the 4 pre-specified items of the *Prospects for Peace (National Unity)* index also get dropped for that reason, but we still consider the outcome pre-registered before analysis (trusting the authors' justification for changing what is in the index slightly). Compare attitudinal outcomes listed in updated PAP (page 8) to attitudinal outcomes in the paper (Table 1, page 2).

Treatment. Participants were randomly assigned to an all-Christian soccer team or a team with 3 Muslims (page 1; sample in main analysis is Christians).

Allport Conditions. All conditions satisfied; from page 1: "The leagues largely met the conditions considered key for activating successful intergroup contact: Teammates had to cooperate to achieve their shared goal, players were subject to the equalizing effect of team sports, and local Christian leaders and organizations endorsed the leagues."

Duration of Contact. From page 1: "for a 2-month league". Players reach a total of about 26 hours on the pitch, plus 0 to 4 hours extra depending on the team's performance (page 8 supplemental materials). So the best guess is 26 hours plus 2 hours (midpoint between 0 and 4).

Days Since Contact Ending and Measurement. According to PAP, the endline survey (i.e. where all attitudinal outcomes plus some behavioural outcomes are obtained) was going to be completed on the day the league ended (i.e. 0 days after contact ending, page 5 original PAP), but the paper says that the attitudinal indices are measured 2 weeks to 5 months postintervention (page 2). Couldn't find more information or explanation. For all outcomes measured at the endline, I assign them the median between 14 and 150 days (= 82).

Attend mixed dinner event was supposed to be measured 2 weeks after the league ends (page 5 PAP), but the paper shows between 3 weeks and 5 months (page 2, median = 64 days). Supplemental material explains that it was 3 weeks for the first wave of the study and 4 months for the second wave. Roughly one third of the sample was in the first wave, so the median is 4 months (= 120 days).

Train with Muslims at least once a week was measured 6 months after the intervention according to the paper (page 2) although PAP specified it would be 2 months (page 5).

Patronize restaurant in Mosul is measured 1-4 months postintervention according to paper (page 2), which lines up with the vouchers being valid for 3 months according to the updated PAP (page 4). Using 2.5 months as the median (= 75 days).

Reported in Abstract. "The intervention improved behaviors toward Muslim peers: Christians with Muslim teammates were more likely to vote for a Muslim (not on their team) to

receive a sportsmanship award, register for a mixed team next season, and train with Muslims 6 months after the intervention. The intervention did not substantially affect behaviors in other social contexts, such as patronizing a restaurant in Muslim-dominated Mosul or attending a mixed social event, nor did it yield consistent effects on intergroup attitudes." Only outcome not mentioned above is the donation outcome.

Specification. *"For the main analysis, I estimated the average treatment effect on each behavioral outcome and attitudinal index, controlling for randomization block and other baseline covariates while clustering standard errors at the team level."* (page 3). This specification lines up with what was specified in the updated PAP (page 1).

Point estimates, sample size, and number of clusters are obtained from Table S-1 (page 29 supplementary materials; there seems to be a typo in the sign of the difference for the donation outcome). Lines up with values presented in page 3 of the paper for behavioral outcomes, and the value for *National Unity* in text. Point estimates for behavioural outcomes are in percentage points, and on standard deviations for attitudinal outcomes.

I find standard errors in the available data because the standard errors reported in brackets look like p-values (less than 1 and strongly correlated with the non-bootstrapped p-values). The relevant analysis is in the paper's replication package, *main-analyses.R* lines 168-200.

SDs. SDs are obtained from Table S-6 in page 32 of supplementary materials (multiplied by 100 for behavioral outcomes since the results for these outcomes are in percentages but the SDs are not). I used the SD in *t1* where both *t0* and *t1* are available. Full sample SD is not available but they do report Treated SD.

A.3.1.6 Friedman et al. (2024) Worker Assignment and National Unity: Are All Stable Matches Socially Stable?

Pre-registration link.

Outcomes. Pre-registration is from December 2018 (no changes to outcome in future versions of the pre-registration) and surveys were conducted in 2019 so outcomes are pre-registered before analysis.

Outcomes in the pre-registration are:

1. Volunteer retention in the program.
2. Literacy and numeracy outcomes of primary students.
3. Volunteer national unity attitudes and behaviors, specifically in terms of inter-ethnic prejudice, levels of identification (ethnicity vs. national) and national pride.
4. Host school and community national unity attitudes and behaviors.
5. Job performance of volunteers in terms of volunteer productivity, and satisfaction with the G-United program on the part of school officials.

However, the pre-registration states that for our comparison of interest, only a subset of outcomes will be analyzed: "In addition to comparing retention, job performance, and measures

of national unity under the random assignment procedure to those under deferred acceptance, we will also compare a subset of outcomes of those assigned through either mechanism to a pure control group. A randomly selected subset of applicants will not be assigned at all, but will instead be placed in a control group, allowing examination of the effect of program participation on applicants' interethnic prejudice, levels of identification (ethnic vs. national) and national pride."

This only includes outcome (3) above, which appears in the paper as *National Unity Index:* "This index is constructed as the first principal component of responses to a feeling thermometer (a set of indicators from a ten-point scale) [explicit evaluation], reporting of ethnic compared with national identity (a set of indicators from a five-point scale) [political and cultural attitudes], agreement with the statement, *It is valuable to interact with people from other ethnic groups* [political and cultural attitudes], and an indicator for whether the respondents' first two reported friends share their mother tongue [behavioral measure but not behavior/incentivized since it is self-report] (page 12)."

Treatment. We will compare people that participated in the program vs pure control. "Greatness-United (GUnited) was a high-prole program of the Government of Kenya that recruited and deployed recent graduates of Kenyan universities to low-performing public schools to administer an early-grade literacy program with an explicit goal of promoting national unity through inter-regional contact. (page 1)" Page 12 describes random assignment; "We estimate the effect of participating in the G-United program using random variation in the timing of offers induced by a random lottery number. As part of the assignment process, all applicants were given a random lottery number, which determined the round in which an applicant received an offer, and subsequently influenced the likelihood of attending training."

Allport Conditions. No discussion in the paper. We can infer that the contact intervention is supported by authority (i.e. government program), with cooperative contact and a common goal (i.e. students learning), but there's no equal status given the different status of teachers vs students.

Duration of contact. Volunteers served for 6 months (page 34). No additional information on weeks or hours, best guess will be 40 hours * 26 weeks = 1040 hours.

Days Since Contact Ending and Measurement. Service ended in October 2018 and participants were contacted for follow-up survey starting June 2019. So 8 months, 240 days.

Reported in Abstract. No mention to National Unity outcome or our comparison of interest.

Specification. We use the 2SLS estimates since those actually use random assignment: "We use two measures of program participation: attendance at the mandatory training (short-run) and completion of service (long-term). We examine effects on social attitudes and county of placement using OLS and 2SLS using the randomly assigned lottery number that increased the likelihood of being assigned early as an instrument. Applicants who attend training score 0.06 SDs higher on an index of national unity attitudes according to OLS estimates, although this difference is not statistically significant (Table 5).31 Those who complete service score 0.16 SDs higher on the national unity index, and this difference is highly statistically significant. These results may reflect selection into participation if more socially motivated applicants are also more likely to participate. However, 2SLS estimates have similar point estimates, but in all cases, they are statistically insignificant due to larger standard errors. (page 12)" The specification includes controls (page 8).

Results are in column 1 of Table 5 (page 27). We use the 2SLS specification for *Attended training*, since the exclusion restriction is more plausible than for *Completed service*.

SDs. Outcome is standardized (see mean equal to 0 in bottom of the table, and in-text results in standard deviations [bottom page 12] line up with coefficients in the table).

A.3.1.7 Freddi et al. (2024) The effect of brief cooperative contact with ethnic minorities on discrimination

Pre-registration link.

Outcomes. Pre-registration lists 2 primary outcomes: A choice in an ultimatum game as responder (i.e. minimum acceptable offer made, page 67), and a choice in a deception game as communicator (i.e. whether or not the student is honest about the outcome of a die roll, page 68). Both games were played with real stakes (page 68). Intervention started on February 14 2019, but outcomes were collected on March 12 2019, after the pre-registration on March 2, 2019 (page 66).

Outcome type is *Generalized to outgroup* for both outcomes because the other player in the game was not someone the participant had interacted with. The ethnic background of the other player was randomly varied and it was either Dutch majority or Moroccan/Turkish descent (page 67). Participants who had a majority partner in the ultimatum game were assigned a minority partner in the communication game, and vice versa (page 68).

Treatment. Treatment is being randomly assigned to an ethnically homogeneous (control group) or an ethnically mixed team (treatment group) for a cooperative task (i.e. participants did this cooperative task and then played the two games to measure the outcomes).

Allport Conditions. All conditions satisfied; from page 75: "First, intergroup contact that satisfies Allport's contact theory conditions (personal interaction, shared goals, a common project, equal status, and approval by a recognized authority) has been shown to have long-lasting effects in other contexts. (...) Also, the fact that the intervention subjects did not have a lower social status than the to be studied subjects in our experiment may have facilitated cooperative and common-goal-oriented interethnic contact and may thus have been a necessary condition for reduced discrimination among the group without prior classroom contact." We can infer support from authority.

Duration of Contact. From page 67: "After listening to an introductory lecture on the topic, the teams were given 15 minutes to discuss and prepare their answer."

Days Since Contact Ending and Measurement. From page 66: "The intervention phase took place on February 14, 2019, and was designed to bring the high school students into contact with peers who are members of an ethnic minority. In the decision-making phase, which took place on March 12, 2019, the students participated in two economic games designed to elicit their attitude toward ethnic minorities."

Reported in Abstract. "We find that, overall, students did not discriminate and that participation in an ethnically mixed team did not have an effect on their behavior." According to the paper, the games measured discrimination: "...we measured ethnic discrimination among the participants. To do so, we had the high school students make choices in two behavioral games..." (page 65).

Specification. Registration doesn't specify how the data will be analyzed. Thus, I'm using the specifications without controls; column 1 of Table 2 (page 71) for the ultimatum game, and column 1 of Table 3 (page 73) for the communication game. In both cases, the outcome of interest is the interaction between the dummy for being treated and the dummy for having a minority partner in the game, *Minority x Treated*. Both outcomes are reverse signed.

SDs. Standard deviations provided by the author. We use the SD of the whole control group (ethnically homogeneous teams).

A.3.1.8 Ghosh (2023) Religious Divisions and Production Technology: Experimental Evidence from India

Pre-registration link.

Outcomes. The PAP registers 3 categories of primary outcomes. There are several versions of the pre-registration, but the primary outcomes and their explanation stay fixed, so they are registered before analysis.

1. High frequency daily team level output data obtained from the firm: Production Data will be obtained directly from the firm which is recorded daily at the firm. Such data are recorded primarily at the line level and not at the section level. Production data for some sections are available which will be used. In addition, during the course of the intervention, production supervisors will record their own assessment of each section's performance daily.

2. Different measures of implicit and explicit out-group perceptions and prejudice across treatment arms: Measures of out-group perception and prejudice constructed using survey data in order to analyze whether contact with individuals from other religious groups affect preferences and how these effects vary across sections that require high and low dependency amongst workers. There will be direct measures of social distance, for example survey questions on cross-religion communication at the workplace and preferences towards having non-coreligionist supervisors.

3. Workers will participate in lab-in-the-field experiments at endline in order to understand mechanisms that result in productivity differences in religiously mixed and homogeneous teams: An Implicit Association Test (IAT) where workers associate identifiable Hindu and Muslims names with positions in the hierarchy at the firm to understand whether workers have bias towards their own religious group occupying higher positions.

For (2) and (3), the paper says (page 26): "The endline survey focused on two main sets of outcomes: 1. Those that capture actual interactions between workers during production and 2. Attitudes towards non-coreligionists co-workers." For each set, the paper further specifies that there are 3 main outcomes each (1-3 in the list below for the first set, and 4-6 for the second set).

I compare the pre-registration to what I found in the paper, and establish that the pre-registered primary outcomes are:

1. Identified teammate as contributing low effort: If a worker identifies his teammate to have not contributed to the team as much as other workers did, or to the extent that is expected, then this outcome is coded 1. We drop this outcome because it's mechanically affected.

2. Blamed by teammate: The outcome variable is coded 1 for teammates who have blamed the respondent at least once during the intervention period. We drop this outcome because it's mechanically affected.

3. Sacrificed (or willing to) relief time for teammate: The outcome variable is coded 1 for teammates that workers already have, or are willing to give up their relief time for. We drop this outcome because it's mechanically affected.

4. Taking Orders: Workers were asked if they are comfortable taking orders from non-coreligionists. Considered an explicit evaluation since they are directly asked about the outgroup.

5. Communicating: Question is whether they find interacting or communicating with non-coreligionists (in general) as comfortable as co-religionists.

6. Co-working: If they prefer to be in mixed or all-Hindu groups if teams were to change again in the future. The outcome is incentivized since workers were told that their answers would be kept in mind for future team changes.

7. IAT.

Outcomes at the line level or line-section level will not be included, because the outcome is mechanically affected by the intervention and not necessarily by contact.

Page 82 lists differences between the pre-registration and the paper and explains why the IAT doesn't appear in the paper: "While the output data were obtained as planned, at endline only a short phone survey could be conducted due to COVID-19 related restrictions. Therefore, survey questions were restricted to those on inter-group relations at the workplace only. Furthermore, the endline IAT could not be conducted either (there was one conducted at baseline)."

Treatment. Treatment is a dummy variable coded 1 if the line-section-level team is religiously mixed, and 0 otherwise.

Allport Conditions. From the paper we can conclude that contact is cooperative: "I use naturally occurring variation in types of collaborative contact (due to production function differences)" (page 6). Furthermore, we can consider that it meets the other three conditions since workers have a common goal of production, equal status within their teams, and support from authorities (e.g. supervisors).

Duration of Contact. According to page 18, the experiment was conducted between November 2019 and March 2020. From the author: Intervention was from the end of November to the end of March, so 4 months. So that is 16 weeks at about 48 hours of contact max (depending on product demand, storage, etc). We will use this upper bound of 768 hours as our best guess.

Days Since Contact Ending and Measurement. Outcomes are obtained at the endline, which I assume to take place 30 days after the intervention, since the endline was collected between April and May 2020 (page 18).

Reported in Abstract. "Despite lowering short-run productivity, mixing improves out-group attitudes for Hindu workers in high-dependency tasks, but there are little or no effects in low-dependency tasks." As detailed in the outcomes section above, outcomes (1)-(3) are about interaction between workers and (4)-(6) are about attitudes, so the latter set is reported.

Specification. The pre-registration doesn't specify how the data will be analyzed.

I select the specification with *Mixed* rather than the ones that look at the interaction between treatment and LD/HD. Point estimates, standard errors, and sample size are obtained from columns 1-3-5 in Table 6 (page 28) for worker interactions, and columns 1-3-5 in Table 7 (page 30) for attitudes. For Identified teammate and Blamed teammate, I reverse the sign. For two attitudinal outcomes (4-5 above), the relevant question is also asked at baseline and is included as a control. I can't find the number of clusters specifically for these outcomes, but I'll assume that all 113 line-section-level teams were included.

SDs. Control groups SDs obtained from the author.

A.3.1.9 Baseler et al. (2023) Can redistribution change policy views? Aid and attitudes toward refugees in Uganda

Pre-registration link.

Outcomes. The outcomes were pre-registered before analysis: Endline data was collected 2021 inwards, and their registration and updated PAP is from October 30, 2020.

According to the PAP, the two main primary outcomes are summary indices of support for inclusive hosting and business outcomes. The PAP was not very detailed on how the index would be constructed, but it seems to come from the items under Domain 1 (page 5 PAP), and according to the paper, "Each summary index represents a weighted average of standardized components within a domain" (page 16). I'm only using the summary index, and not the components, as a primary outcome, in addition to business profits (i.e. self-reported business profits minus the additional value of unpaid family labor used for the business, page 5 PAP).

The intervention consisted of a mentorship program that matched business owners with experienced refugee business owners in the same sector. Table B11 shows the number of mentorship meetings held by year across Refugee and Ugandan Mentorship arms (page 22). The table shows that not everyone had the same number of meetings, so it seems that the degree of contact varied across participants. The average number of meetings is 2.3 (obtained from Table B11). There's no evidence that the number of meetings was balanced across treatments. However, it seems that the number of meetings may have had no impact for either treatment (Table C31, online appendix). Since I'm using the comparison between being mentored by a Ugandan vs Refugee, I'm assigning High vs no outgroup contact as the comparison type.

I also assume that the Yarid facilitator was randomly Ugandan or Refugee across treatments, especially because it seems that there's an interaction between being mentored by a refugee and having a refugee facilitator (Table A5, page appendix).

Treatment. Treatment group is being mentored by a refugee and the control group is being mentored by a Ugandan. Experimental sample is Ugandans.

Allport Conditions. From page 8: "Our project experimentally induced short term, collaborative contact through a mentorship program and builds on this literature by comparing the effects on

attitudes to programs focusing on economic incentives." We can also infer that participants had a common goal (i.e. business success) and are supported by authority. There's no equal status given the mentor-mentee interaction.

Duration of Contact. The intervention started in January 2020 but then stopped in mid-March 2020. They resumed and delivered the intervention from March to May 2021. Thus, I'm assuming the duration of contact was 5 months (2.5 months before the pandemic and 2.5 months after, although it may have been a little longer/less). I'm not considering the pause in the interventions in the duration of contact, or that people had a different number of meetings.

Participants could have up to 10 meetings (up to 6 in 2020 and up to 4 in 2021). Before the pandemic, the program included up to six in-person meetings between the mentor and mentee, roughly once per week, each facilitated by a YARID staff member who provided guidance and translation if necessary (page 13). After the pandemic, mentorship meetings went from in-person to remote, and YARID provided up to four facilitated mentorship meetings regardless of the number of meetings that were held prior to COVID-19 (page 14). Before the pandemic conversations lasted an average of 43 minutes, compared to 23 minutes after (footnote 21, page 14). As mentioned above, the mean number of meetings was 2.3. Out best guess is then 30 minutes*2.3 = 69 minutes, rounded to 1 hour.

Days Since Contact Ending and Measurement. They conducted a midline survey over the phone in October 2020 (0 days since intervention ongoing), plus three additional follow-up surveys after interventions were completed: a phone survey in August 2021 (90 days), and two in-person surveys in May 2021 (0 days) and March 2022 (300 days). Note of Figure 2 (page 21) clarifies that the index measure was not collected on the second phone survey (i.e. August 2021), but was collected otherwise. Nothing suggests that the profit measure was not collected in all surveys. The PAP specifies that they will run additional specifications that allow for time-varying effects with separate coefficients for the number of months since the treatment (page 16 PAP), which reflects on Figure 2 for the index (but there's no table for it in case we wanted to use that estimation; there's only C31 with number of meetings). Furthermore, their specification's outcome is for individual i at time t, meaning that they pool results across different time points.

Reported in Abstract. "We find minimal impacts of intergroup contact, implemented as business mentorship by an experienced refugee." No mention to the specific outcome where they find the minimal effects, but we consider both reported as there's minimal effects on both.

Specification. They estimate ITT effects with an ANCOVA specification, with the value of the outcome at baseline, baseline controls, treatment assignment dummies , survey round fixed effects, and strata fixed effects. Standard errors are clustered at the individual level.

The specification in the PAP lines up with the one in the paper (page 15), assuming that the errors clustered at the enterprise level means the same as the individual level. Results for *Integration Policies Index* are in column 5 of Table 2 (page 20), and results for *Profit* are in column 1 of Table 5 (page 26). Point estimates are obtained from subtracting the point estimate in *Mentored by Ugandan* to the one in *Mentored by Refugee*, and the p-value is obtained from *R-Mentee = U-Mentee* in order to infer the standard error. We obtain the t-statistic from the p-value and degrees of freedom, and then we divide the point estimate by the t-statistic to obtain the standard error.

For the degrees of freedom, the sample size is 3051 (and 4029 for profit). Our best estimate of degrees of freedom is 3015 (and 3993), but since N is over a hundred, the degrees of freedom

won't make much of a difference.

Number of clusters is not specified in the table, but I assume it's the number of enterprises = 1406 (page 11).

For *Sample Size for relevant treatment arms*, the full experimental sample is 1406 Ugandan microenterprise owners and the index (profit) outcome is measured 3 (4) times. However, 1406*3(4) does not equal the sample size in the table, probably because of endline attrition. Thus, I multiply the *Number of Clusters for relevant treatment arms* (169 mentored by Refugee and 168 mentored by Ugandan; in Table B10) by the average number of observations per enterprise in the table. For index outcomes this is (169+168)*(3051/1406), and for profits outcome it is (169+168)*(4029/1406).

SDs. Index is standardized, so the full sample SD is 1. Control groups SDs obtained from the author.

Additional Bundled Effects: Outgroup contact bundled versus pure control and Ingroup contact bundled versus pure control

Treatment. For the comparison between outgroup contact versus pure control, we can compare those mentored by a refugee to those in the pure control group. For the comparison between ingroup contact versus pure control, we can compare those mentored by a Ugandan to those in the pure control group.

Reported in Abstract. "We find minimal impacts of intergroup contact, implemented as business mentorship by an experienced refugee."

Ingroup versus pure control comparison not in the abstract.

Specification. Same as above. We obtain results from Tables 2 and 5, row *Mentored by Refugee* and row *Mentored by Ugandan*.

The *Number of Clusters for relevant treatment arms* is 169/168 (mentored by refugee/Ugandan) plus 265 (pure control). For *Sample Size for relevant treatment arms*, I multiply the average number of observations per enterprise with the number of enterprises: (169+265)*(3051/1406) and (168+265)*(3051/1406).

SDs. Will use SDs from previous comparison.

A.3.1.10 Bezabih et al. (2024) Inter-group interaction and attitudes to migrants

Pre-registration link.

Outcomes. Pre-registration is from January 10, 2020 and intervention started January 25, 2020, so outcomes are pre-registered before analysis. There's only one outcome in the pre-registration, *Attitude to migrants*, and details obtained from paper fully line up with PAP (page 7 paper and PAP page 4). The variable is an index from responses to the questions below by summing the responses (treating Don't Know as missing), and rescaling the sum to a number between zero and one (higher value is higher agreement):

1. "To what extent do you agree with the following statement: 'Refugees who are currently living in refugee camps in Ethiopia should be allowed to freely work and live outside of the camp.'"

2. To what extent do you agree with the following statement: 'Refugees who are currently admitted to Ethiopia should be allowed to settle in my home community permanently if they are not able to return to their home country.'"

3. "To what extent do you agree with the following statement: 'If given a chance to settle, a refugee can be as good a citizen as someone who is born locally.'"

Since (1) and (2) are about policies, outcome type is *Political and cultural attitudes*, but (3) is *Explicit evaluation*.

Treatment. The relevant sample is host community members (there's results for migrants but the pre-registered outcome is for hosts), which were randomly assigned to four treatment groups and a pure control. In the first three treatments, host community members were paired with a randomly selected migrant from a refugee camp to play an incentivized guessing game. The framing of the game was neutral in the first treatment. In the second and third treatments, they subtly introduced an economic framing and an ethnic identity framing to the guessing game, respectively. The fourth treatment paired host community members with other host community members to play the neutral version of the game. Host community members in the pure control group did not interact with anyone, instead proceeding straight to the survey posed to all the other treatment groups post-interaction (page 4). The relevant (unbundled) comparison for us are those that played the neutral game with a migrant (treatment group) vs with another host (control group). Comparison type is *High vs no outgroup contact*.

Allport Conditions. All conditions satisfied; from page 12: "The guessing game was included and designed to make the interaction adhere as closely as possible to the first three of the Allport (1954) conditions; equal status, common goals of the interaction, and cooperation rather than competition." We conclude that it is also supported by authority.

Duration of Contact. All treatments involved 15-minute interaction; 10 minutes informally, 5 minutes playing the game (page 11).

Days Since Contact Ending and Measurement. Immediately post-interaction (page 10).

Reported in Abstract. "However, we see similar effects on attitudes to migrants in the treatment where hosts interacted with other hosts, suggesting that the effects are driven by human interaction in general, rather than by interacting specifically with a migrant." Result for our comparison of interest.

Specification. Treatment effects are estimated through ordinary least squares (OLS) estimation (with robust standard errors), with dummy variables indicating each of the 4 treatments. The estimated coefficients thus capture the mean attitudes in the treatment groups compared to the host control group (page 15). They pre-specified the model with and without controls, without stating a preference, so we will use the one without controls. Plus, for the effect of intergroup contact, they test the effect of the host-migrant neutral treatment relative to that of the host-host neutral treatment (i.e. $\beta T1 > \beta T4$) with a pre-specified one-sided t-test and using the null hypothesis $\beta T1 = \beta T4$ (page 16).

Results are obtained from Table 1 (page 23). To obtain the point estimate we subtract the point estimate in *Treatment host-host neutral* from the point estimate in *Treatment host-migrant neutral*. We back out the standard error (with the same method as described before) using the p-value reported in-text (0.316 one sided so we don't divide it by 2; page 22), the sample size

(N = 600), and the number of covariates in the regression (k = 5, including intercept). For sample size relevant treatment arms we add the N in each treatment, following Figure 1 (page 10): There's 120 host participants in each treatment.

SDs. Standard deviation for full host sample is obtained from Table A3 (page 3 appendix).

Additional Bundled Effects: Outgroup contact bundled versus pure control and Ingroup contact bundled versus pure control

Treatment. For the comparison between outgroup contact versus pure control, we compare the effect of neutral game played with a migrant (first treatment arm) to the pure control (no game played). For the comparison between ingroup contact versus pure control, we compare the effect of neutral game played with another (fourth treatment arm) to the pure control (no game played).

Reported in Abstract. Comparison outgroup contact versus pure control: "The results show that interaction with a migrant significantly improved attitudes towards them compared to no interaction."

Comparison ingroup contact versus pure control: "However, we see similar effects on attitudes to migrants in the treatment where hosts interacted with other hosts, suggesting that the effects are driven by human interaction in general, rather than by interacting specifically with a migrant."

Specification. Same as above. Results are obtained from Table 1 (page 23), from row *Treatment host-migrant neutral* and row *Treatment host-host neutral*.

SDs. SD for the full sample as above.

A.3.1.11 Dahl et al. (2021) Does integration change gender attitudes? The effect of randomly assigning women to traditionally male teams

Pre-registration link.

Outcomes. The following paragraph in the paper explained what is covered by the pre-registration (page 15): "In addition to the survey data, we have a set of outcomes collected by the military, including promotion outcomes (after the end of boot camp), occupations (assigned after boot camp) and service evaluations (conducted near the end of service). We likewise add administrative data from the Norwegian registers on education, occupation and workplace characteristics for the years after military service. For confidentiality reasons, the military worked directly with Statistics Norway to create a merged dataset for us. The proposed analyses of these additional data, including coding choices, were described in a pre-analysis plan registered at the AEA RCT Registry (AEARCTR-0005987) before the data was received." Then, footnote on page 15 clarifies that military occupation was not pre-registered: "We did not know that it was possible to merge in military occupations when we wrote up our pre-analysis plan, but added this variable when we became aware of it."

The Primary Outcomes specified in the pre-registration are "Grades and other measures of achievement as well as share of women in field of study or occupation". We obtained the PAP from the authors, which allowed us to establish the outcomes below were pre-registered as primary outcomes.

1. Fraction women in chosen education.

2. Fraction women in chosen occupation.

3. Summary variable of evaluations conducted by the military at the end of mandatory service. This variable is created by averaging the mean of 4 binary variables (equals 1 if the soldier is rated as exceeds expectations or excellent, one for each category of the assessment).

**Promoted to Vice Corporal* is not considered primary because PAP says it's part of "other measures of achievement". According to PAP, their main variable of interest for the long run analysis of the share data was a combination of Share of women in field of study or occupation (page 4 PAP). This is not available in the paper, so we consider (1) and (2) above to be the pre-registered primary outcomes (since it was pre-registered that they would also be analyzed separately).

Treatment. Treatment is being assigned to a squad with women or not.

Allport Conditions. All conditions satisfied; from page 28: "In this setting, men and women were equal in rank and had to complete a similar set of tasks. Moreover, men and women were placed into teams which required cooperation to reach common goals, such as the completion of a training exercise. These features combine to create exactly the type of setting predicted by contact theory to result in changed attitudes." We can infer support from authority.

Duration of Contact. Contact occurs during boot camp, which lasts 8 weeks (page 8). Given that it is a fully immersive experience, we will consider the full length of contact.

Days Since Contact Ending and Measurement. For the outcome on service evaluations, I use 10 months (300 days), since they are conducted near the end of service.

For *Fraction of women in chosen education/occupation*, it seems to be measured at different endpoints for different individuals, see paragraph from page 25 below:

We first calculate the fraction of women in every field of study, including both college majors and vocational training, in the entire Norwegian population using 4 digit Norwegian Standard Educational Codes. We then define the fraction of women in an individual's chosen education based on their first year of enrollment in higher education after 2014 (i.e., after mandatory military service is over). Seventy-eight percent of our sample pursues some type of further education.We likewise calculate the fraction of women in every occupation in the entire Norwegian population using 4 digit International Standard Classification of Occupations Codes. We then define the fraction of women in an individual's chosen occupation in the first year they are employed after 2014. All but three of the individuals in our estimation sample held a job, even if just part time. Finally, we calculate the fraction of women at a workplace based on the establishment an individual works at, using the job with the highest earnings in the first year they are employed after 2014.

We will use 10 months (page page 8), which is after the end of mandatory service, and the soonest the participant could be studying or working, but this is a minimum estimate.

Reported in Abstract. "...there is no long-term effect on choosing fields of study, occupations or workplaces with a higher fraction of women in them after military service ends" covers outcomes (1) and (2). "Contrary to the predictions of many policymakers, we do not find that integrating women into squads hurt male recruits' performance or satisfaction with service, either during boot camp or their subsequent military assignment" covers outcome (3).

Specification. The pre-registration did not specify how the data would be analyzed, so the specifications below are as outlined in the paper (for the most part they didn't have different specifications for the same outcome). In all specifications standard errors are clustered by boot camp squad. They don't specifically mention the number of clusters, but I'll assume it is the 153 squads in their main sample (page 11), but this may be inaccurate, especially since some regressions have a smaller sample size.

For *Fraction of women in one's military occupation*, the specification includes troop fixed effects and I choose the one without control variables (column 7, Table IV).

Across both survey waves, *Answer to "I feel qualified for further military service."* is operationalized as a dummy = 1 if answer is *Strongly agree"*, and *Answer to "Overall, how satisfied were you with military service?"* is a dummy equal to 1 if the answer is *"Good"*. *Promoted to Vice Corporal* equals 1 if promoted. For these outcomes, there's no results without controls. Results are in Table V, and include troop fixed effects.

For *Fraction of women in chosen education/occupation*, results are shown in columns 5-7 of Table VII, with controls only.

Results for answers to questions in third wave survey plus overall evaluation are in columns 1-2-7 of Table VIII, with controls only.

SDs. Obtained standard deviations from authors.

A.3.1.12 Clochard (2022) Improving the Perception of the Police by the Youth

Pre-registration link.

Outcomes. Pre-registration says "See pre-analysis plan" under primary outcomes, but PAP is not linked in the pre-registration. I was able to find it with a link in the paper, which also clarifies that pre-registration was before data collection.

Primary outcome variables in PAP (page 5):

1. Trust partner: Amount sent in trust game to the person paired with.
2. Trust police: Amount sent in trust game to "a random policeman".
3. IAT: Implicit Association Test aiming at observing perceptions of the police, coded as higher value being less bias against the police.

Will not use (1) since it is mechanically affected by treatment. The games were played with tokens that translated to a grade for the student (i.e. real stakes).

Treatment. There's two treatment arms in the paper, *Photo* and *Contact*. We are interested in the *Contact* treatment. For the treatment, participants met their pair face to face, either a police officer (treatment group) or a student (control group), and both alternately answer progressively more personal questions (page 9).

Allport Conditions. No discussion in the paper. We can infer that the intervention had support from authority, equal status within the context of the experiment, but there's no collaboration or common goal (i.e. answering questions doesn't have an objective).

Duration of Contact. The original protocol is adapted so that discussions last 10 minutes (page 9).

Days Since Contact Ending and Measurement. Outcomes are measured right after the end of the intervention (page 11).

Reported in Abstract. "However, the effect fails to translate to an increase in trust in the police in general." covers outcome (2) but there's no mention to IAT, biases, or prejudice to cover outcome (3).

Specification. The empirical strategy in the PAP includes age, education, and gender of participants, plus socio-professional characteristics of the parents, and for the possibility of having been a victim of crimes and misdemeanors (pages 6-7). In the paper there's an additional control for an instructional manipulation check (page 10). Characteristics of parents is omitted and explained in page 40. Standard errors are clustered at the class level.

Results are in Table 1 (page 13), including standard deviation of the control. Sample size relevant for treatment arms differs from sample size and is obtained from figure 1 in page 8. However, some of the missing observations in the regressions (i.e. 359 instead of the full sample of 366 in Table 1) may be from the relevant treatment arms, so 129 is our best estimate.

Since the labelling of the table was not clear, we obtained point estimates and standard errors from the authors. We asked the author to regress the outcome on an indicator for being paired with a police officer, restricted to students assigned to the contact treatment arm.

It appears randomization is at individual level, so we won't use the number of clusters (although table says clustered at class level).

SDs. Control SD reported at the bottom of Table 1. However, this is not the control in our relevant comparison.

Results confusion. The results in Table 1 don't seem to line up with results in Figure 2 (to do the comparison, divide coefficients in *Contact x Police* by the SD, and compare to the difference between the red and green bars). Doesn't seem to align either with raw means figures in Appendix E. Author clarified over email that it may have to do with the addition of controls given the small sample size.

<u>Additional Bundled Effects:</u> Outgroup contact bundled versus pure control and Ingroup contact bundled versus pure control

Treatment. For the comparison between outgroup contact versus pure control, we can compare those who had a conversation with the police to those in the pure control group. For the comparison between ingroup contact versus pure control, we can compare those who had a conversation with another student to those in the pure control group.

Reported in Abstract. "Results indicate a positive effect of contact on trust at the individual level, i.e. toward the specific police officer met. (...) However, the effect fails to translate to an increase in trust in the police in general." No mention of IAT.

The ingroup versus pure control group comparison is not in the abstract.

Specification. Same as above. For the comparison between outgroup contact versus pure control, we initially obtain results from Tables 1 (page 12), adding up the coefficients on *Contact*, *Police*, and *Contact x Police*. However, we then obtain slightly different results from the author and we use those. For the comparison between ingroup contact versus pure control, we obtain results from Tables 1 (page 12), row *Contact*. *Sample size relevant treatment arms* is obtained from Figure 1; 92 participants in the control group plus 42/87 that had contact with the police/another student. Since almost all the sample (N = 366) is in Table 1 (N = 359), I don't rescale.

SDs. Control SD reported at the bottom of Table 1.

A.3.1.13 Loiacono and Silva-Vargas (2023) Can work contact improve social cohesion between refugees and locals? Evidence from an experiment in Uganda

Pre-registration link.

Outcomes. I'm using the first registration from February 2021, since there were no changes to primary outcomes, which are: *Compound definition of social cohesion (between refugees and local workers) that comprises three major indicators: attitudes towards the out-group, implicit and explicit biases and behaviours in real and hypothetical scenarios.* Intervention happened in October 2021 (page 12) so outcomes were pre-registered before analysis.

*Authors shared the PAP with us, but it is from January 2022, where the endline had already been collected for local workers/firms. Doesn't change much except that there's outcomes in Social Preferences that were not in the original pre-registration or paper. We are not including these outcomes since they were pre-registered after the endline for local workers.

Outcomes in the paper that are pre-registered (page 13):

1. Implicit bias: Index averaging Work IAT and General IAT.
2. Explicit Bias Index: constructed from measure of explicit stereotypes and measure of attitudes.
3. Willingness to have a business partner from the outgroup (behavior in hypothetical scenario).
4. (Among refugees) Willing to work in a similar internship matching program (behavior in real scenario).
5. (Among refugees) Willing to work with a Ugandan firm (behavior in real scenario).

   Plus, the outcome below is not in the paper, but could be considered pre-registered since it is a real behavior:

6. Donation to charities: Participants are asked how much (of their earnings in a real lottery) they want to donate to each charity, one that helps Ugandans and one that helps refugees.

For (3), the question is *"Imagine you start a new business, and you can choose between different business partners that have a lot of experience in the sector. How many partners between 0 and 6 would you choose? Of these, how many would be refugees? Of these, how many would be Ugandans?"* (page 50). Since it's hypothetical I will choose generalized outcome type.

For outcomes (4) and (5), the question is *"We would like to know your interest in future projects that might give you the possibility to be matched with Ugandan or refugee firms in Kampala. If you are interested, you can register by sending an SMS to the phone number we will give you. In the message, you need to include (1) your full name, (2) the ID number we will give you and (3) your preference between being matched to a Ugandan firm with Ugandan*

employees or refugee firm with refugee employees (include only one preference). Please only register yourself, not other people! All firms are the same in terms of wages and hours worked" (page 51). For (4), since it measures wanting to participate in the program in general, with outgroups or not, will assign other type. For (5), doesn't seem to include people met, so will assign generalized type.

Treatment. There were 3 possible treatments, plus a pure control group. Respondents either received direct contact [i.e. a one-week internship to skilled refugees in Ugandan firms that were willing to participate], indirect contact [i.e. documentary about characters from both groups], or both [there was cross-randomization, so some participants received both]. For the paper's analysis, they consider one treatment group (respondents that received direct contact, indirect contact, or both), and the control group is composed of refugee workers that are not matched to any firm, local workers that are not matched to work together with refugees, and workers that watch the placebo video (page 9). We are interested in the effect of direct contact only, called *Only exposure* in the paper. The control group is those that received no treatment at all.

Allport Conditions. All conditions satisfied; from page 7: "The direct contact respects Allport's four conditions. First, to respect the equal status condition, we focus on firm workers from two groups - refugees and locals - that work on similar tasks within a firm. This eliminates any potential hierarchy difference between the employees. For institutional support, we focus only on firms that are willing to participate in the program, thus endorsing the contact between employees. The third and fourth conditions are respected because workers work for the same firm and in the same department, and thus, cooperate towards common goals."

Duration of Contact. Internships lasted 1 week* (page 8).

*Page iii of Loiacono's thesis suggests that the intervention was the same as the one in the first paper of the thesis, and from that paper (page 17) we can obtain that the median duration of the internships was 7 days, each intern worked an average of 7 hours a day, and managers at the firm spent more than 5 hours supervising the intern every day. So our best guess is 7 days*6 hours (midpoint 5-7) = 42 hours.

Days Since Contact Ending and Measurement. The 1-week internship happened in October 2021. The endline of firms and local workers happened between November and December 2021, and the endline for refugees happened between July and August 2022 (page 37). Thus, will estimate days to be 30 for locals and 210 for refugees.

Reported in Abstract. The abstract refers to the effects on their main treatment of interest (i.e. the pooled treatments), and there's no mention of the effects of exposure only.

Specification. For the main analyses, they pool together all treatments, but in the appendix they present the results for separate treatments (they registered that they pool all treatments after the intervention had happened). Results are in Table A3, A4, and A5 (pages 44-46), row *Only Exposure*.

The results for the separate treatments are only shown with one specification, with refugee strata and robust standard errors. For outcomes (1)-(3) it is not clear if they have baseline controls or not; they are not listed in the table note, but according to the text, they control for the baseline value of the outcome when possible (page 18), and these three outcomes are measured at baseline (i.e. appear in baseline balance checks; Table 1 and Table 2).

For *Sample Size and Number of Clusters for relevant treatment arms*, ideally we could add the participants in the only exposure treatment plus those in the control, but they don't report

the number of people in each treatment (exposure vs video vs both).

SDs. Explicit bias index is normalized (see mean of 0). Remaining SDs provided by the authors.

A.3.1.14 Greene et al. (2024) Interacting as Equals: How Contact Can Promote Tolerance Among Opposing Partisans

Pre-registration link.

There's a PAP but it's not public, and since outcomes are described in the pre-registration we will not request.

Outcomes. Below is a list of the primary outcomes in the pre-registration, along with how they were called in the paper:

1. Tolerance → In the paper it's *Tolerant behavior index*, which adds and standardizes *Sharing* and *Willingness to dialogue*, measured both at endline and follow-up (page 10).

2. Preference for democracy → *Democracy preferred* and *Majority Vote*; two survey items in the endline survey only (page 13 Supplementary Materials).

3. Pro-social, pro-democracy behavior as in declared willingness to be a poll worker → *Poll worker*; survey item in the endline survey only (page 14 Supplementary Materials).

4. Willingness to participate in future meetings, and willingness to participate in mix-partisanship meetings → *Willingness to dialogue*: willingness to take part in a future online meeting with other participants, which we indicated would include opposing partisans.

5. Willingness to donate to the aforementioned NGO → *Donations to anti-corruption NGO*, measured only at endline (page 14 Supplementary Materials). I consider this outcome as "Other" type, but this NGO has been heavily criticized by the leader of the Morena party (and this is told to participants), so could be considered *Generalized to Outgroup* but only for those that sympathize with Morena. The question is phrased hypothetically, so it's not behavioral or incentivized.

6. Willingness to donate in dictator games to the fellow party sympathizers, vis a vis people who sympathize with a different party → *Sharing*: choose to donate to an anonymous participant with opposing political sympathies.

7. Generalized trust and trust in fellow country people → *Trust people* and *Trust a fellow Mexican*, measured only at endline (page 14 Supplementary Materials).

Will not include (4) and (6) given that they are in the index in (1).

Treatment. The paper describes two treatments: contact with equal status and contact with unequal status. We will use the treatment with equal status (since one of the Allport conditions for contact to work is equal status), and this was pre-registered to be more effective: "Collaborative contact under equality in the interaction increases the variables in H1 more than under inequality in the interaction (E > U)" (from pre-registration). Treatment is bundled with the pair

interaction. For outcomes measured immediately after the intervention and in the follow-up, we will include both as separate outcomes.

Allport Conditions. "During the contact interaction, we held constant across contact treatment arms the presence of common goals and the incentive to collaborate" (page 7). Plus, "we manipulated relative status" (page 7) but we focus on the treatment with equal status. We can infer that the contact is supported by authorities.

Duration of Contact. Collaboration between pair members lasted ten minutes (page 7).

Days Since Contact Ending and Measurement. Endline survey directly after the intervention (0 days), plus follow-up survey approximately 3 weeks later (21 days) (page 6).

Reported in Abstract. "Interacting under both equal and unequal status enhanced tolerant behavior immediately after contact; however, three weeks later, only the salutary effects of equal contact endured". Only result in abstract; covers tolerant behavior index at follow up and at endline.

Specification. There's no details in the pre-registration about the specification or analysis. From page 18: Results are ITT effects, and the specification has an indicator for equal status and one for unequal status, such that the omitted group is the control. The specification includes block fixed effects and pre-treatment covariates (including the outcome measured at baseline). Standard errors are robust and clustered at the pair level.

Tolerance index results are in the left of Figure 2 (page 12), and column 5 of Table T13 for the follow-up (page 66 Supplementary Materials). For the endline they are in Table T17 (page 70 Supplementary Materials; table note says sample size is complete pairs at endline, but based on Table T1 it has to refer to individuals).

Results for *Democracy preferred*, *Majority Vote*, and *Poll Worker* are in Table T18 (page 71 Supplementary Materials). Results for *Donations to anti-corruption NGO*, *Trust people* and *Trust a fellow* are in Table T19 (page 72).

For all of the above I inferred the standard error from the 95% confidence interval, which lines up with the standard errors reported in a previous version of the paper. Number of clusters is sample size divided by two, because the main analysis sample only includes individuals in pairs where both completed the study (page 5). Sample sizes for different treatments are in Table T-1 (page 50 Supplementary Materials).

SDs. Tables include Control SD at the bottom.

A.3.1.15 Clochard et al. (2023) Low-Cost Contact Interventions and Inter-Ethnic Trust: Evidence from Senegal

Pre-registration link.

Outcomes. There are two outcomes in the pre-registration and paper: *Trust* and *Prejudice*. *Trust* is how much you send to your partner in the dictator game, and is thus ruled out for being mechanically affected by treatment. *Prejudice* is the share of a fictitious endowment sent to the member of the participant's own ethnic group above a certain threshold (more details in page 10). The variable is defined as missing for the 17% of participants not belonging to the two main ethnic groups (Wolof and Pulaar), and is coded differently than in the PAP and in previous versions of the paper.

Treatment. Participants were randomized into the *Contact* treatment, *Photo* treatment, or *Control*. In the *Contact* treatment, participants are randomly assigned to have a discussion with a research assistant from their same ethnic group or a different one (Wolof or Pular, page 7).

Allport Conditions. All conditions satisfied; from page 9: "Importantly, the procedure we used meets the four conditions for effective contact in Allport (1954). The pairs have equal status because they both answer the same set of questions. The contact is positive because they meet university student assistants who are instructed to be friendly. The contact is supported by authorities because the experiment is conducted in public buildings, with the approval of local leaders. Lastly, the pairs share a common goal to have the discussion proceed smoothly."

Duration of Contact. 10 minutes (page 7).

Days Since Contact Ending and Measurement. Outcomes were collected right after the intervention (page 9) and then again at a one-month follow-up phone survey (page 30). The results presented in the paper are for the immediate results, but the phone survey result for *Prejudice* can be found in the appendix (page 31). They only surveyed participants in the *Contact* or *Photo* treatments, but we can still draw the comparison for *Contact*.

Reported in Abstract. "Contact is found to be effective in increasing interethnic trust toward the individuals met during the intervention, in line with previous results from longer interventions. However, the results do not generalize to the collective level." General enough that it can cover the outcome in the short run and long run.

Specification. Page 7 of the PAP describes the specification, which is the same as the one in the paper. They control for age, education, gender and ethnicity (page 11).

Results in the short-run are obtained from column 2 in Table 1 (page 14). Point estimate is obtained by subtracting the point estimate in *Contact x Same ethnicity* from the point estimate in *Contact x Different ethnicity*. For the long-run they are in column 2 of Table F1 (page 31; there seems to be a typo in the title), by subtracting the point estimate in *Contact x Same ethnicity* from the point estimate in *Contact* (this seems to refer to *Contact x Different ethnicity*). For both outcomes, the sign of the point estimate is reversed. There's no p-value for *Contact x Same ethnicity = Contact x Different ethnicity*, so we can't obtain the standard error.

The sign for *Prejudice* is reversed because "The variable is coded as positive (the subject showing more prejudice) if the amount sent to the fictional person from their own ethnic group is higher than this threshold."

We obtained from the authors the point estimate (which lined up with what we had, minus some rounding difference) and standard error.

Sample size relevant for treatment arms differs from sample size and is obtained from Figure 1 in page 4 (= 341). However, not everyone in the treatment is in the results for *Prejudice* because the prejudice variable is missing for subjects who do not belong to one of the two main ethnic groups, so I rescale: I multiply N by the number of participants in the relevant treatment arms (138 + 203), and divide that by the total N for all conditions for short term (895), and the N for Contact and Photo treatments for long run (because only those two groups received the follow-up).

SDs. Standard deviation for the whole sample in the short run is obtained from the bottom of Table 1. For the long-run, it is obtained from Table B1, row *Long-term in-group bias* (page 24; not clear why the N doesn't line up with the one in Table F1).

Additional Bundled Effects: Outgroup contact bundled versus pure control and Ingroup contact bundled versus pure control

Treatment. For the comparison between outgroup contact versus pure control, we can compare those who had a conversation with an outgroup member to those in the pure control group. For the comparison between ingroup contact versus pure control, we can compare those who had a conversation with an ingroup member to those in the pure control group.

We will get the results for the short-run outcome only, since the follow-up survey was not done for the control group.

Reported in Abstract. "Contact is found to be effective in increasing interethnic trust toward the individuals met during the intervention, in line with previous results from longer interventions. However, the results do not generalize to the collective level." General enough that it can cover the outcome for the outgroup comparison.

Specification. Same as above. Results in the short-run are obtained from column 2 in Table 1 (page 14), row *Contact x Different Ethnicity* and row *Contact x Same Ethnicity*. For sample size relevant treatment arms, I obtained it from Figure 1 (N = 254 in the control group plus N = 203/138 contact with different/same ethnicity group) and then rescale it as explained above.

SDs. Control SD reported at the bottom of Table 1.

A.3.1.16 Abril et al. (2023) Building Trust in State Actors: A Multi-Site Experiment with the Colombian National Police

Pre-registration link.

Outcomes. The pre-registration lists the following primary outcomes:

1. Public trust in the police, measured through one single public trust question in the citizens survey. "The National Police of Colombia is an institution in which I can trust."
2. Demand for policing services, measured through a costly request (we asked residents whether they would support a new tax directed at funding the police).
3. Police beliefs, measured as their trust in citizens and their second-order beliefs on public trust. This turns into two outcomes in the paper:
   (a) Officers' trust in citizens
   (b) Officers' beliefs about citizens' public trust

Outcomes (1) and (2) are measured for citizens, and (3a) and (3b) are measured for police officers.

*Outcome 3 was changed after the intervention ended but note at the bottom clarifies that: "We updated the description of the police measures after the intervention finished, but not as an ex-post change."

Treatment. Treatment is assigned at the police quadrant level. First treatment arm received the core components of the COP Initiative. Second treatment arms received the same as the first, with the addition of an information campaign. The third treatment arm is a pure control.

Our relevant comparison compared the first two treatment arms with the pure control arm, and treatment is bundled.

The COP Initiative consisted of retraining officers in adopting procedural justice principles in interactions with citizens, plus increasing interaction with citizens on a randomly-selected street block.

Allport Conditions. No discussion in the paper. We can infer that contact is supported by authority and is collaborative with a common goal (i.e. stay safe), but participants have unequal status given (i.e. police have the authority to enforce the law, etc).

Duration of contact. The intervention lasted 6 weeks (page 12), but contact may have lasted 0 with a specific citizen. For our best guess, we will use what is most likely an overestimate (15 minutes), but that should keep this study under light touch interventions.

Days Since Contact Ending and Measurement. Intervention was implemented until late April 2022 and endline for citizens happened in late April and May (page 14), assuming a typo). I use half a month (15 days) as a mid-point. For the police, data was collected during implementation (page 15), so 0 days.

Reported in Abstract. "The intervention improved policing frequency, perceptions of fair treatment, and public trust. (...) We find no impacts on officers' trust in citizens or beliefs about public trust, implying that institutional change may require more profound efforts." Outcomes (1), (3a), and (3b) reported in the abstract.

Specification. They use OLS to estimate ITT effects, including quadrant-level and individual-level covariates (page 18). They have city, poverty tercile, and baseline trust triplet fixed effects (i.e. strata fixed effects). Errors are clustered at the quadrant level. They have a specification where they pool both treatment arms.

Results are in Table 3 column 2 (page 20). Number of clusters is obtained from the number of quadrants in Table 1 (page 16).

SDs. Standard deviations obtained from the author.

A.3.1.17 Burlacu et al. (2024) Exploring the Impact of a Multifaceted Intervention on Knowledge, Attitudes and Behaviors towards Persons with Visual Impairment

Pre-registration link.

Outcomes. Pre-registration is from March 23, 2022. According to the timeline of the study, the intervention had already begun in February and the follow-up was implemented between March and April (page 7), so we can assume the outcomes were pre-registered before analysis.

The primary outcomes in the pre-registration are, followed by how they appear in the paper:

1. Money passed in the dictator game to another visually impaired student: Students played the role of the dictator for 3 rounds and were asked if they would like to share part of it with another anonymous student (randomly selected from another school in the province, so not someone they met) that was a blind student (primary outcome), a generic student, or a student with motor impairment (secondary outcomes) (page 12). Primary outcome is *Giving in the DG (%) Visual Impairment.* The choice could be implemented, so there were real stakes.

2. Willingness to pay to interact socially with a person with visual disabilities: *WTP* in the paper, measure WTP to participate in a short individual meeting at school with a person with visual impairment (page 11). The choice could be implemented, so there were real stakes.

3. Beliefs on performance of individuals with visual disabilities in various tasks: memory, math (summation and multiplication), 400 meter sprinting: Under *Incentivized beliefs* in the paper (page 12), *Sprint, Memory, Summation,* and *Multiplication.* These outcomes were incentivized for accuracy.

4. Beliefs on life satisfaction of individuals with visual disabilities: Under *Incentivized beliefs* in the paper (page 12), *Life Satisfaction.*

All outcomes are *Generalized to outgroup.*

Treatment. The RCT design was between and within class. Two classes were selected to serve as a pure control - all students received the intervention only after the follow-up survey. Students in the other four classes (called main classes in the paper) were randomly assigned at the individual level to receive the intervention either in between the two surveys (treatment group) or after (control group). We focus on the main classes comparison as this was pre-registered as the main comparison of the paper. (In principle we could also look at the between-class effects, but with only two control group classes, such effects would be very imprecisely estimated. In addition, these effects are not reported in the paper – the control group classes are used only to test for spillover effects, through a comparison with control students in treatment classes).

The intervention was composed of the "informational treatment" and the "simulation treatment." The informational treatment aimed to increase knowledge and understanding of visual impairments, and was delivered in class by a sighted facilitator. The simulation intervention was in a "restaurant' in complete darkness where students were served by blind waiters and had the opportunity to interact with the blind waiters (page 6). The treatment is bundled.

Allport Conditions. All conditions satisfied; from page 6: "At the same time, we also expected that the implemented simulation treatment would enable participants to experience positive inter-group contact by concretely satisfying the key conditions identified in Allport (1954): equal status, common goals, no inter-group competition, and authority sanction."

Duration of Contact. Each activity lasted 50 minutes (page 5); only the simulation treatment involves contact, so 50 minutes.

Days Since Contact Ending and Measurement. Follow-up survey implemented 2-3 weeks later, so will use 18 days.

Reported in Abstract. "Moreover, the intervention does not improve outcomes measured through incentivized choices, such as the willingness to pay for social interaction with persons with visual impairment, beliefs regarding their performance and outcomes in various domains, and altruism towards them." The underlined phrases cover outcomes (2), (3)-(4), and (1), respectively.

Specification. All outcomes are normalized to vary between the theoretical minimum and maximum (page 8). They estimate ITT with an indicator for treatment, controls (including the

outcome measured at baseline when available), and school fixed effects (page 15), with robust standard errors. No details of the specification in the pre-registration.

Results for outcomes (1) and (2) are in columns 2 and 1 of Table 4, respectively (page 20). For outcomes (3) and (4), they investigate potential treatment effects non-parametrically because the empirical distribution of beliefs deviates from the Gaussian distribution substantially for some of the choices (page 21). "For all outcomes, the median values by group are almost identical. Smirnov-Kolmonrov tests fail to reject the null hypothesis, with associated p-values larger than 0.5 for all outcomes." We received point estimates and standard errors from the author. For sample size we use Table A3 (page 36).

SDs. For outcomes (1) and (2), Table 4 reports control SD. No SD for outcomes (3) and (4). Received SDs from author.

A.3.1.18 Barros (2024) The Power of Dialogue: Forced Displacement and Social Integration amid an Islamist Insurgency in Mozambique

Pre-registration link.

Outcomes. Primary outcomes were pre-registered in August 2022, and intervention occurred between August and October 2022 (page 16), so outcomes are considered pre-registered before analysis. The primary outcomes in the pre-registration are (further explanations in the pre-registration):

A. Tolerance towards IDPs and locals: survey questions and lab-in-the-field games (joy of destruction).

B. Trust towards locals / IDPs: survey questions and lab-in-the-field games (trust, public goods).

C. Social Cohesion: survey questions and lab-in-the-field games (public goods).

D. Integration of IDPs in local community: survey questions and follow-up surveys tracking individuals' connections.

E. Preference / Bias towards insurgents and religious extremism: survey questions (religious extremism), list experiment (preference for insurgents), and implicit association test (bias towards insurgents).

I went through all the outcomes in the paper (outcomes described in page 61), and listed them below if they fell under one of the categories pre-registered above. When there's an index available, I prioritize that and indicate whether it was pre-specified. All outcomes are measured at post-meeting and follow-up, with results reported separately.

1. *Tolerates IDPs staying in host neighborhood* (under A). Index not pre-specificed, constructed by averaging the responses of two binary survey questions that measure whether locals think that IDPs should be moved away from host neighborhoods or sent back to their homelands (page 61).

2. *Positive beliefs about IDPs in host neighborhood* (under B). Index not pre-specificed, constructed by averaging the responses to three survey questions that measure the extent to which locals appreciate IDPs (page 62).

Outcomes above are measured for locals only. Results are in Table 1 (page 29). Coded as *Generalized to outgroup* since the questions are not about specific individuals met.

3. *Anti-social game (Destroys endowment)* (under A).

4. *Monetary contribution in public good game* (under B and C).

5. *Trust game - Donation* (under B).

6. *Trust game - Retribution* (under B).

   All games are measured for both locals and IDPs, with real money, and with results for each reported separately. During the post-meeting activities, individuals played games with other players from the same cohort (page 65), so outcome type is *Includes specific people met*. During follow-up activities, individuals were totally randomized and assigned to groups of other players of the same neighborhood, so outcome type is *Generalized to outgroup*. Results for locals and IDPs are in Table D1 and D2, respectively (pages 72 and 73).

7. *Trust in IDPs* (under B).

8. *Trust in Locals* (under B).

9. *Feels better integrated* (under C).

10. *Participation in neighborhood life* (under D).

    (7) is measured for locals and (8) for IDPs through a survey question. Both (9) and (10) are measured for IDPs only. (7) and (8) above are a single question, not indices (page 62); coded as *Generalized to outgroup* since the questions are not about specific individuals met. Social cohesion outcomes don't appear in survey questions as described in the pre-registration. While (9) is not explicitly described under C, it seems to be a good fit for a measure of social cohesion.

11. *Discrimination against IDPs* (under D).

12. *Feels discriminated by locals* (under D).

    (11) is measured for locals with a list experiment where the sensitive sentence was: "I do not like that IDPs are living in my neighborhood" (page 63), so outcome type is *Political and Cultural Attitudes*. (12) is measured for IDPs with a list experiment where the sensitive sentence was: "I feel discriminated by the local population of this neighborhood", so outcome type is *Explicit Evaluations*.

    Results for (8)-(10) and (12) are in Table 3 (page 35). Results for (7) and (11) are in Table 1 (page 29).

13. *Religious tolerance* (under E). Index created from averaging the responses to two survey questions (page 63). Pre-registration didn't explicitly mention the index, but did say they would use survey questions to measure this.

14. *Preference for insurgents - IAT* (under E).

15. *Preference for insurgents - List Experiment* (under E).

    Outcomes (13)-(15) are measured for both locals and IDPs, separately, but the analysis is restricted to the Muslim population. Results are in Table 6 (page 42). For these I used Other outcome type (the religious extremists are not the ingroup/outgroup, both locals and IDPs are predominantly muslim).

    For trust outcomes (B), pre-registration stated that they would construct a trust index by aggregating participants' trust level in different individuals. This is not found in the paper, but there's additional tables in the appendix (Tables D3 and D4) with levels of trust for several different groups. Since trust towards the outgroup is covered above (outcomes (7) and (8); the results presented are the same as the ones in Tables D3 and D4) and there's no index, we are omitting the rest of the trust measures towards different groups.

    For (D), pre-registration stated that there would be an index capturing the strength of IDPs connections (i.e. "importance of the people IDPs know, how regular they contact these people; whether IDPs contact more regularly other IDPs or if they also have good connections with locals"). Couldn't find this index in the paper, although Table 5 (page 40) has outcomes for intra-cohort networks generated by community meetings. We decided to include one of these outcomes even if it differs from what was pre-registered. The outcomes in this table capture the percentage of cohort members that kept in touch with the respondent. We report results for *Anyone in cohort* only since it sums *Persons not known* and *Persons already known*. Given that this outcome was pre-registered for IDPs, I only include the results for them (columns 1 in Panel B; but the table includes results for locals as well). Outcome Type is *outcome unrelated to prejudice or intergroup relations*.

16. *Anyone in cohort*

Treatment. Treatment is attending a community meeting between hosts and IDPs. Meetings have a randomly assigned community leader as a moderator (page 17). The control group doesn't get a meeting at all, so treatment is bundled with the meeting itself.

Allport Conditions. All conditions satisfied; from page 17: "These conditions, as applied to the design of the community meetings, were equal status of both groups, meaning there was no hierarchical relationship during intergroup contact; cooperation, meaning both groups engaged with each other in a noncompetitive environment; common goals, such that both locals and IDPs engaged in the meeting with the same objectives; and support from authorities, meaning that the meetings were regulated by an entity respected by both groups."

Duration of Contact. Meetings lasted approximately 3 hours. I obtained the duration of meetings from page 1 of protocol of community meetings (appendix A).

Days Since Contact Ending and Measurement. The community meetings took place between August and October 2022. Participants are surveyed 2 to 3 days after the meeting (page 21), and then 2 to 3 months after community meetings (page 22). Pre-registration said they predicted follow-ups at 3, 6, and 9 months, but the last two don't seem to have materialized.

Reported in Abstract. "Analysis of survey data, list experiments, the Implicit Association Test, and lab-in-the-field games shows that the community meetings produced immediate and

sustained positive effects on the relationship between hosts and IDPs." Very broad, but it covers all outcomes since they are all obtained through the data mentioned. "Religious tolerance also improved, and religious-extremist beliefs decreased, highlighting the potential of intergroup contact to support counterinsurgency efforts." Further covers outcomes (13)-(15).

Specification. There's no specification in the pre-registration. The specification (page 27) includes neighborhood dummies, individual demographic characteristics measured at baseline, controls for meeting characteristics, and the outcome measured at baseline when available. Errors are clustered at the cohort level (participants in the same meeting). There's no results without controls. Couldn't find the number of cohorts for different outcomes. Figure 10 (page 20) says total number of cohorts for treat and control, but sample sizes don't always line up with that. We will assume 108 cohorts (54 treatment and 54 control) for all outcomes (even when the sample is smaller it seems unlikely that they would have lost a whole cohort).

For outcomes (11)-(12) and (15) I reverse the sign of the point estimate, which is obtained from row *Sensitive x Treated*. I also reverse the sign for (3) and (14).

SDs. Received SDs for the control group from the author.

A.3.1.19 Chaudhry and Hussain (2024) The economic effects of inter-sectarian contact

Pre-registration link.

There is a little ambiguity about whether their sample includes both Sunnis and Shias, seemingly due to some typos (e.g. abstract says Shia mosques were visited by Sunnis, while the rest of the paper says that Sunni mosques were visited by Shias; and on page 17, on top of the coupons it says "Sunni/Shia respondent"). Given that these appear to be typos, we conclude that the survey participants are all Sunnis.

Outcomes.The pre-registered outcomes are: (i) do you trust the opposite sect and (ii) "our experimental game's outcome" [incentivized voucher to buy books]. I consider (1)-(2) and (4) below to fall under (i). Pre-registration is from August 2022 and the endline was collected in September 2022, so outcomes are considered pre-registered before analysis.

1. Business, Shias: Answer to the question "What do you think about entering into business with Shias?" (page 16). Answer: 0 (very bad) to 5 (very good).

2. Hiring Change: Answer to question "What do you think about recruiting Shia/Sunni Workers?" (page 18). Answer: 0 (very bad) to 5 (very good). Coded as change between baseline and endline.

3. Book Choices: Change in demand for Sunni books from baseline to endline (positive point estimate means they demand more books of their own sect).

4. Plumber choice: The dependent variable is a binary variable which is 1 when respondents, Sunnis, choose discounted plumbing services from a member of the opposite, Shia, sect (the names of the plumbers allow for clear sectarian identification) and 0 otherwise.

Treatment. The sample corresponds to worshippers in mosques that belong to the Sunni sect (page 11). They have three different mosque-level treatments, plus a control arm (i.e. worshippers in mosques with no intervention): having Shia worshippers pray in the Sunni mosque

(*Contact*), having the imam deliver a message in support of unity (*Leadership*), and both (*Combined*) (page 13). We will do two comparisons, since they both compare the same degree of intergroup contact: First, the comparison between *Contact* and the control arm. Second, the comparison between *Combined* and *Leadership* (i.e. effect of contact conditional on imam delivering message of unity). I consider both of these *High vs no outgroup contact*, given that there's so many ingroup worshippers in the mosque that the control group does have equivalent intergroup contact, but with an ingroup (i.e. it is not a pure control group).

Allport Conditions. No clear discussion on the paper. We can infer they have equal status (all prayers in the mosque), support from authority (even more so in the Leader treatment, but still in the other conditions since worshippers are allowed in the mosque), without collaboration or common goal.

Duration of contact. They send volunteers every day over a twelve-day period during the second-to-last prayer of the day to every mosque (page 13). Depending on how often a participant goes to the mosque, they will have different exposure. Muslim prayers last at most 10 minutes, so assume 120 minutes of contact, but this is a high estimate.

Days Since Contact Ending and Measurement. Endline data collection is one month after the intervention (page 15).

Reported in Abstract. "We find that the combined treatment (but not the stand-alone treatments) reduces prejudice: more Sunni worshipers choose to hire a Shia plumber and purchase books about Shias." Covers (3) and (4) outcomes for both comparisons.

Specification. Regression includes binary variable indicators for each treatment, strata fixed effects, and standard errors clustered at the strata level (of which there are 7, clarified by authors). They report results with and without controls (page 19), without stating a preference, so I report results without controls.

Results for outcomes (1) and (4) are in columns 3* and 1, respectively, of Table 2 (page 21), and results for outcome (3) are in column 1 of Table 3 (reverse point estimate for outcome (3); page 23). For the first comparison, we use the point estimate and standard error in *Contact* row, and for the second comparison we subtract the point estimate in *Leadership* row from the point estimate in *Combined* row (SEs requested from authors). Number of clusters is in *Number of Mosques.*

Results for outcome (2) are in column 2 of Table 12 (page 37), with no mention in the main paper. We omit this table, and thus the results for this outcome, because the author explained that this table includes observations where there was a mistake in treatment assignment.

For the number of clusters we use the number of mosques (since that is the unit of randomization).

Can't find sample size relevant treatment arms, but we know the number of mosques in each treatment: 8 in the control group, 6 in the contact only, 5 in leadership only, and 5 in combined (page 5). Assume that the sample size per treatment arm is proportional to the number of mosques per treatment arm (i.e. divide sample size over number of mosques to get number of observations per mosque, and then multiply by the number of mosques that received the treatments in the comparison of interest).

SDs. No standard deviations in the paper; obtained from the author.

A.3.1.20 Ghosh et al. (2024) Creating Cohesive Communities: A Youth Camp Experiment in India

Pre-registration link.

Outcomes. Outcomes were pre-registered before the launch of the endline survey (which is where most outcomes are collected). For the one-year-later phone survey, 50 participants had already been surveyed when outcomes were pre-registered (considered pre-registered).

For the endline survey, there are four categories of pre-registered outcomes, with several items underneath. These categories are: Social preferences, Willingness to interact, Identity, and Political and Social Attitudes. The outcomes below are the ones for which there are results for the contact treatment and fall under one of these categories. The pre-registration didn't specify the indices.

1. Dictator game (stranger)
2. Number of outgroup friends
3. Willingness to play: We measure willingness to interact with the outgroup using self-reported friendships and an incentivized willingness to "play" measure.
4. National identity index:
   (a) Self-report measure of which group they feel most attached to (political and cultural attitudes).
   (b) Choose a magnet with Indian flag or religious symbol (behavioral measures).
5. Attitudes index:
   (a) Inter-religious attitudes with two yes-no questions: (i) would you be willing to marry a [Hindu/Muslim] when you're older? (behavioral measures) and (ii) would you support giving Indian citizenship to a [Hindu/Muslim] immigrant? (political and cultural attitudes)
   (b) Attitudes towards foreigners using feelings thermometer ratings (from 0 to 100) toward Nepalese/Bangladeshi and Pakistani people for Muslims/Hindus (explicit evaluations).
   (c) Attitudes toward politicians, we take the mean of thermometer ratings for Mahatma Gandhi and reverse-coded ratings for Narendra Modi (political and cultural attitudes).
   (d) Attitudes towards democracy, we asked respondents which type of political system they think is the best form of government (political and cultural attitudes).

For the follow-up phone survey, the pre-registered primary outcomes are:

1. Well-being
2. Number of outgroup friends

Treatment. Those assigned to the camps were randomized into teams of ten; in each camp, six teams with five Hindus and five Muslims (high contact for Hindus, low contact for Muslims), and six teams with eight Hindus and two Muslims (low contact for Hindus, high contact for Muslims) (page 13). Thus, comparison is *High versus low outgroup contact*.

Allport Conditions. From page 3: "Second, ethnically mixed camps bring children into close collaborative contact with ethnic outgroups, which can improve intergroup relations." Then we can infer that contact is supported by authorities, that campers have equal status, and given the nature of the activities, have a common goal (i.e. like sports).

Duration of contact. Camps lasted 12 days, and met for 4 hours each day, giving a total of 48 hours of activities (page 11).

Days Since Contact Ending and Measurement. First endline at 6 weeks: "first endline between four and seven weeks after the camps had concluded, with the median respondent completing the survey 5.9 weeks later" (page 13). Phone survey administered 12 to 13 months after camp's conclusion (page 17; 1 year + 2 weeks = 380 days).

Reported in Abstract. "Meanwhile, additional camp elements have heterogeneous effects: rituals have more positive impacts for the Hindu majority than the Muslim minority, while higher intergroup contact backfires among Hindus but not Muslims." Too unspecific; previous sections of the abstract refer to the effect of camps but not contact specifically.

Specification. Page 19 in the paper (lines up with pre-registration): To analyze the effects of contact the sample is restricted to those assigned to camps. The specification includes an indicator equal to one for individuals randomized into a team with high exposure to outgroup individuals, and equal to zero otherwise. Regression includes camp x religion fixed effects, and the baseline version of the outcome when it is available (otherwise, they do not include baseline controls). Standard errors are at the camp-team-level, with 24 clusters.

Results for the contact intervention are obtained from Table S11 (page 11 Online Appendix).

Effects of contact on well-being and number of outgroup friends at the second endline are not in the paper, but we estimate them using our data following the same specification as above.

Additional Bundled Effects: Outgroup contact bundled versus pure control

Outcomes. In contrast to the contact comparison, the results for the bundled treatment include the results for each of the indices encompassing the pre-registered categories of outcomes, so we will use the following outcomes at endline 1: Social preferences index, Willingness to interact index, National identity index, Attitudes index. Plus the following outcome at endline 2: Number of outgroup friends, Well-being.

The social preferences index includes the dictator game (behavioral measure) and public goods game (behavioral measure). Given that the public goods game is played with their camps team, the index is coded as *Includes specific people met.*

Treatment. We compare those who attended camps with those in the pure control group, since all campers had either high or low intergroup contact.

Reported in Abstract. "We find that camps reduce ingroup bias, increase willingness to interact with outgroup members, and enhance psychological well-being. Campers continue to have more than twice as many outgroup friends than control participants one year after the camps ended." We can infer from "ingroup bias in the dictator game" (page 21) that ingroup bias refers to social preferences.

Specification. Specification includes dummy for being in a camp plus baseline covariates

and randomization strata fixed effects, with robust standard errors (page 18). Results are in Figure 1 (page 20) and Table S6 (page 9 appendix) for endline 1 outcomes, and in Table 4 for endline 2. ***SDs.*** We use SDs from the previous comparison, and we obtain from our own data the SD for *Social preferences index*.

A.3.2 EGAP Registry

A.3.2.1 Scacco and Warren (2018) Can Social Contact Reduce Prejudice and Discrimination? Evidence from a Field Experiment in Nigeria

Pre-registration link.

We use the available data.

Outcomes. Outcomes are described in PAP from January 2020 (PAP date shows as 2020, but talks about 2015 in the future tense, so it seems that PAP is from earlier but released later on). According to EGAP registration, researchers had not accessed outcome data, so outcomes are considered pre-registered before analysis. Outcomes (1)-(3) below measure prejudice, which was pre-registered as *Prejudice* in the PAP (page 3 PAP). The authors conducted an exploratory factor analysis to determine if all items could be combined into a single scale, but found that the components retained three dimensions corresponding to the three indices. These outcomes are *Generalized to outgroup* (inferred from survey questions in page A10 of the appendix). Outcome (4) was described in the PAP under *Discrimination*, and outcome (5) was described in the PAP under *Conflict-related behaviors and attitudes.* Outcomes (4) and (5) are *Includes specific people met* because participants played against another random participant in the study, which could be their classmate (and the enumerator indicated when this was the case; page 661). The final outcomes in the PAP were *Cooperation* and *Trust*, which are not referenced in the paper.

1. Negative Attributes Index: Asks how well negative adjectives describe the outgroup.
2. Positive Attributes Index: Asks how well positive adjectives describe the outgroup.
3. Out-group Evaluation Index: Asks respondents to rate how lazy/ignorant/not generous they think the outgroup is.
4. Number of bills given in dictator game.
5. Number of bills destroyed in destruction game.
6. Cooperation: willingness to contribute to public goods, both in lab-in-the-field behavioral games and in actual giving to the Nigerian Branch of the Red Cross/Red Crescent Society.
7. Trust: extent to which subjects report that they trust others, including outgroup members.

*PAP mentions a follow-up survey, for which the authors were going to submit a separate pre-analysis plan (page 2). This is not in the paper.

Treatment. The paper has three different comparisons, in the context of providing a computer education program: No program vs program, Homogeneous classroom vs Heterogeneous classroom, and Co-religious partner vs Non-co-religious partner (within heterogeneous classrooms) (page 659). Given that both the second and third can be considered *High vs no outgroup contact* comparisons, we'll report results for both. The control group would be the homogeneous classroom in the first comparison, and non-co-religious partner in the second.

Allport Conditions. From page 659: "Within classrooms, UYVT participants were randomly assigned to a partner from their own or the other religious group, with whom they worked in close cooperation on course assignments and custom-designed partner activities." From this, it is clear there was cooperation and a common goal. Plus, "Although the core claim of social contact theory—that positive and equal-status social contact with members of the out-group should decrease prejudice—is widely applied in peacebuilding programs, to our knowledge, the theory has never been directly tested using an empirically rigorous field experiment in an ongoing conflict environment." Finally, we can infer that there's support from authority.

Duration of contact. Sixteen weeks from September to December 2014 (page 657). "Each section met twice weekly for a total of four hours per week over sixteen weeks" (page 660).

Days Since Contact Ending and Measurement. Endline survey in January 2015, so assuming 30 days (page 657) (January and February 2015 according to PAP, page 2).

Reported in Abstract. "After sixteen weeks of positive intergroup social contact, we find no changes in prejudice, but heterogeneous-class subjects discriminate significantly less against out-group members than subjects in homogeneous classes." Outcomes (1)-(3) measure prejudice so these are reported, but the mention to discrimination (outcomes (4) and (5)) is only made for classroom comparison.

Specification. All estimates are OLS with robust standard errors. For outcomes (1)-(3) the specification only includes a dummy for the treatment comparison. For outcomes (4) and (5) the specification includes round-of-play fixed effects, a dummy for treatment comparison, an interaction term between treatment and outgroup player (i.e. individuals played 10 rounds of each game and for each round of play, subjects were randomly assigned to another individual in the study, either an in-group or an out-group member), and errors are clustered at the individual level.

Results are obtained from columns 4 and 7 of Tables 2 to 5, plus Table 7 (pages 666-669 and page 673).

Point estimate for the destruction game is reversed following the description in page 664. For all other outcomes positive means "good" (page 664).

SDs. I take SDs from available data (*data_APSR.dta* and *data_APSR_long.dta* for game outcomes), control SD is SD for those in homogeneous classrooms for first comparison and those in homogeneous pairs for second comparison.

Additional Bundled Effects: Outgroup contact bundled versus pure control and Ingroup contact bundled versus pure control

Treatment. For the comparison between outgroup contact versus pure control, we can compare those not in the program (pure control group) and heterogeneous deskmate (maximum level of intergroup contact). For the comparison between ingroup contact versus pure control, we can compare those not in the program (pure control group) and homogeneous classrooms (minimum intergroup contact in treatment).

Reported in Abstract. This comparison is not in the abstract.

Specification. Same as above. Results are obtained from Tables A53-A57, column 10/1 outgroup/ingroup comparison.

SDs. SDs obtained from available data (*data_APSR.dta* and *data_APSR_long.dta* for game outcomes). For the control group SD, we obtain the SD of the outcome for those not in the program.

A.3.2.2 Broockman and Kalla (2016) Durably reducing transphobia: A field experiment on door-to-door canvassing

Pre-registration link.

We use the available data.

Outcomes. According to pre-registration, registration prior to researcher analysis of outcome data (according to PAP, they had looked at outcome data but without the real treatment indicator), so outcomes are considered pre-registered before analysis.

1. Transgender Tolerance Scale: In the PAP they pre-registered as the main outcome a General Trans Acceptance Attitudes index (PAP page 3). This is almost the same in the paper except that they removed the law questions from it (supplementary materials page 5). Measured at 4 different time points. Includes both explicit measures and political and cultural attitudes.

Treatment. The paper's main treatment of interest is conversation targeting antitransgender prejudice, vs placebo conversation (i.e. recycling; page 222). Our treatment of interest is conversation with transgender (they revealed their identity) vs nontransgender canvass leader, regardless of the type of conversation, so treatment is *High versus no outgroup contact*.

Both the intervention (conversation about transphobia) and the placebo (conversation about recycling) had trans and non-trans canvassers. The paper is not super explicit about it, but I could confirm it with the data (I did tab identiy_canvasser treatment indicator and all cells were non-empty). It's not clear if they disclose their identity in the placebo but we keep the placebo to obtain the results anyways (and this way we are consistent with other papers, given that we do not check in the other papers whether the identity of the outgroup was explicitly revealed).Canvassers were randomly assigned: "The groups of households (turf) were then randomly assigned to pairs of canvassers by having canvassers pick a number corresponding to a turf out of a hat. Then, canvass leaders flipped a coin to determine which canvasser would knock on A doors and which on B doors" (page 3 supplementary materials). However, results are conditional on the participant opening the door, and this is the only data available with canvasser identity.

Allport Conditions. No discussion in the paper. We can infer equal status (canvassers and participants play different roles but there's no difference in status), support from authority, and cooperation (i.e. participants answering to canvasser questions rather than closing the door) without a common goal.

Duration of contact. 10 minute conversation (page 224).

Days Since Contact Ending and Measurement. Follow-up surveys began 3 days, 3 weeks, 6 weeks, and 3 months after the intervention, and were open for approximately 2 weeks (supplementary materials page 4; use days at which they began).

Reported in Abstract. "These effects [reduced transphobia] persisted for 3 months, and both transgender and nontransgender canvassers were effective." The 3 months covers all 4 time periods of the tolerance scale, but no mention here of the effect of contact per se.

Specification. The specification is OLS with cluster-robust standard errors, clustering on household, residualizing using pre-treatment covariates from the baseline survey and voter list, and adjusting for the contact rate (i.e. compiler average causal effect estimation rather than ITT, doesn't change t-stat or p-values but increases point estimates slightly; supplementary material page 11, lines up with PAP).

Table S5 (page 35 supplementary materials) is the only result that uses the identity of the canvasser. From what I understand, this table estimates average treatment effects for trans canvassers and non-trans canvassers separately (see *Broockman-Kalla-SM.R* line 1243). Notice the caption in Table S5: there's no comparison of treatment effects by identity of the canvasser.

We use the available data (obtained from running *Broockman-Kalla-SM.R* until line 410, before "Estimation Procedures") to obtain the comparison of trans vs non-trans canvassers, pooling all treatments together (including placebo; see code for regressions result from treatment and placebo conditions separately). We regress the outcome on the identity of the canvasser with controls and clustering standard errors on canvasser ID*; point estimate is not affected significantly from adding the treatment indicator. To check we have the right variables and understand their specification, we replicate Table S5 as closely as possible except that we don't use their method to calculate cluster robust standard errors (it says "from Mahmood Arai" in their code) nor we adjust the estimate and standard errors using the contact rate**, so we obtain the same t-stat but different point estimates and standard errors (our p-values are also double because theirs are one-sided).

*Ideally, we would cluster at the turf level, which was the level of random assignment of canvassers. However, this variable is not available. The table below shows a balance check we did for each of the baseline control variables (and the outcome at baseline) on the indicator for whether the canvasser is trans (we used the same dataset as described above).

**We do not do the contact rate adjustment (IV re-scaling) because with intergroup contact, you are "treated" as soon as you open the door and meet the canvasser. Whereas with the effects of canvassing, you are plausibly only treated once you hear what the canvasser has to say (hence why they rescale to get the effects of the canvassing).

SDs. Standard deviation for control (i.e. placebo) group is obtained from Table S23 (supplementary materials page 55).

We obtain standard deviations from the available data (same dataset as described above) for the control group and full sample in our sample of interest (i.e. if e(sample) == 1 after running our specification).

Table A3: Balance Table Broockman and Kalla (2016)

	Coef	Cluster SE	N
miami_trans_law_t0	0.24	0.20	482
miami_trans_law2_t0	0.08	0.17	482
therm_trans_t0	3.42	2.47	482
gender_norms_sexchange_t0	0.20	0.14	482
gender_norms_moral_t0	0.13	0.14	482
gender_norms_abnormal_t0	0.05	0.12	482
ssm_t0	-0.09	0.25	482
therm_obama_t0	6.16	4.33	482
therm_gay_t0	4.06	2.98	482
vf_democrat	0.06	0.06	482
ideology_t0	-0.02	0.15	482
religious_t0	0.04	0.18	482
exposure_gay_t0	-0.02	0.03	482
exposure_trans_t0	-0.01	0.03	482
pid_t0	0.14	0.27	482
sdo_scale	0.00	0.04	482
gender_norm_daugher_t0	-0.02	0.12	482
gender_norm_looks_t0	0.01	0.14	482
gender_norm_rights_t0	-0.02	0.11	482
therm_afams_t0	2.90	2.53	482
vf_female	-0.00	0.04	482
vf_hispanic	-0.00	0.08	482
vf_black	0.05	0.08	482
vf_age	0.30	2.20	482
survey_language_es	-0.04	0.02	482
cluster_level_t0_scale_mean	0.12	0.10	482
transtolerancedvt0	0.14	0.11	482

A.3.2.3 Grady et al. (2023) How contact can promote societal change amid conflict: An intergroup contact field experiment in Nigeria

Pre-registration link.

There are 2 PAPs associated with this pre-registration, one from 2018 and one from 2023. Below I refer to the one from 2018. The PAP is not specific about what are primary/main outcomes so we consider all the outcomes mentioned as primary. There's also a PAP Deviation document from 2023, which I refer to below.

The study was originally meant to be an RCT at the community level and at the individual

level, but due to compliance issues, the design was RCT only at the community level (Study 1; page 4 PAP). Thus, I'm only considering community level outcomes.

Outcomes. Below are the 4 outcomes in the paper with how they are broken up and presented. Outcomes are described on page 4 in the paper. For survey items, compare the questions in each index in the paper (Table S37) vs in the PAP (page 19).

1. Intergroup contact: They measured intergroup contact with behavioral monitoring of farmer–pastoralists interactions in markets and social events*, a survey index, and a survey experiment.

    (a) Self-reported Contact: This corresponds to the survey index. The survey outcome measures are registered in the PAP as "Social Contact", although in the PAP there were a few more questions included.

    (b) Contact Willingness: "A survey experiment, which we call the percent experiment, informed us about respondents' willingness to engage in contact, depending on the presence of outgroup members". This was pre-registered as "randomization experiment" (page 5 PAP deviation). Asks participants willingness to engage in hypothetical behaviors.

    (c) *Outgroup Event Attendance

    (d) *Farmers in Market

    (e) *Pastoralists in Market

    *The behavioral monitoring was also pre-registered, but it wasn't originally meant to be part of their analysis: "We collected more behavioral observation data than we anticipated. Instead of collecting behavioral observation data only in treatment sites, we also collected it in control sites" (page 4 PAP deviation). However, that same document states that they decided to gather observational data in control sites after baseline survey data collection. So since presumably the change was made before analysis, these are eligible too. These and (1) will be considered *Includes specific people met* because in theory all 3 can increase because person X met outgroup member person Y as part of the intervention, and then continues to see them after.

2. Perceptions of physical security: Survey questions combined into an index, where high values indicate security. This is in the PAP as "Perceptions of insecurity due to conflict", although in the PAP there were a few more questions included.

3. Intergroup attitudes: They measure intergroup attitudes with a survey index and an endorsement experiment.

    (a) Self-reported Attitudes: This combines elements under "Social Cohesion", "Outgroup Trust", and "Social Distance" in the PAP.

    (b) Endorsement Experiment: Appears in PAP. We asked respondents how much they would support a water policy if it was endorsed by a farmer organization (asked of

pastoralists), if it was endorsed by a pastoralist organization (asked of farmers), or if no endorsement was mentioned (the control condition posed to both pastoralists and farmers).

4. Intergroup cooperation: They measure intergroup cooperation with donations in a natural-field public goods game (pre-registered in PAP).

   (a) Public Goods Donation: Proportion of individuals who donated to the public good.

   (b) Public Goods Amount: Average donation amount in each community.

All of the above except (2) (considered *Other*) are considered *Generalized to outgroup*, especially since the analysis is at the community level and the majority of respondents did not interact with the outgroup through the intervention.

Beyond these, below are other constructs in the PAP that are not in the paper.

- Perceptions of economic benefit.
- List experiment .
- Threat perception.
- Violent conflict history.
- Shared resources.
- Dispute resolution.
- Acceptability of violence.

Treatment. "We randomly assigned communities with ongoing farmer–pastoralist violence to receive a contact-based intervention or serve as a control group. The intervention formed mixed-group committees [each joint project committee included an even number of farmers and pastoralists, as well as women and youth representatives, and totaled between 12 and 15 members] and provided them with funds to build infrastructure that would benefit both communities (...). The program also provided mediation training to each community's leaders and held forums where the groups discussed the underlying drivers of conflict" (page 2). Treatment group is communities that received the intervention and control group is communities that didn't. Treatment is bundled because contact with outgroups through the intervention is bundled with the many aspects of the intervention.

Note that the majority of survey respondents did not have intergroup contact: "In intervention sites, community members who did not participate directly in the contact interventions composed the vast majority of the sample. This design gives us two datasets to analyze. First, we create community-level survey data by aggregating the survey respondents within each community at baseline and endline. Second, we have observational data for social and market behaviors for each site at baseline and endline (page 3)." Only 52 of over 1000 respondents in the

treatment group had exposure to mediation [second aspect of the intervention] at endline (page 2 supplementary materials).

Allport Conditions. All conditions satisfied; from page 3: "The intervention was designed with contact theory in mind. Specifically, groups 1) cooperated with, 2) equal status to achieve, 3) shared goals with, and 4) support of local authorities."

Duration of contact. After 18 mo, they surveyed another approximately 50 randomly selected respondents per community in an endline survey (page 3). Reiterated as an "18 month program" in supplementary materials (page 2). However, since most participants didn't have contact through the intervention at all, we assign it 15 minutes to keep it as a light touch intervention (which is the case for most respondents).

Days Since Contact Ending and Measurement. For endline outcomes (includes public goods game*), above it states that the endline happened after 18 months of the intervention, so 0 days. For observational outcomes, "In the months immediately after the baseline survey and immediately before the endline survey, we collected observational data on farmer–pastoralist interactions in shared markets and at social events (page 3)." So 0 days as well.

*There's no explicit mention of whether the percent experiment and the endorsement experimented were implemented at endline, but it seems likely to be the case.

Reported in Abstract. "We find those who lived in the communities that received the intervention had more positive intergroup attitudes and feelings of physical security, as well as were more likely to engage in voluntary intergroup contact measured through self-reports and observed behavior in markets." Covers (3a), (2), (1a), and (1e) ((1d) and (3b) not covered because results are not significant).

Specification. (Page 3; changed a little but not significantly from original PAP) They use linear regression, with randomization inference for p-values, bootstrapping for standard errors, fixed effects for states, and one-sided tests. When treatment groups are balanced on the baseline outcome, they use the baseline outcome as a covariate to predict the outcome at endline. When treatment groups are not balanced on the baseline outcome, they use the change score of the outcome as the dependent variable (and baseline measure is not a covariate). For the observations of market behavior and social events, they cluster errors at the site-level because there's multiple observations (i.e. measured at several timepoints) per site.

Results are only available in Figure 1. Authors shared a table with the results. For standard errors, we divide the whole standardized interval by 1.96*2. Since the interval provided for *Outgroup Event Attendance* doesn't line up with the figure, we eyeball it from the figure.

SDs. Results in Figure 1 are standardized with the baseline value of the outcome (page 4).

A.3.2.4 Zhou and Lyall (2023) Prolonged contact does not reshape locals' attitudes toward migrants in wartime settings

Pre-registration link.

We use the available data.

Outcomes. The original focus of the RCT was on the economic intervention and outcomes, so the outcomes on intergroup contact were pre-registered as secondary. We will consider these outcomes anyway given that the intergroup contact intervention was also secondary in the pre-

registration. There are 5 PAPs, but from the very first, the outcome "attitude towards refugees" is included. Then in the April 2016 PAP, there's more details on the precise survey questions used. Registration is prior to any research activities so outcomes are considered pre-registered before analysis.

The outcomes on attitudes towards refugees are the answers to Q4.7-Q4.10 (page 10 PAP April 2016), plus Q4.6 which is about interaction with immigrants. Below are the questions, which is how they appear both in the paper ("outcome measures", page 6) and PAP. The scales are such that higher values are more positive attitudes towards refugees, and the questions are only asked to locals (refugees are not asked questions about locals).

1. Interaction with migrants: Thinking about the past six months, approximately how much interaction did you have with refugees/migrants outside of the INVEST program in your community?

2. Perception of migrants: In general, what kind of effect do you feel that refugees/migrants have on your community?

3. Migrants more likely to support violence: Some people believe that these refugees/migrants are more likely to support violence than local residents of Kandahar. Others disagree. Do you...

4. Migrants take jobs away from locals: Some people believe that these refugees/migrants will take employment away from native residents of Kandahar. Others disagree. Do you...

5. Migrants are a burden on resources: Some people believe that these refugees/migrants will become a burden on government resources by requiring welfare assistance. Others disagree. Do you...

Treatment. The intervention consisted of Technical Vocational Education and Training (TVET) and unconditional cash transfer (UCT), where only TVET involves intergroup contact. Participants were assigned to TVET treatment - UCT treatment, TVET treatment - UCT control, TVET control - UCT treatment, and TVET control - UCT control (i.e. pure control group; page 5). All courses had an average of 40 students each and were naturally mixed in local–migrant composition (page 6). The treatment group includes those that received TVET and the control group includes those that didn't. The treatment is bundled with the educational and training components (but orthogonal to the unconditional cash transfer component).

Allport Conditions. All conditions satisfied; from page 3: "INVEST participants were young people who experienced equality within the classroom; collaborative classroom tasks and shared broader goals, including graduation; and substantial support from local and international authorities."

Duration of contact. Participants could choose to enroll in 3 or 6 month courses (prior to randomization; page 5). 360-720 hours according to the abstract ("the most sustained duration of intergroup contact... experimentally evaluated to date"). We use the average; 540 hours.

Days Since Contact Ending and Measurement. Endline survey after 6-month courses ended (2.5 months after the 3-month courses ended), and a second endline 8 months later (page

6). According to the study timeline in Appendix F (page 7 supplementary materials), it seems to be 8 months for the 3-month courses, not the 6-month ones, so for the 6-month ones is 5 months. For endline 1, since for 3-month courses is 75 days and for 6-month courses is 0 days, I use the average of 38. For endline 2, the average is (240 + 150)/2 = 195 days.

Reported in Abstract. "While the program provided the most sustained duration of intergroup contact (360–720 h) experimentally evaluated to date, we find no evidence of reported behavioral or attitudinal change by locals (N = 1276) toward migrants generally, regardless of classroom demographics or course duration." All outcomes considered reported in abstract.

Specification. They estimate "ITT effects using a non-parametric analysis approach based on the difference-in-means estimator that accounts for the block randomization design" (page 13).

Results are in Figure 3 (page 16). To obtain table results, we use the available code: follow instructions in their read me file to compile *INVEST_ProlongedExposure_Paper_Final.Rnw*, and then obtain results for endline 1 in *RefAttPlot1.Rdata* and *RefAttPlot2.Rdata* for endline 2). Sample size is obtained from the available data (INVEST_Panel.csv). We use the point estimate and standard error for "All" for each of the 5 outcomes, at endline 1 and endline 2. Considered as obtained in the main paper since the estimates were available in the figure, just not with the precise number.

SDs. No SDs for the outcomes (there's only SDs for the full sample at baseline). Obtained from data (*INVEST_Panel.csv*).

A.3.2.5 Kalla and Broockman (2020) Reducing Exclusionary Attitudes through Interpersonal Conversation: Evidence from Three Field Experiments

There are 2 EGAP registrations associated with this paper. The first experiment corresponds to the 2018 pre-registration, and the second and third correspond to the 2016 pre-registration. Both pre-registrations are prior to realization of outcomes, so outcomes are considered pre-registered before analysis.

We use the available data.

EXPERIMENT 1:

Outcomes. PAP lists 5 outcomes:

1. Site-Specific Outcome
2. Anti-Immigrant Prejudice Index
3. Anti-Immigrant Policy Index
4. Perspective Taking
5. Active Processing

Results by canvasser immigrant status are only presented for an index of all the items in the prejudice and policy indices (*Overall Index*) as measured in the first post-treatment survey (outcomes 2 and 3 above, page 19 online appendix). The remaining outcomes are not considered as

"Not in the paper" because this heterogeneity analysis was not the main focus of the paper. The pre-registration doesn't specify which outcomes would be tested in the heterogeneity analysis (it only says "Canvasser Heterogeneous Treatment Effects", page 3 PAP).

Treatment. Voters were randomly assigned to receive a long treatment conversation on immigration (Full Intervention), a short treatment conversation on immigration (Abbreviated Intervention), or a placebo conversation (Placebo) (page 415). The canvasser could be an immigrant or not.

The paper is not explicit about it, but it seems that canvassers in the placebo condition could be immigrants as well, because the results are presented relative to the placebo (Table OA12, page 20 appendix); this can also be further checked in the data.

Canvassers were randomly assigned (page 9 supplementary materials): "The groups of households (turf) were then randomly assigned to pairs of canvassers by having canvassers pick a number corresponding to a turf out of a hat. Then, canvass leaders flipped a coin to determine which canvasser would knock on A doors and which on B doors." However, results are conditional on the participant opening the door, and this is the only data available with canvasser identity.

I've checked the procedures in the paper and it doesn't seem that canvassers had to systematically disclose whether they are immigrants or not. However, part of the treatment procedure was to exchange narratives, and the canvasser would share their immigration story, but it doesn't seem that this was necessarily their personal story (in the case they are an immigrant). See the description of the procedure in the appendix (page 3):

Exchange narratives about personal experience with immigration. *The canvasser then asked the voter if they know anyone who is an immigrant and, in particular, an unauthorized immigrant. If the voter knows someone, the canvasser would have the voter talk about how they know this person, their immigration story, and how it must feel to be an immigrant. Whether or not the voter knows an immigrant, the canvasser would always share their immigration story. This might be a personal story or about a friend or family member. The canvasser would end this section by asking the voter if there is anything about the story that they can relate to, encouraging perspective taking.*

However, when presenting the results for this analysis the paper says (page 418): *There was little meaningful treatment effect heterogeneity by canvasser or voter attributes; the conversations were broadly persuasive regardless of which canvassers or voters were involved. Online Appendix Table OA.12 shows that the effects of the Full Intervention are similar regardless of whether the canvasser is an immigrant or is not an immigrant. The clearly significant effects for non-immigrant canvassers mean the effects cannot be attributed to mere contact and that voters need not be prompted to take canvassers' own perspective for the intervention to be effective.*

Allport Conditions. Footnote in page 413 says "voters' contact with canvassers met few of the conditions", but there's no further details. We can infer equal status (canvassers and participants play different roles but there's no difference in status), support from authority, and cooperation (i.e. participants answering to canvasser questions rather than closing the door) without a common goal.

Duration of contact. On average 11 minutes in the Full Intervention, 5 minutes in the Abbreviated intervention, and 1 minute in the Placebo condition (page 414). Since we pool conditions, I average it to be 5 minutes.

Days Since Contact Ending and Measurement. Follow-up surveys began 4 days, 30 days, and 3-6 months after the conversation (page 415), but results for this analysis are only shown for first follow-up.

Reported in Abstract. No, because there is no mention about the immigrant vs non-immigrant comparison.

Specification. As pre-registered, the specification is OLS with cluster-robust standard errors, clustering on household and also including the pre-treatment covariates (page 14 online appendix). The only results by identity of canvasser are in Table OA12 (page 20 online appendix). Index is coded such that higher values indicate more tolerance (page 13 online appendix). We use the available data (immigration_data.csv) to fully replicate the table, and then set up a regression to obtain the effect of canvasser identity on the overall index (clustering on canvasser characteristics* and including covariates and site fixed effects). Results are similar if we add treatment indicators or if we obtain the effect separately for treatment vs placebo conditions.

*Ideally we would cluster at the turf level, at which canvassers were randomly assigned, but this data is not available. We cluster on a variable grouping canvassers gender, latino, and age (canvasser characteristics available). Below is a balance check of covariates on the indicator for the canvasser being an immigrant (using immigration_data.csv).

SDs. We obtain standard deviations from the available data (using immigration_data.csv) for the control group (i.e. placebo conversation) and full sample in our sample of interest (i.e. if e(sample) == 1, after running our specification).

EXPERIMENT 2:

*We will not consider the third experiment because it did not randomize intergroup contact (intervention is phone call but there's no variation in the identity of the caller).

Outcomes. Several variables combined into one scale measuring the overall effect of the conversations on prejudice towards transgender people (PAP page 3), called *Overall Index* in the paper.

Treatment. Participants were randomly assigned to *Video Narratives Only, Participants' and Video Narratives*, or *Placebo* conditions (page 419).

The scripts suggests that canvassers had a space to share they're transgender in the treatment conditions, but not clear whether all trans canvassers disclosed their identity, and whether they did so in the placebo (I believe they didn't, based on the script). We nevertheless keep the placebo when obtaining the results (consistent with other papers where we do not require outgroup members to disclose their identity).

Our comparison of interest is having a conversation with a trans vs non-trans canvasser, which is orthogonal to the type of conversation, so treatment is *High versus no outgroup contact*.

Canvassers were randomly assigned (page 33 supplementary materials): "The general survey recruitment procedures and experimental design were identical to Experiment 1 except as otherwise noted below." However, results are conditional on the participant opening the door, and this is the only data available with canvasser identity.

Allport Conditions. Footnote in page 413 says "voters' contact with canvassers met few of the conditions", but there's no further details. We can infer equal status (canvassers and participants play different roles but there's no difference in status), support from authority, and cooperation (i.e. participants answering to canvasser questions rather than closing the door) without a common goal.

Table A4: Balance Table Kalla and Broockman (2020) Experiment 1

	Coef	Cluster SE	N
t0_imm_better_worse	0.02	0.06	1572
t0_imm_police	0.06	0.10	1572
t0_imm_driverslicense	-0.07	0.11	1572
t0_imm_daca	0.01	0.08	1572
t0_imm_citizenship	0.13	0.08	1572
t0_imm_deportall	0.06	0.11	1572
t0_imm_attorney	0.01	0.09	1572
t0_imm_prej_living	-0.05	0.08	1572
t0_imm_prej_neighbor	-0.04	0.08	1572
t0_imm_prej_speaking	0.02	0.07	1572
t0_imm_prej_workethic	-0.00	0.03	1572
t0_imm_prej_fit	-0.05	0.08	1572
t0_imm_know	-0.03	0.03	1572
t0_social_distance_immigrant	-0.02	0.12	1572
t0_therm_illegal_immigrant	0.98	1.86	1572
t0_therm_legal_immigrant	1.37	1.21	1572
t0_college_educ	0.01	0.02	1572
t0_asian	0.01	0.02	1572
t0_latino	-0.00	0.03	1572
t0_black	0.01	0.01	1572
t0_white	-0.01	0.03	1572
t0_born_in_us	0.01	0.01	1572
t0_factor_undoc_immigrant	0.01	0.07	1572
t0_factor_lgbt	0.04	0.06	1572
t0_factor_trump	0.01	0.07	1572
vf_age	-1.21	1.11	1572
vf_voted08	-0.02	0.03	1572
vf_voted10	-0.01	0.03	1572
vf_voted12	0.01	0.03	1572
vf_voted14	0.01	0.03	1572
vf_voted16	-0.01	0.02	1572
vf_female	0.07	0.03**	1572

Duration of contact. 7.7 minutes in *Video Narratives Only* condition and 10.5 minutes in *Participants' and Video Narratives* condition (page 419).

Days Since Contact Ending and Measurement. Follow-up surveys began one week and one month after the intervention (page 419). But results for this analysis are only shown for

first follow-up.

Reported in Abstract. No, because there's no mention to transgender vs non-transgender comparison.

Specification. As pre-registered, the specification is OLS with cluster-robust standard errors, clustering on household and also including the pre-treatment covariates (page 35 online appendix).

Table OA52 presents ATE results on the overall index in the 1 week survey by the identity of the canvasser (online appendix page 45; we were able to replicate the table results with the available data). Since we are interested in the effect of contact, we set up our own regression of the outcome on the identity of the canvasser, with controls, site fixed effects, and errors clustered on canvasser ID*; adding the treatment indicators doesn't affect our relevant point estimate much. We obtain the data from running *SMTrans_replication.R* until line 214, before "Estimation Procedures". We only considered canvassers with known gender identity (i.e. no "missing data"). We include the placebo condition in our specification, see code to see the results separately for the placebo and the treatment conditions.

*Ideally, we would cluster at the turf level, which was the level of random assignment of canvassers. However, this variable is not available. The table below shows a balance check we did for each of the baseline control variables on the indicator for whether the canvasser is trans (using the same dataset as described above).

SDs. We obtain standard deviations from the available data (same dataset as described above) for the control group and full sample in our sample of interest (i.e. if e(sample) == 1 after running our specification).

A.3.2.6 Paler et al. (2020) How Cross-Cutting Discussion Shapes Support for Ethnic Politics: Evidence from an Experiment in Lebanon

Pre-registration link.

Outcomes. Registration is prior to researcher access to outcome data so outcomes are considered pre-registered before analysis.

Appendix A panel A in PAP has a full list of outcomes (page 31 PAP). Below is a list with all of these outcomes and how they appear in the paper:

1. Closeness to same class/same sect → *Sectarian and class in-group (1-7)*
2. Closeness to same class/other sect → *Sectarian out-groups and class in-group (1-7)*
3. Closeness to other class/same sect → *Sectarian in-group and class out-group (1-7)*
4. Closeness to other class/other sect → *Sectarian and class out-groups (1-7)*
5. Group dist rich/poor same sect → *Closeness btwn diff classes in same sect (1-7)*
6. Group dist rich diff sect → *Closeness btwn poor of diff sects (1-7)*
7. Group dist poor diff sect → *Closeness btwn rich of diff sects (1-7)*

Table A5: Balance Table Kalla and Broockman (2020) Experiment 2

	Coef	Cluster SE	N
t0_transpolicy_lgbtdiscrim	-0.11	0.08	980
t0_transprej_comfortwork	-0.15	0.13	980
t0_transprej_moralwrong	0.22	0.16	980
t0_transprej_moralgenderchange	0.33	0.16**	980
t0_transprej_restroom	0.01	0.19	980
t0_transprej_friend	-0.24	0.18	980
t0_transvid_lgbtdiscrim	-0.06	0.10	980
t0_transvid_fire	-0.13	0.09	980
t0_transvid_school	-0.09	0.13	980
t0_transvid_comfortbathroom	0.08	0.21	980
t0_transvid_restroom	-0.04	0.19	980
t0_transvid_predator	-0.16	0.22	980
t0_transvid_teacher	0.04	0.15	980
t0_therm_trans	-3.53	1.92*	980
t0_trans_factor_hh	-0.07	0.09	980
t0_partisan_factor_hh	0.07	0.09	980
t0_religion_monthly	0.01	0.03	980
t0_social_trans	-0.05	0.03	980
vf_afam	0.02	0.02	980
vf_white	-0.05	0.03*	980
vf_dem	0.02	0.03	980
vf_rep	0.01	0.04	980

8. Cooperation → Total group contribution round 1 (*Round 1*), total group contribution round 2 (*Round 2*), difference between round 2-round1 (*Difference*).

9. Resource Allocation, specifically Allocation to non co-sectarian districts → *Share for non-cosectarian districts* (there's several others in the paper, but this is the only one considered main in the pre-registration).

10. Support for multi-sectarian policies → Not in the paper.

11. Support for multi-sectarian political action → *Proportion Signed*

12. Support for sectarian political action → Not in the paper.

(1)-(4) capture the extent to which an individual feels close to (identifies with) a particular social group, and is measured on a scale from 1-7 where 7 means they wholly identify with the other group (page 57). Last bullet point on page 18 of the appendix details how (5)-(7)

are constructed, but in sum, they are similar to the exercise above, except that they capture perceived closeness between the groups themselves rather than between the individual and the group; a more positive number means that they perceive the two groups compared to have more in common.

We drop (1) because it is not clear we would expect or want an effect on ingroup identity, so we focus just on the outgroup attitudes associated with the outgroup dimension being compared.

For (8), they have several outcomes, but only the three listed above are group-level outcomes. These are the ones reported in the paper and are the only ones that seem relevant, but they are ruled out since they are mechanically affected.

For (11) it was pre-registered as an index with survey items plus the proportion that signed a petition (*pet_sign_final* in Appendix A in the PAP). The index or the individual survey items are not in the paper, but they did note in the pre-registration that they would analyze components separately (i.e. this covers the analysis of the petition measure). See the description of the petition in the paper (page 49): *"The petition, sponsored by LCPS, embodied many of the issues that emerged from the protests by denouncing the sectarian status quo; calling for electoral reforms to reduce the influence of sectarian parties; and demanding more policy-making on the basis of economic and programmatic priorities."*

(10) and (12) are not in the paper.

All outcomes are *Generalized to outgroup* type. (5)-(7) is *Generalized to outgroup* because in at least one of the categories in one of the groups of the comparison they'll have an outgroup.

Treatment. Participants were randomly assigned to six-person discussion with different sectarian and class compositions; (1) same-sect, same-class, (2) mixed-sect, same-class, (3) same-sect, mixed-class, and (4) mixed-sect, mixed-class (page 43). Comparison is *High versus no outgroup contact*.

The paper has multiple types of contact (i.e. sect/class, conditional on same/mixed class/sect), but the paper doesn't describe any as better than the other (see page 5 PAP), so we will record all comparisons. We will do a sect comparison and a class comparison, with the outcomes that are relevant for each. For each outcome in a comparison, we will do the sect/class comparison conditional on same/mixed class/sect.

For sect comparison, the relevant outcomes are (2), (4), (6), (7), (9), and (11). (6) and (7) will be included for now, but since participants can be comparing two groups of outgroups, it's unclear what could be expected from the treatment.

For class comparison, the relevant outcomes* are (3), (4), and (5).

*(11) not included because it isn't obvious that cross-class exposure should affect signing. Also cross-class exposure may affect each class group differently; rich people should sign more when exposed to poor people, because the petition is about policy based on economic need, but this can be the opposite for poor people exposed to rich people.

Allport Conditions. No discussion in paper. We can infer support from authority, and equal status, but no common goal or cooperation. Note that: "Nevertheless, we intentionally did not include a collaborative exercise — as is common in intergroup dialogue and positive contact interventions" (page 47).

Duration of contact. Discussions lasted 60 minutes (page 45).

Days Since Contact Ending and Measurement. Follow-up survey immediately after the discussion (page 48).

Reported in Abstract. "Our evidence suggests that cross-sectarian discussion resulted in less support for sectarian politics but only when individuals also belonged to the same economic class." The outcome referred to in the abstract is *Proportion Signed,* given that "We capture our main outcome of interest — support for sectarian versus cross-sectarian, programmatic politics — using two measures rooted in news headlines at the time of the study. Our main behavioral measure is willingness to sign a petition condemning the role of sectarianism in Lebanese politics and demanding a programmatic alternative" (page 49). This is the only outcome mentioned in the abstract.

Specification. Paper's specification is weighted least squares regression with dummies for each discussion group type (omitted group is same-sect, same-class), controls, and randomization block fixed effects (i.e. set (same-sex), sect, and class, plus recruiter and neighborhood when possible), and errors clustered at the discussion level (page 50). This differs from the pre-registered specification in that the pre-specified one had dummies for mixed-class and mixed-sect, and the interaction between the two (page 23 PAP). Since it's just a different way of presenting the same results, we will use the specification in the paper.

Results with their main specification are in Appendix J1 (page 29 appendix), which has an indicator for each discussion group type. Results in the main paper are in Figure 1 for outcome (11), Figure 2 for outcome (9), and Figure 4 for outcomes (1)-(4) (outcomes (5)-(7) are not in the main paper). There are also results with their pre-registered specification showing the interaction effect of the mixed-sect and mixed-class treatments in Appendix J2 (page 32 appendix; outcomes (5)-(7) are not available with this specification).

Results for outcomes (9) and (11) are in Table J1 (page 29 appendix), and results for outcomes (1)-(7) are in Table J2 (page 30 appendix). Neither of these tables specify the sample size, but I assume this is equivalent to the sample size for dependent variables listed in Table G1 (page 21 appendix; corroborated with Tables J3 and J4 in Appendix J2). Errors are clustered at the discussion level: there were 120 discussions, 30 for each type (page 43). I could not obtain sample size or number of clusters for relevant treatment arms, but we could assume N it's 713/4 = 178 participants in each group: there were 30 6-person discussions of each type, so there were 30*6 = 180 participants in each treatment, but 713/720 participants completed the study.

For sect comparison conditional on same class (compares (1) same-sect, same-class to (2) mixed-sect, same-class), the point estimate is B1 and the standard error is SE1. For sect comparison conditional on mixed class (compares (3) same-sect, mixed-class to (4) mixed-sect, mixed-class), the point estimate is B3 - B2, and we impute the standard error as the average of SE2 and SE3.

For class comparison conditional on same sect (compares (1) same-sect, same-class to (3) same-sect, mixed-class), the point estimate is B2 and the standard error is SE2. For class comparison conditional on mixed sect (compares (2) mixed-sect, same-class to (4) mixed-sect, mixed-class), the point estimate is B3 - B1, and we impute the standard error as the average of SE1 and SE3.

SDs. Standard deviation for the full sample is obtained from Table G1 (page 21 appendix).

A.3.2.7 Asimovic et al. (2024) Estimating the effect of intergroup contact over years: Evidence from a youth program in Israel

Pre-registration link.

Outcomes. Updated PAP is from January 2020, and it states that data had not been consulted yet, so outcomes are considered pre-registered before analysis. The outcomes are measured with a series of indicators which are summarized in indices.

1. Outgroup regard (called "Prejudice" in PAP). Includes the following: social distance (behavior), support for peace process (political and cultural attitudes), perspective taking (behavior), hostile attribution by subjects/peers (indirect measure), optimism about peace (unrelated), ingroup identity steem (unrelated).

2. Self-esteem.

3. Ingroup regulation. Includes: effort to persuade (behavior), ingroup censuring (behavior), and perspective sharing (unrelated).

Table 4 (page 18) lists the items in outcome (1) and (3).

Treatment. Participants are recruited and randomly assigned to the program (treatment group) or put on a waiting list (control group), so treatment is bundled (page 10). Participants in the treatment group are invited to join a team from the same residential community, ethnic group, and gender. The team practices weekly with their ingroup, but after a month joint practices between an Arab-Palestinian team and a Jewish-Israeli team begin (5-8 times throughout the season) (page 7).

The paper shows results both for short-term with the RCT design, and multi-year exposure with a fusion design (i.e. not an RCT). We use the RCT results only.

Allport Conditions. All conditions satisfied; from page 8: "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 providing authority sanctioning."

Duration of contact. The intervention lasts a season, which seems to be 7 months (page 11). There's 5 to 8 joint practices. No further duration of length. For our best guess will do 2 hours * 6.5 practices.

Days Since Contact Ending and Measurement. Endline is collected at "the end of the season" (page 11). Could not find any further details so I'll assume 0 days.

Reported in Abstract. "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." No mention of self-esteem (outcome 2).

Specification. ITT linear regression specification with treatment dummy, individual covariates, and year-specific and site-specific fixed effects (page 6 appendix). Follows PAP.

Results for *Outgroup regard* are obtained from column 1 in Table 5 (page 22). Results for *Ingroup regulation* are obtained from column 1 in Table 7 (page 24). Results for *Self-esteem* are in column 3 in Table 23 (appendix page 16).

SDs. SDs for the control group are at the bottom of the tables.

A.3.2.8 Rossiter (2023) The Similar and Distinct Effects of Political and Non-Political Conversation on Affective Polarization

Pre-registration link.

Outcomes. Pre-registration is for study 2 and from September 2020; study 2 was conducted in the fall of 2020 (page 21), but pre-registration specifies that it was done prior to any research activities, so outcomes are pre-registered before analysis. There are 4 outcomes in PAP (page 7):

1. Outparty affect: Difference in pre- and post-treatment feeling thermometer ratings for Republicans/Democrats across the country.

2. Outparty trait stereotypes: Agreement with whether 4 positive and 4 negative traits describe Republicans/Democrats. The paper has two indices not described in the PAP, *All Negative Traits* and *All Positive Traits*, with each of the 4 negative/positive items. We will use these indices since they contain all (and only) the pre-registered items.

3. Future outparty contact: Willingness to have a conversation with Democrat/Republican, both for (a) non-political conversation (family) and (b) political conversation (immigration). *Future non-political/political conversation.* The question is asked in a hypothetical manner, without stakes.

4. Bipartisan views: Alignment of goals (i.e. compatible goals) between Democrats and Republicans, measured among (a) partisans and (b) elites. *Perception of bipartisanship among partisans/elites.*

All outcome types are *Generalized to outgroup.*

Treatment. Participants are randomly assigned to partnerships: one Democrat and one Republican in each. Partnerships are then randomly assigned to treatments. Participants can have either imagined or actual contact, and discuss either a political or a non-political topic (2 x 2 study; page 20). Treatment group is being assigned to have actual contact and control group is being assigned to have an imagined conversation. Treatment is bundled with social interaction (i.e. participants in the treatment group have intergroup contact through talking to an outgroup contact but also have a real, not imagined, conversation).

Rossiter does predict ex ante that non-political conversations will be a more effective form of contact: "Therefore, in general, I expect that non-political conversations will be more effective than political conversations at improving outparty affect and use of negative trait stereotypes to describe the outparty" (page 3 PAP). For now, we will report outcomes for political (in Dropped Effects sheet, because non-political was predicted to be more effective) and non-political conversations separately, conditional average treatment effects (CATE) in the paper. But they also

report average treatment effect (ATE) estimates of actual cross-partisan conversation, relative to imagined conversation, pooling both types of conversation topics (page 21).

Allport Conditions. From page 7: "However, Allport's influential "contact hypothesis" suggests that improved intergroup relations can result from intergroup contact if it meets several conditions—equal group status within the contact situation, common goals, intergroup cooperation, and the support of authorities, law, or custom (Allport 1954). Yet, cross-partisan conversation as a form of contact would presumably lack several of these conditions. For example, partisans engaging in an everyday conversation are not likely to be pursuing a shared goal, nor does the current American political environment and its elites necessarily support positive interactions amongst partisans." Based on this, the intervention doesn't have a common goal or cooperation, but in the context of this experiment (in contrast with everyday interaction) we can infer support from authority and equal status.

From page 28: "A second limitation is that this experimental design involved only two individuals, one from each party. While this helps satisfy one of Allport's conditions for contact to improve outgroup prejudice—equal status in the contact situation—not all conversations will avoid having a minority group or minority opinion apparent in the interaction."

Duration of contact. 8 minutes (page 11, which describes study 1, but page 21 says "following the same procedures used in Study 1...").

Days Since Contact Ending and Measurement. After completing their assigned task, participants proceeded to a posttreatment survey to measure outcomes (page 11).

Reported in Abstract. "Across two experiments, I find that conversation, whether politically-charged or not, decreases affective polarization. However, I find talking politics has distinct democratic benefits, providing greater opportunity to learn about the outparty and increasing willingness for future political conversations." I consider outcomes (1) and (3b) as reported in the abstract, even if (3b) refers to political topic treatment specifically.

Specification. The specification in the paper deviates from the one pre-specified because they include all participants in partnerships where both individuals completed the study, rather than participants in which the entire block of eight participants completed the study (see note 14 page 21). For the first two outcomes, they provide results with the pre-registered specification in the appendix, so I'm reporting those results. All models cluster standard errors for participants that had the actual conversation.

Results for outcome (1) are in Table A11 (page 15 appendix(*)). Results for outcomes under (2) are in Table A15; point estimate is reversed for *All Negative Traits*. N clusters is 3*N/4(**). For both tables(***) the first row under the outcome gives the effect for non-political topic and the one below for political topic. Outcome (1) and (2) is in the paper for a different sample, but we still consider it reported in the paper. Results for outcomes (3) and (4) are in Figure A4 (these are results with full partnerships rather than full blocks). Point estimates and standard errors obtained from the author. From Table A19 we can see that this analysis uses the full main analysis sample (N = 740) and from Table A9 we know that N = 378 for non-political conversation and N = 362 for political conversation.

*An interesting aside: page 14 appendix suggests that the control group still got an effective "treatment": "It is not that non-political conversation was a weak treatment—it increased outparty affect by an average of 8.55 degrees in this sample. Instead, imagined conversation was a strong baseline condition, increasing outparty affect by an average of 6.7 in this sample. Thus,

the effect of non-political conversation, relative to imagined conversation, was not significant."

**For each N/4 full blocks in the specification (out of 8 participants in each full block, 4 had a political/non-political topic), 2 people had an actual conversation and 2 had an imagined conversation. Since errors are clustered for actual conversation partners, that gives 3 clusters for each full block. Will assume the same for the results from Figure A4, which are not in the sample restricted to full blocks.

***According to pre-registration, they were going to use randomization inference. Tables report difference-in-means estimates, but they also include a p-value obtained with randomization inference.

SDs. Obtained SDs for the control group from the author. These are slightly different from our control group; the SDs are from the main analysis sample (where full partnerships completed the study), rather than for the sample in the appendix results (where full blocks completed the study), but we will use the SDs provided. Figure A4 is standardized, so we have full sample SD for outcomes (3) and (4).

A.3.2.9 Rossiter and Carlson (2024) Cross-Partisan Conversation Reduced Affective Polarization for Republicans and Democrats Even After the Contentious 2020 Election

Pre-registration link.

Paper published online as a short article on June 5. Notes below updated with that version. We use the available data.

Outcomes. Pre-registration is from January 2021, and according to it, the study started in February 2021. Primary outcomes are obtained from PAP under "Main Outcomes of Interest" (page 8 in PAP). For all of these, the outcome is the change between each participant's pre-treatment and post-treatment responses (page 3).

1. Outparty Affect: Feeling thermometer ratings for Republicans across the country / Democrats across the country.

2. Social Polarization: Average likelihood of engaging in different activities with someone from the outparty (i.e. a Democrat/Republican; "How likely is it that you would engage in the following activities?").

3. Election Integrity: Rate on a scale for how they think the elections were run and administered.

Treatment. Participants randomly assigned to cross-partisan partnerships. Partnerships were then randomly assigned to one of two conditions. Participants in treatment group partnerships were told their partner's partisanship, read a brief overview of the 2020 election, and discussed the election for eight minutes. Participants in the control group had an identical prompt but were asked to complete a short essay alone (page 3).

Allport Conditions. From page 2: "Specifically, "losers" will improve attitudes toward the out-party less because the conversation makes it hard to dispel emotional reactions to threat,

especially if perceptions of unequal status trickle into the interpersonal setting (Allport 1954). Because election outcomes alter partisan groups' status, and elections can produce differential feelings of threat, we expect threat experienced by electoral losers to decrease how much cross-partisan conversation reduces affective and social polarization relative to electoral winners." The above suggests that there was not equal status. We can infer that the contact treatment was supported by authority, but participants did not have a common goal nor they needed to cooperate.

Duration of contact. 8 minutes.

Days Since Contact Ending and Measurement. From page 3: After their conversation or short essay, participants completed a survey to measure our outcomes. Three days later, we followed up with participants to examine the durability of treatment effects for outparty affect.

Reported in Abstract. "However, for both sides, cross-partisan conversations reduced out-party animosity for at least three days, reduced social polarization, but did not increase perceptions of election integrity." The underlined phrases correspond to outcomes (1)-(4), respectively.

Specification. From page 3: Linear regression with cluster-robust standard errors for conversation partners* and block fixed effects. Sample average treatment effects (SATE) of conversation, relative to no conversation. Sample is restricted to all participants in partnerships that completed the experimental task and post-treatment survey.

*Table notes clarify that standard errors are clustered at the partnership level for individuals assigned to the conversation condition.

Results are obtained from Table A16 in Appendix N (page 48), columns 1, 3, and 5. Number of clusters is inferred from N in the treatment group (294) divided by two, plus N in the control group (284) (page 3). This lines up with what I find in the available dataset (*results.csv*).

Result for the follow-up is obtained from Table A15 in Appendix M (page 47). According to the paper and pre-registration, for this analysis they were going to include all participants that complete the follow-up, regardless of whether their partner also completed the item (page 4 appendix). However, their code and sample size (i.e. *outparty_change_t2* in *results.csv* is nonmissing for 481 observations, not 410) suggests that they restricted the sample to full clusters as in previous analyses. For N clusters, I obtain them from the available data (*results.csv*).

SDs. I couldn't find any standard deviations in the paper and outcomes don't seem to be standardized. I obtain the SD for the control for those in a full cluster from the available data (*results.csv*). Control group is those that completed an essay on their own ($z = 0$).

A.3.2.10 Porat et al. (2024) The Costs of Collaboration: Evidence From Two Field Experiments in Jerusalem

There are 2 EGAP registrations associated with this paper. The first experiment corresponds to the 2022 pre-registration, and the second corresponds to the 2023 pre-registration. According to the second pre-registration, the first one was a pilot, and the second intended to replicate in a larger sample, using a more scalable intervention.

EXPERIMENT 1:

Outcomes. Pre-registration is prior to realization of outcomes, so outcomes are pre-registered before analysis. The PAP specifies two sets of primary outcomes, one on learning and achieve-

ment, and another on intergroup relations (page 3). In the paper, outcomes are described on page 25.

Learning and achievement:

1. Learning: In the PAP, individual assignment handed in by students and teacher assessment of individual assignment. In the paper, the assignment was graded by the research team and the outcome is called *Assignment Score*.

2. Contribution to joint work: *Contribution* in the paper and corresponds to student's report of their own contribution. Drop because it is mechanically affected.

3. Comfort: Grouping of questions on comfort of working with partner and speaking English with partner. Drop because it is mechanically affected.

4. Partner allowed my participation: Not in the paper. Would drop anyway because it is mechanically affected.

5. I allowed partner participation: Not in the paper. Would drop anyway because it is mechanically affected.

6. Self-efficacy: Grouping of "I can succeed in this course" and "I am confident in my English abilities."

Intergroup relations:

1. Motivation to work with partner again: In the paper it is presented as a learning outcome and called *Future Work*. Drop because it is mechanically affected.

2. Motivation to work in heterogeneous groups in the future: Not in the paper. "To what extent would you like to work with a teammate whose mother tongue is [outgroup language]?" Explicit evaluation because they ask if they would "like".

3. Motivation to work in homogeneous groups in the future: Not in the paper. "To what extent would you like to work with a teammate whose mother tongue is [ingroup language]?"

4. Feeling thermometer: Feeling thermometer for Jews and Arabs. In the paper, the outcome is called *Prejudice Towards Outgroup* in the table, but the authors confirmed over email that it measures prejudice reduction (i.e. bigger numbers mean more positive feelings towards the outgroup).

**Course score* is pre-registered as secondary in experiment 1.

Treatment. Intervention consisted of randomly assigning students within classrooms to work with a student from the other ethnicity (i.e. Jewish-Palestinian, treatment group) or from their same ethnicity (control group); these are the groups, in some cases they had to form triads to accommodate uneven numbers (page 22). All classes taught the same lesson plan according to the course level, and these plans were designed by the researchers to conform to the four conditions for optimal contact laid out by Allport (page 23).

Allport Conditions. All conditions satisfied; from page 6: "All participating classes taught the same lesson plan, designed by the second and third authors to conform to Allport's conditions for optimal contact".

Duration of contact. Intervention was implemented in two consecutive class sessions in the first two weeks of the semester (page 22). No information on duration of the sessions, but we will assume one hour.

Days Since Contact Ending and Measurement. Survey one-to-two weeks after the second class (i.e. end of the intervention; page 25).

Reported in abstract. "We find some evidence that collaborative learning results in higher achievement among Palestinian students. However, the impact of such learning on the achievements of majority Jewish students tends to be negative. Moreover, the learning experience itself is generally viewed in a negative light, especially among majority Jewish students, who prefer to work with in-group members and feel less comfortable in heterogeneous pairs. Patterns for Palestinian students, though less pronounced, also tend to be negative, with students who worked in heterogeneous groups reporting a reduced sense of belonging and confidence in class. The only exception we find to this largely negative pattern is that, consistent with prior work, both Jewish and Palestinian students who worked in heterogeneous groups are significantly more likely to forge relationships with outgroup members." I consider achievement to encompass outcome (1), and consider it reported in the abstract for experiment 1 even though the results for Jewish students in experiment 1 are positive not negative (notice the wording "tends to"). Underline covers outcome (6) for Palestinian students.

Specification. No details of the specification in the paper, but according to PAP (page 10), they will analyze Jewish and Palestinian students separately, with and without controls, and with class fixed effects. For Jewish students, they will follow LSDV method with weighted dummies. There's no mention of clustering in the PAP or paper, but tables in the paper have N Clusters, and according to the note in Table 1 (page 10), standard errors are clustered at the group level.

Results are presented separately for Jewish and Palestinian students, and I pick the specification without controls following our rules. Results for (6) and (10) are in Table 4 (page 16) and Table 5 (page 19), respectively,

SDs. Received SDs for the control group (for Jewish and Palestinians separately) from the authors.

EXPERIMENT 2:

Outcomes. Pre-registration is prior to researcher access to outcome data, so outcomes are pre-registered before analysis. Found the PAP in the pre-registration files; it clarifies that the intervention began but they had not finished collecting outcome data nor had they analyzed any data. The primary outcomes listed in the PAP are (page 3; compare to the ones in the paper in page 25):

Learning and Achievement:

1. Learning: According to the PAP this would be grade in individual assignment, but this is not in the paper. There's only *Course Score,* but this is registered as a secondary outcome.

2. Contribution to joint work: Drop because it is mechanically affected.

3. Comfort with partner: Drop because it is mechanically affected.

4. Partner allowed my participation: Drop because it is mechanically affected (though not in the paper anyway).

5. Self-efficacy: Grouping of "I can succeed in this course" and "I am confident in my English abilities" (in PAP it also included their estimate of their grade in the course).

Intergroup Relations:

1. Motivation to work with partner again: Drop because it is mechanically affected.

2. Motivation to work in groups: *General attitudes towards group work*. Combines "I enjoy working with people from class" and "I enjoy working in small groups during class." No results for it in the paper.

3. Motivation to work in heterogeneous groups in the future: Not in the paper.

4. Feeling thermometer: In the PAP it is rating for the following groups, Jews, Arabs, Ethiopians, and Ultra Orthodox. In the paper, the outcome is called *Prejudice Towards Outgroup* in the table, and is a feeling thermometer for Arabs and Jews. As above, authors confirmed that a more positive number means more positive feelings.

Treatment. As in experiment 1, students were randomly assigned to work with a student from the other or their same ethnicity. However, the intervention was implemented in four class sessions instead of two, and the lesson design was up to the teachers (page 24). Researchers did meet with the teachers, and explained and asked them to apply Allport's theory for collaborative work.

Allport Conditions. From page 24: "Since the design of the lessons was left to instructors, we wanted to ensure that they knew how to apply Allport's four conditions. Thus, we prepared a short manual explaining Allport's theory and how they could apply each condition during collaborative work".

Duration of contact. The intervention took place in four class sessions that the teachers selected (page 24). No details on duration of class sessions, we will assume 1 hour.

Days Since Contact Ending and Measurement. Survey during the final week of the semester (page 25). Since different classes had the intervention in different class sessions, then this measure varies with class. The PAP states that the intervention happened during the last 6 weeks of the semester, so assuming that on average the intervention took place in the middle of this period (day 21), and the survey was administered on the last day (day 42), days since contact is 21.

Reported in Abstract. "Patterns for Palestinian students, though less pronounced, also tend to be negative, with students who worked in heterogeneous groups reporting a reduced sense of belonging and confidence in class." I consider confidence to refer to self-efficacy, which indeed has negative results for Palestinian students.

Specification. No details of the specification in the paper, here is what we have from the PAP (page 10): They do analysis separately for Jewish and Palestinians, and at the individual

level. They will estimate difference-in-means with and without covariates. Covariates include: gender, age, degree, prior acquaintance with partner. For the Jewish participants, they will also include political ideology and will use the LSDV method, running an OLS regression with weighted dummies (according to the probability of random assignment). There's no mention of clustering in the PAP or paper, but tables in the paper have N Clusters, and according to the note in Table 1 (page 10), standard errors are clustered at the group level.

Results are presented separately for Jewish and Palestinian students, and I pick the specification without controls. Results for *Self-Efficacy* are obtained from column 3 in Table 4 (page 16). Results for *Prejudice Towards Outgroup* are in column 3 of Table 5 (page 19).

SDs. Received SDs for the control group (for Jewish and Palestinians separately) from the authors.

A.3.2.11 Adamu et al. (2024) The Effect of Social Ties on Engagement Cohesion: Evidence from Ethiopian University Students

Pre-registration link.

Outcomes. Pre-registration states that registration is prior to realization of outcomes. Plus, PAP is from November 2022 and the endline was conducted between October and November 2022, so will consider outcomes as pre-registered before analysis. Indices are pre-registered (PAP page 13).

1. Political Engagement Index. Constructed from survey responses on the following topics: I) behavioral measures of messages to government ministries, II) contacting a government official, III) signing a petition and IV) intending to or becoming a member of a political party (hypothetical, but the outcome is still behavioral/incentivized because the others were not).

2. Civic Engagement Index: Includes 8 measures of civic engagement (behavior), only one of which is hypothetical (i.e. "intention to"), so outcome is behavioral/incentivized.

3. Political/Ethnic Tolerance Index: module of survey questions that measure varying aspects of tolerance, including attitudes towards other political parties and ethnic groups (explicit evaluations), supporting political violence against other groups (political and cultural attitudes), and disapproval of political compromise (unrelated).

4. Social Cohesion Index:

    (a) Ranking of Ethiopian identity compared to regional and ethnic identities (political and cultural attitudes).

    (b) Perceptions of diversity as strength (political and cultural attitudes).

    (c) Perceptions of Ethiopian unity (political and cultural attitudes).

    (d) Support for leaders from other ethnic groups (political and cultural attitudes).

    (e) Trust in students from other ethnic groups (explicit evaluations).

(1) and (2) are Unrelated because they measure behaviors not related to intergroup relations.

*There's also results for separate items but will focus on the indices.

Treatment. Treatment is being randomly assigned to attend TEF (Tolerant Engagement Forum) or not. The forum involves presentations and networking, followed by a second session of structured dialogue in groups of approximately 10 people, with a minimum of 3 women and 3 ethnic minorities. Above that minimum, they randomly varied the number of women and minorities in each group to assess the impact of greater diversity on outcomes (page 18). However, I could not find results with this variation. Thus, I had to use TEF (treatment group) vs No TEF (control group) and treatment is bundled.

Allport Conditions. No discussion in the paper, but we can infer support from authority, equal status (i.e. all university students), and given the nature of the discussion and forum overall, we can infer that students were in a collaborative environment with a common goal of increased tolerance.

Duration of contact. One-day TEF workshops (page 4), couldn't find more information on hours. For our best guess will do 8 hours.

Days Since Contact Ending and Measurement. Endline survey approximately 4 months after the TEF workshop (page 21).

Reported in Abstract. "Four months post-intervention, we observed increases in both self-reported and behavioral measures of civic engagement, effects that increase with the formation of new social ties." Covers outcome (2).

Specification. They estimate ITT, including baseline values of the outcome, pre-treatment controls, and block fixed effects interacted with the treatment indicator (page 24; according to email they did demean before interacting with treatment).

Results appear in solid line top estimates in Figures A11-4-7-A14 (pages 26 and 33 and pages 21 and 24 appendix). There's no table format results but authors provided the point estimate and standard error for each outcome (in txt files).

SDs. SDs for the control group obtained from the author.

A.3.2.12 Mousa et al. (2024) Counselling, Intergroup Contact, and Refugee-Native Integration in Lebanon

Pre-registration link.

Outcomes. Pre-registration prior to realization of outcomes. As pre-specified, they use a series of survey items and detect latent clusters, then using factor analysis to keep certain indices. The outcomes in the paper end up being the following, all of which are constructed from pre-specified measures:

1. Social proximity:
    (a) How much do you think you have in common with kids from Lebanon/Syria? (explicit evaluation)
    (b) In general, Lebanese people are friendly toward Syrians/Lebanese. (explicit evaluation)

(c) My Syrian friends would be supportive if I became close friends with someone Lebanese/Syrian. (unrelated)

(d) I can imagine becoming close friends with someone Lebanese/Syrian. (political and cultural attitudes)

(e) My family would be supportive if I became close friends with someone Lebanese/Syrian. (unrelated)

2. Conflict knowledge: Conflict is something that should not happen/a normal part of life. (unrelated)

3. Conflict Skill: Let's say two of your friends got into an argument and they ask for your help resolving it. How comfortable would you feel about stepping in to help? (unrelated)

4. Emotional Skill: When a friend is sad, I usually understand why. (unrelated)

5. Outgroup event RSVP: Accept invitation to outgroup cultural event. (behavioral measure)

6. Outgroup event attendance: Actually attending outgroup cultural event. (behavioral measure, and behavior/incentivized)

For (5) and (6) we use *Includes specific people met* because other program attendants could also be in the event.

Treatment. The Family Psycho-Social Support Program (FPSS) program provides services focusing on improving mental health and well-being (page 7). Sessions are provided in groups of 8-15 students. For the intervention, they randomly assign individual participants to: (1) attend either heterogeneous or homogeneous classrooms in the program; and (2) receive either an empathy curriculum, or a placebo curriculum focused on health and nutrition (page 9).

We will focus on the effects for the youth, since they're the ones that receive the treatment and have any intergroup contact. We will use the comparison between homogeneous classrooms (control group) vs heterogeneous classrooms (treatment group).

Allport Conditions. No discussion in paper. We can infer cooperation since participants are taught to cooperate, and the common goal of course activities: "Here, we propose that explicitly training participants how to empathize and cooperate — with curricula tailored to the local setting — is better suited to activate the effects of contact without such empathy-related content" (page 5). We can also infer equal status within the intervention and support from authority.

Duration of contact. The FPSS program sessions met once a week for 12 weeks (page 10), and each session was 2-3 hours long (information obtained from authors). Our best guess is 12*2.5 hours.

Days Since Contact Ending and Measurement. Endline survey (measuring attitudes) administered 2 weeks after the program ends, and behavioral measures obtained 3 weeks after the program ends (page 9). For RSVP, "Invitations are sent immediately after the program ends" (page 13), so 0 days.

Reported in Abstract. "We find that contact significantly reduces prejudicial attitudes toward the out-group but find that it also depresses participation in future contact — such as attending events celebrating the outgroup's culture 1-2 months after treatment". Social proximity and outgroup event attendance considered reported in the abstract.

Specification. From page 13: "We estimate average treatment effects by regressing the outcomes on the single and combined treatment indicators, controlling for randomization block fixed effects, program cycle, age, gender, nationality, education, a dummy for whether the respondent is working, and the outcome question measured at baseline wherever available, to increase precision. Instead of the standard OLS estimator, as specified in the pre-analysis plan, we employ the Lin estimator which is much more robust against unequal assignment probabilities across groups, which were evident especially for nationalities, due to unbalanced registration numbers (Lin 2013). We cluster standard errors at the classroom level for every model including curriculum treatment, as this treatment is administered at the group level. For the models that only include contact treatment, we apply heteroskedasticity-consistent standard errors on the individual level (HC2; Long and Ervin 2000), a deviation from the pre-analysis plan that is methodologically sound and that only results in slight differences in standard error magnitude for contact models."

Results for the effect of contact treatment are in Figure 1 (page 18), and in Table A1.

SDs. Standard deviations are in Table A5 (page A31). We don't have the control group SD for our relevant comparison because it would combine the Empathy and Control groups, but we do have it for the full sample in All.