Integrating diversity education programs in school curricula can reduce prejudice: evidence from field experiments in Israel

Chagai M. Weiss[a,1], **Shira Ran**[b], **and Eran Halperin**[c,b]

[a]Stanford King Center on Global Development; [b]aChord Center, Hebrew University of Jerusalem; [c]Department of Psychology, Hebrew University of Jerusalem

This manuscript was compiled on December 29, 2022

Intergroup prejudice is pervasive in many contexts worldwide, leading to discrimination and conflict. Existing research suggests that prejudice is acquired at an early age and that durably improving intergroup relations is extremely challenging, often requiring intense interventions. Building on existing theoretical frameworks in social psychology and inspired by the Israeli TV series "You Can't Ask That," which depicts charismatic children from minority groups discussing sensitive topics at the core of intergroup relations, we develop a monthlong diversity education program. Our program exposed students to the TV series and facilitated follow-up classroom discussions in which students constructively addressed various sensitive topics at the core of intergroup relations. Through two field experiments implemented in Israeli schools, we show that integrating our intervention into school curricula improved Jewish students' attitudes towards minorities and increased take-up of a pro-diversity bracelet up to 13 weeks post-treatment. We further provide suggestive evidence that the intervention was effective by encouraging students to take their outgroups' perspectives and address an element of scalability by delegating implementation responsibilities to classroom teachers in our second study. Our findings suggest that psychologically informed intensive interventions that introduce changes in school curricula are a promising route to reducing prejudice at a young age.

Prejudice | Intergroup Relations | Conflict | Field Experiments |

Prejudice, conceptualized as "a negative bias towards a social category of people," is a common feature of intergroup relations around the world (1). Existing studies suggest that intergroup prejudice is acquired at an early age and is highly resistant to change (2, 3). The prevalence and stability of prejudice worldwide are concerning because negative sentiments towards outgroups have been linked with a host of adverse phenomena, including discrimination (4, 5), officers' use of force (6), and intergroup conflict (7, 8).

Acknowledging the detrimental consequences of intergroup prejudice, scholars and practitioners have long been interested in developing approaches to reduce prejudice and improve intergroup relations (1). Despite a common understanding that prejudice is a consequence of large-scale social forces such as intergroup competition (1), exclusionary institutions (9), and intense socialization experiences (2, 3), most existing approaches for prejudice reduction focus on light-touch interventions (i.e., short and inexpensive treatments), tested in controlled laboratory environments rather than field-based settings, focusing on short-term effects and lacking rigorous evaluations of long-term durability. Recent meta-analyses suggest that these approaches have modest effects (10, 11), and though a growing number of intensive field-based studies report encouraging findings regarding the potential for reducing prejudice and discrimination (12–18), several landmark studies testing canonical theories of prejudice reduction have yielded modest or null effects (19–21).

Building on the understanding that intergroup prejudice is, at least in part, a consequence of early-age socialization in education systems (2, 3, 22), and inspired by recent calls to combine psychological insights with structural intervention for prejudice reduction (11), we develop an intensive diversity education program to familiarize Jewish Israeli elementary school students with different social groups in Israeli society and reduce their prejudice towards minorities. Specifically, we designed a month-long diversity education program based on the Israeli TV series "You Can't Ask That."[*] Each episode of the TV series "You Can't Ask That" depicts charismatic children from different social groups responding with sophistication and humor to questions from home audiences regarding core issues of intergroup relations that children would never consider asking outgroups directly (see Figure 1).[†]

Our educational program focused on three episodes depicting Arab, visually impaired, and immigrant children,[‡] and

[*]The TV series is a Hebrew adaptation of an Australian show, which has been translated and aired in multiple countries. See the Israeli TV series website here: https://testkankids.kan.org.il/program/?catid=1527.

[†]Show participants voluntarily applied to participate in the show, and were selected by the producers to participate in a given episode.

[‡]Immigrants were children of Filipino foreign workers, many of whom are undocumented immigrants in Israel.

Significance Statement

Existing research suggests that intergroup prejudice is acquired at an early age and that durably reducing prejudice is extremely challenging. We develop a monthlong diversity education program, which exposed children to charismatic outgroups through the Israeli TV series "You Can't Ask That," and facilitated constructive classroom discussions regarding sensitive topics at the core of intergroup relations. Through two field experiments implemented in Israeli schools, we show that integrating our intervention into school curricula reduced Jewish students' prejudice towards minorities and increased take-up of a pro-diversity bracelet up to 13 weeks post-treatment. Our findings emphasize how psychologically informed, intensive interventions that introduce changes in school curricula can reduce prejudice at a young age.

C.M.W., S.R, and E.H. designed research; performed research and project administration; analyzed data; and wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: cmweiss@stanford.edu.

included four sessions in which Jewish elementary school students watched the TV series and engaged in guided follow-up classroom discussions aimed to engage with the show's content and teach students about intergroup similarities, intragroup heterogeneity, and the value of taking others' perspectives. Taking into account recent critiques of the prejudice reduction literature, which suggest that many interventions overlook inequality and group grievances in the persuit of intergroup harmony (23, 24), the core objective of our diversity education program was to expose students to charismatic outgroups, constructively discuss various sensitive issues at the core of intergroup tensions, and in doing so, improve intergroup attitudes and behaviors. By combining psychologically informed intragroup conversations with parasocial (i.e., media-based) exposure to outgroups, we ensure that the discussion of conflictual topics minimally burdens disadvantaged group members and still reduces prejudice towards minorities.

To test the effects of our diversity education program, we implemented two field experiments in Israel. Experimentally integrating our intervention into school curricula and measuring elementary school students' attitudinal and behavioral prejudice up to thirteen weeks post-treatment, we consider the short and longer-term effects of our diversity education program. Doing so, we follow recent calls to rigorously evaluate the long-term attitudinal and behavioral consequences of structural intensive prejudice reduction interventions (10, 11). Our findings suggest that early childhood education that facilitates positive exposure to outgroups and constructive intragroup discussions of topics at the core of intergroup relations can durably reduce prejudice towards multiple minority groups.

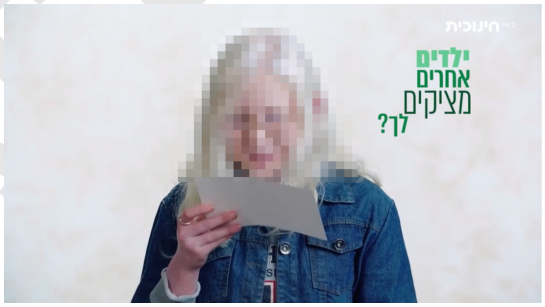
The Prejudice Reduction Intervention

Our diversity education program leveraged the TV series "You Can't Ask That" to expose Jewish Israeli students to charismatic minority group members and broach constructive classroom discussions about sensitive topics at the core of intergroup relations. The show's main objective is to provide home audiences with an opportunity to ask members of different social groups forthright questions regarding taboo topics and to generate a constructive discussion (in the studio) about intergroup grievances, disagreements, inequality, and experiences of discrimination. Notably, all show participants chose to participate in these conversations and underwent a selective application process. Thus, the show provides a platform to address some of the most sensitive and contentious issues relating to intergroup relations without imposing undesired and challenging conversations on unwilling minority group members. Moreover, by virtue of the platform that does not entail direct intergroup contact, minority group members can freely make their arguments with no interruptions or judgment from majority group members.

Recent studies suggest that addressing sensitive and conflictual topics during prejudice reduction interventions is important for meaningful social change (12, 23, 24), but doing so might generate backlash (25). Therefore, we designed psychologically informed follow-up classroom activities to help children unpack the show's content and ensure that our diversity education program would effectively reduce prejudice. The main goal of our classroom activities was to connect the sensitive topics discussed in the TV series with common mechanisms of prejudice reduction emphasizing intragroup heterogeneity (i.e., differences within social groups), intergroup similarities (i.e., similarities between social groups), and the value of taking an outgroup members' perspective (i.e., putting oneself in another person's shoes).

(a) Arab Children

(b) Immigrant Children



(c) Visually Impaired Children

Fig. 1. This figure depicts snapshots from the TV series You Can't Ask That. Panel (a) portrays an Arab child discussing their complex national identity, panel (b) depicts an immigrant child discussing their fear of being deported, and panel (c) depicts a visually impaired child discussing their experiences with bullying.

For example, we designed activities encouraging discussions of group grievances that emphasize outgroup heterogeneity. Specifically, our activities encouraged students to reflect on how different show participants hold varying and, at times, contradicting positions on any given social or political issue. The realization that outgroups vary with regard to their grievances and political preferences can emphasize that a given outgroup is not homogenous, and this, in turn, can lead to prejudice reduction (26).

Similarly, our program included constructive activities focusing on intergroup disagreements that emphasize elements of intergroup similarities. For example, when unpacking challenging questions relating to intergroup conflict, our follow-up classroom discussions emphasized how despite many differences, ingroups and outgroups often share similar motivations,

emotions, and feelings around conflictual issues. Realizations regarding intergroup similarities originating from such activities can improve majority group members' attitudes toward minorities (27–29).

Finally, the activities we designed encouraged students to discuss various experiences of discrimination and inequality described in the show. In doing so, our program emphasized the value of taking other people's perspectives. By encouraging students to put themselves in others' shoes (30), we sought to raise awareness of inequality and discrimination, and increase students' capacity for intergroup empathy, in order to reduce their prejudice towards minorities (14, 31).

Our intervention included four meetings. In the first three meetings, students watched a group-specific episode (15 minutes) featuring Arab, visually impaired, or immigrant children and engaged in follow-up classroom activities and discussions (30 minutes). In the final meeting, students watched a recap from all three episodes (15 minutes) and engaged in an overview discussion (30 minutes). As noted above, we designed all classroom discussions to focus on the TV series' central theme—discussions of topics at the heart of intergroup relations—and to connect these topics with the core psychological mechanisms noted above, relating to information about intragroup heterogeneity, intergroup similarities, and the value of taking other people's perspective. An elaborate description of the TV series is provided in Appendix A1, and an overview of our educational program is provided in Appendix A2.

Testing the Intervention

To test the effects of our intervention, we implemented two field experiments in Israeli elementary schools. As depicted in Figure 2, in both experiments, we collected baseline survey data from students, then block-randomized classes into treatment and control conditions by grade, and collected endline surveys and behavioral measures.

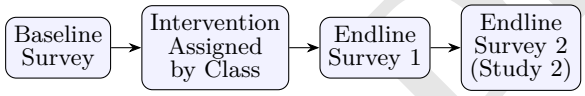


Fig. 2. Study Procedure.

Study 1.

Research Design. In our first study, which served as a preliminary test of our intervention, we implemented a field experiment with 12 classes in grades 4-6, in a school located in central Israel. After receiving IRB approval from the Hebrew University of Jerusalem, permission from Israel's ministry of education, and informed consent from students' parents, our field experiment followed the steps depicted in Figure 2.[§] Following a baseline survey with 270 students, we block randomized classes into treatment and control conditions by grade, resulting in 6 treated and 6 untreated classes. Students in treated classes participated in our month-long educational curriculum, which a professional educational practitioner delivered, and the control group did not participate in any activity.

[§] In line with our IRB approval, and in agreement with the ministry of education, all students in treated classes participated in our intervention, however only students for whom we received parental informed consent participated in our surveys. We obtained informed consent from over 70% of students. Given our design, parental consent is orthogonal to treatment and does not threaten internal validity.

Notably, the start of our intervention coincided with a cycle of intense violence between Jews and Palestinians. Between May 10-21, 2021, intense missile fires and inter-communal clashes disrupted life in many cities across Israel, including our intervention site. Violence was so intense that some schools closed for several days, but during the study period, our partner school operated in regular capacity, and we concluded implementing our intervention amongst treated classes in the first week of June 2021.

A week post-treatment, we began collecting endline surveys. The main outcomes we measured in our pre-and post-treatment surveys included attitudes and behaviors relating to intergroup prejudice and support for diversity. Specifically, we collected information about students' outgroup affect towards Arab, immigrant, visually impaired, and Ultra-Orthodox children (the latter group was not mentioned in the intervention), contact intentions with Arab, immigrant, visually impaired, and Ultra-Orthodox children, perceptions of intergroup similarity with Arab, immigrant, visually impaired, and Ultra-Orthodox children, a five-item index of students' support for diversity, and a behavioral measure of registration for a future intergroup contact event. In our main analyses, we aggregate measures of group-specific affect, contact intentions, and similarity into general outgroup indices and report average treatment effects on group-specific measures in Appendix A4.2.2. We describe the survey wording we used to collect our main outcome measures in Appendix A3.

Estimation Strategy. We estimate OLS regressions in which we regress standardized outcomes ($\mu = 0$, $\sigma^2 = 1$) over our treatment indicator, controlling for respondents' gender, assignment block, and pre-treatment outcome measures.[¶] Given the modest number of clusters in our data, we employ a wild-cluster bootstrap procedure to cluster our errors at the classroom level (32). Our main estimating equation is:

$$y_{ic} = \beta Z_c + \phi \boldsymbol{X}_{ic} + \epsilon_{ic} \quad [1]$$

In our analyses, we focus on identifying β, representing the average treatment effect of the intervention on students' post-treatment attitudes and behaviors.

Results. In Figure 3, we report the effects of the intervention on primary outcomes of interest. The results in Figure 3 suggest that our intervention substantially affected students' attitudes. Indeed, students' positive affect towards outgroups increased by over a third of a standard deviation, resembling an eight points shift on a 0-100 outgroup feeling thermometer. This effect is almost double the magnitude of the average effect of well-powered interventions reported in a recent meta-analysis of prejudice reduction experiments (11).

We also find that the intervention increased students' intentions to engage in contact with outgroups and support for diversity by over a fifth of a standard deviation and hieghtened students' perceptions of similarity with outgroups by over a third of a standard deviation. Finally, though the point estimate on our behavioral measure of registration for an intergroup contact event is positive, it is imprecisely estimated. Therefore, we conclude that the intervention substantially affected students' attitudes but did not shape the behavior we measured in this first study.

[¶] For our behavioral measure, which was not collected pre-treatment, we control for pre-treatment thermometer, diversity, similarity, and contact intention indices.

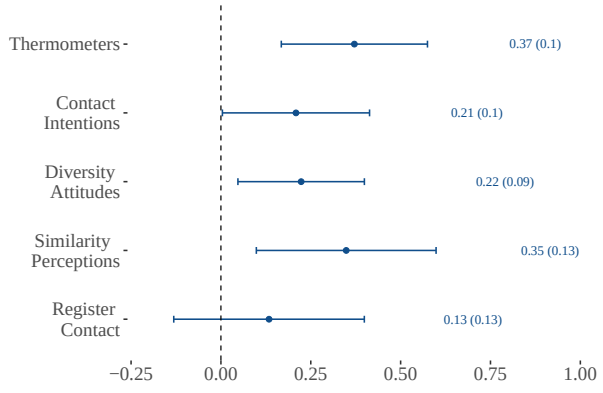


Fig. 3. Exposure to the intervention improved children's attitudes in Study 1. This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on students' attitudes and behaviors 1-2 weeks post-treatment. Point estimates and standard errors (in parentheses) are reported along each estimate.

We subject our results to several diagnostic and robustness checks. In the Appendix, we report balance checks (Appendix A4.1), examinations of attrition that confirm that treatment status does not predict non-response to post-treatment surveys (Appendix A4.2.1), estimations of alternative models with disaggregated outcomes (Appendix A4.2.2), estimations of alternative specifications employing randomization inference (Appendix A4.2.2.), and explorations of effect moderation based on pre-treatment prejudice (Appendix A4.2.2). Finally, since the implementation of our intervention coincided with a cycle of Jewish-Palestinian violence, we pay close attention to outcomes relating to Arab outgroups. In Appendix A4.3 we demonstrate that whereas attitudes towards Arabs improved between baseline and endline amongst treated subjects, similar attitudes were impaired amongst students in the control group. We cautiously attribute the negative trend amongst students in the control group to the cycle of violence that coincided with our intervention and suggest that educational programs can be employed in times of intense intergroup conflict to counteract the deterioration of intergroup attitudes and behaviors and promote more favorable intergroup relations.

Study 2.

Research Design. Our second study is similar to Study 1, with several notable improvements relating to sample size, treatment modality, and outcome measurement. In terms of sample size, we focus on 767 Israeli students (grades 4-6) in five schools located in central Israel. After surveying all consenting students,[‖] we block randomized a subset of classes (29/46) into treatment. Block randomization was implemented by grade within each school.

In terms of treatment modality, in Study 2, treatment was delivered organically by school teachers to assess one important dimension of scalability relating to treatment implementation (33). Thus, teachers from treated classes were provided with all the necessary materials to implement the intervention and were instructed to deliver four sessions of the intervention over four weeks. The materials provided to teachers included a short document describing the theoretical rationale of the intervention, classroom slides, and a general guide describing the activities to be implemented in each class. A majority of teachers also participated in a one-hour Zoom information session in which an educational practitioner described the intervention and answered any questions raised by teachers.

Finally, in terms of outcome measurement, in Study 2, we focus on short and long-term effects. A week after treated classes in a given school completed all four sessions of the intervention, we returned to the school in order to administer our first post-treatment survey. The overwhelming majority of treated respondents participated in the first post-treatment survey 1-2 weeks following the intervention, and a very small minority of students were surveyed up to 6 weeks post-treatment due to technical challenges in sampling students. Eight weeks after all treated classes in a given school completed the four sessions of the intervention, we returned to administer our second post-treatment survey. In practice, students in all but one school responded to this survey 8-13 weeks after exposure to the intervention. Given scheduling challenges prior to the summer break, one of our schools did not participate in the second post-treatment survey. We show in Appendix A5.2.1 that this attrition is orthogonal to treatment and does not pose a threat to internal validity.

In addition to the primary outcomes we collected in Study 1, in Study 2, we collected survey measures eliciting students' beliefs about outgroupgroup heterogeneity, and appreciation for taking the perspective of outgroup children. Together with our measure of intergroup similarities, these additional measures allow us to explore whether our treatment shaped the core psychological mechanisms underlying classroom activities. Finally, in our second wave of Study 2, we collected a behavioral measure of support for diversity. As compensation for participation in our surveys, we provided children with the possibility to select one of two gifts: A bracelet with a pro-diversity statement or a bracelet with a personal reassurance statement. Inspired by previous studies, we measure whether each student selected the pro-diversity bracelet and interpret this selection as an act of signaling support for diversity (34). Like our analyses in Study 1, our main analyses in Study 2 focus on aggregate indices rather than group-specific survey measures. We elaborate on our survey methodology, the procedures we used to collect our main outcomes, and the timing of outcome collection in Section A3 of the Appendix.

Estimation Strategy. We estimate an OLS regression depicted in equation 2. Following our pre-analysis plan, we interact mean-centered covariates with our main treatment indicator to increase precision (35), cluster errors by class, and employ weights that account for varying treatment assignment probabilities across blocks (36). Covariates include gender, assignment block, and pre-treatment outcome measures,[**] and all outcomes are standardized ($\mu = 0$, $\sigma^2 = 1$). Our estimating equation is:

$$y_{ics} = \beta Z_{cs} + \phi \boldsymbol{X}_{ics} + \gamma(Z_{cs} * \boldsymbol{X}_{ics}) + \epsilon_{ics} \quad [2]$$

In our analyses, we focus on β, representing the average

[‖] Like in Study 1, all students in treated classes participated in our intervention. However, only students for whom we received parental informed consent participated in our surveys. We obtained parental informed consent from 69% of students. Given our design, parental consent is orthogonal to treatment and does not threaten internal validity.

[**] For our two behavioral measures which were not measured pre-treatment, we adjust our model with pre-treatment thermometer, diversity, and contact intention indices.

treatment effect of the intervention on students' post-treatment attitudes and behaviors.

Results. In Figure 4, we report the effects of our intervention on primary outcomes 1-2 (8-13) weeks post-treatment. Our estimates suggest that the treatment substantially affected students' intergroup attitudes and increased the take-up of a pro-diversity wristband but did not affect students' registration for an intergroup contact event. Despite delegating the responsibility of treatment implementation to teachers, the magnitude of effects reported in Figure 4 remains substantively large.

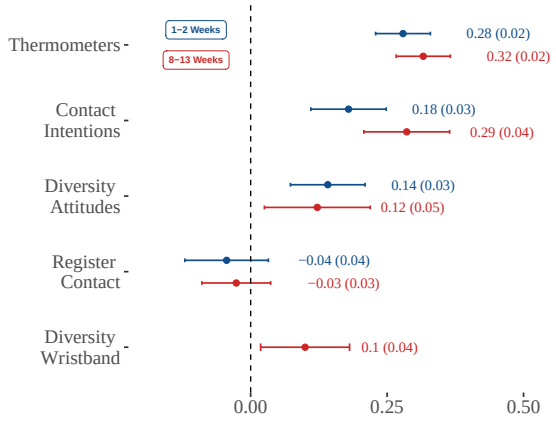


Fig. 4. Exposure to the intervention in Study 2 improved children's attitudes and increased the take-up of a pro-diversity wristband up to 13 weeks post-treatment. This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on students' attitudes and behaviors 1-2 (8-13) weeks post treatments in blue (red). Point estimates and standard errors (in parentheses) are reported along each estimate.

Indeed, exposure to our month-long intervention increased students' positive intergroup affect toward outgroups by almost a third of a standard deviation in the short and longer term. Though treated students' registration for an intergroup contact event was not affected by treatment, their self-reported intentions to engage in intergroup contact increased by a fifth of a standard deviation 1-2 weeks post-treatment and by almost a third of a standard deviation 8-13 weeks post-treatment. As we further discuss in Section A3.4 of the appendix, the null effect on actual registration for an intergroup contact event might be explained by a ceiling effect. Indeed, even among students that were not exposed to the intervention, registration for a contact event as part of our post-treatment survey was substantially high. Finally, we find that the treatment increased students' appreciation for diversity by over a tenth of a standard deviation and that this appreciation translated into students' behaviors. That is, treated students' were more likely to select a pro-diversity wristband that signals to their peers that "*In our school everyone belongs*," over a personal reassurance wristband as compensation for participation in the survey.

What psychological mechanisms might account for the effectiveness of our intervention and explain the success of providing students with meaningful exposure to charismatic outgroups followed by intragroup conversations and activities that constructively engage with sensitive topics at the core of intergroup relations? To answer this question, we turn to survey items measuring the three psychological mechanisms emphasized in our program: intragroup heterogeneity, intergroup similarity, and perspective-taking. In figure 5, we provide a suggestive test of mechanisms by examining how our treatment influenced students' beliefs that different outgroups are heterogeneous, that students are similar to various outgroups, and that it is important to try and take the outgroup's perspective.

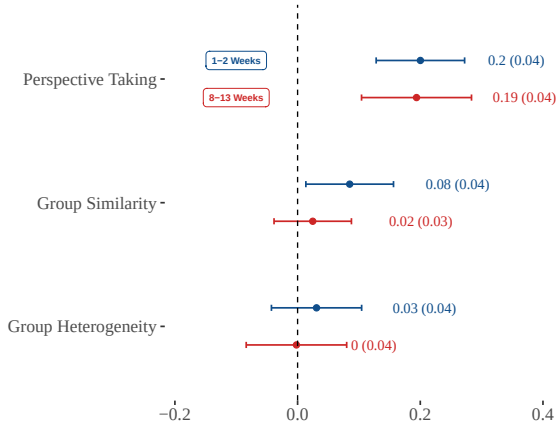


Fig. 5. Exposure to the intervention in Study 2 increased students' ability and willingness to take outgroup perspectives, had a limited impact on perceptions of intergroup similarities, and had no effect on perceptions of intragroup heterogeneity. This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on measures of mechanisms 1-2 (8-13) weeks post treatments in blue (red). Point estimates and standard errors (in parentheses) are reported along each estimate.

The results in Figure 5 suggest that our intervention had consistent and large effects on students' willingness to take their outgroups' perspective, smaller and inconsistent effects on students' perceptions of intergroup similarity, and no noticeable impact on perceptions of intragroup heterogeneity. We cautiously interpret these additional results to suggest that our intervention, which exposed students' to charismatic outgroup children and facilitated constructive classroom discussions regarding a variety of sensitive topics, was effective because it helped students realize the value of taking the outgroup's perspective. In that sense, we construe our overall evidence to suggest that rather than generating backlash amongst students, constructive classroom engagement with sensitive topics at the core of intergroup relations, combined with meaningful mediated outgroup exposure, facilitated increased appreciation for understanding other groups' circumstances and, in turn, improved intergroup attitudes and pro-diversity behavior.

Like in Study 1, we subject our results to similar diagnostic and robustness checks in Appendices A5.1-A5.2. Moreover, since one school did not participate in our second endline survey, one might worry that the short and long-term effects in Figures 4-5 are not easily comparable. To address this issue, we show in Appendix A5.2 that a similar pattern of results emerges when focusing only on respondents participating in both surveys. Finally, in Appendix A5.2, we demonstrate that our results are consistent when estimating the empirical specification employed in Study 1.

Discussion

We developed a diversity education program that leveraged the Israeli TV show "You Can't Ask That" to expose Jewish Israeli students to charismatic minority group members and facilitate constructive classroom discussions about sensitive topics at the core of intergroup relations. Through multiple field experiments in Israel, in which we experimentally integrated our program into school curricula, we show that our intervention had substantial immediate and longer-term effects on Jewish students' attitudes and some pro-diversity behavior. Our findings contribute to several theoretical and applied questions.

First, we contribute to the literature on prejudice reduction by developing a theoretically informed intensive intervention. Building on the understanding that prejudice, as well as other social and political attitudes, are often acquired during early-age socialization (2, 3, 22), we developed a diversity education program for elementary school children. Our program combined ongoing mediated exposure to charismatic outgroups alongside constructive engagement with sensitive topics at the core of intergroup relations aimed to increase children's appreciation of diversity and reduce their prejudice towards multiple outgroups. In developing our intervention, we depart from ongoing trends in the prejudice reduction literature focusing on nudge-like interventions (11), and join other recent studies focusing on early childhood education as a valuable platform to improve intergroup relations (17). Our findings emphasize that psychologically informed intensive education programs are a promising route to reducing prejudice at a young age.

Second, we join a growing body of research that employs natural and field experiments to gain insight into prejudice reduction and conflict resolution (12, 13, 15, 16, 19, 20, 38). Through multiple studies, we test our intervention in a naturalistic setting, measuring attitudes and corresponding behaviors amongst our population of interest up to thirteen weeks post-treatment. Moreover, by implementing multiple studies and introducing a central design modification in which we delegate the responsibility of treatment implementation to teachers, we address one of several central concerns regarding intervention scalability (33). Indeed, we show that our intervention was effective even when implemented by teachers themselves (rather than a trained practitioner), who vary with regard to their commitment to reducing prejudice and adhering to the standard protocol of our diversity education program. We encourage future research to build on these promising results and examine other elements of scalability, including spillover effects, and the potential attenuating effects of counter-political reactions to diversity education programs (33).

Despite these contributions, our findings are not without limitations. First, like many field experiments (19, 20, 39), the geographical scope of our research is somewhat limited. However, we emphasize that generalizability is rarely established through a single study, as it often entails cumulative efforts as part of a broad research program (40). In Appendix A5.3, we quantify the sensitivity of our analyses to external validity bias (41), and find encouraging evidence regarding treatment effect homogeneity and the likelihood that our evidence would generalize to substantially diverse target populations.

Second, our intervention is a bundle of multiple components, including exposure to charismatic outgroups and constructive discussions of sensitive topics. Though we elaborate on the theoretical framework underlying our intervention and explore the mechanisms through which we expect our intervention to work, our empirical focus is on evaluating the overall effect of the intervention rather than identifying the relative importance of each particular component. This general empirical focus is motivated by the understanding that effective prejudice reduction interventions may very well require the "mixing of ingredients from multiple theoretical perspectives" (11, p. 555). Like recent studies, (12, 17), our intervention is likely effective due to its multiple components complementing each other. Thus, we encourage future research to build on our work and further clarify the role of different mechanisms in generating the effects of our intervention. However, the key takeaway of our study is that psychologically informed intensive education programs that expose students to charismatic outgroups and constructively discuss sensitive topics at the core of intergroup relations can reduce prejudice in divided societies.

ACKNOWLEDGMENTS. These studies were pre-registered on OSF and As.predicted (study 1: https://osf.io/kdt8y, study 2: https://aspredicted.org/37w7m.pdf), and were approved by the IRB office at the Hebrew University of Jerusalem, as well as by Israel's Ministry of Education. aChord center and specifically Ronit Hanzis, Ido Oren, and Shir Tankel provided excellent support in designing and implementing the intervention. We thank Idit Mistriel, Ryan Enos, Josh Kertzer, Nicholas Sambanis, Alex Scacco, Macartan Humphreys, Nahomi Ichinao, Anna Wilke, Jonathan Homola, Thomas Zeitzoff, Melissa Sands and workshop participants at American University, Harvard MEI, Harvard WoGPoP, Harvard Department of Psychology, the University of Pennsylvania's Conflict and Identity Lab, WZB, POLMETH XXXIX, and the American Political Sceince 2022 Annual Meeting for helpful comments and suggestions.

1. EL Paluck, DP Green, Prejudice reduction: What works? a review and assessment of research and practice. *Annu. review psychology* **60**, 339–367 (2009).
2. DO Sears, CL Funk, Evidence of the long-term persistence of adults' political predispositions. *The J. Polit.* **61**, 1–28 (1999).
3. DO Sears, S Levy, *Childhood and adult political development.* (Oxford University Press), (2003).
4. RD Enos, N Gidron, Exclusion and cooperation in diverse societies: Experimental evidence from israel. *Am. Polit. Sci. Rev.* **112**, 742–757 (2018).
5. K Peyton, GA Huber, Racial resentment, prejudice, and discrimination. *The J. Polit.* **83**, 000–000 (2021).
6. JK Swencionis, ER Pouget, PA Goff, Supporting social hierarchy is associated with white police officers' use of force. *Proc. Natl. Acad. Sci.* **118**, e2007693118 (2021).
7. D Bar-Tal, Societal beliefs in times of intractable conflict: The israeli case. *Int. J. Confl. Manag.* (1998).
8. N Kteily, E Bruneau, A Waytz, S Cotterill, The ascent of man: Theoretical and empirical evidence for blatant dehumanization. *J. personality social psychology* **109**, 901 (2015).
9. A Acharya, M Blackwell, M Sen, The political legacy of american slavery. *The J. Polit.* **78**, 621–641 (2016).
10. EL Paluck, SA Green, DP Green, The contact hypothesis re-evaluated. *Behav. Public Policy* **3**, 129–158 (2019).
11. EL Paluck, R Porat, CS Clark, DP Green, Prejudice reduction: Progress and challenges. *Annu. review psychology* **72**, 533–560 (2021).
12. D Broockman, J Kalla, Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* **352**, 220–224 (2016).
13. DD Choi, M Poertner, N Sambanis, Parochialism, social norms, and discrimination against immigrants. *Proc. Natl. Acad. Sci.* **116**, 16274–16279 (2019).
14. J Kalla, D Broockman, Which narrative strategies durably reduce prejudice? evidence from field and survey experiments supporting the efficacy of perspective-getting. *Am. J. Polit. Sci.* (2021).
15. DD Choi, M Poertner, N Sambanis, The hijab penalty: Feminist backlash to muslim immigrants. *Am. J. Polit. Sci.* (2021).
16. CM Weiss, Diversity in health care institutions reduces israeli patients' prejudice toward arabs. *Proc. Natl. Acad. Sci.* **118** (2021).
17. S Alan, C Baysan, M Gumren, E Kubilay, Building social cohesion in ethnically mixed schools: An intervention on perspective taking. *The Q. J. Econ.* **136**, 2147–2194 (2021).
18. Y Hasson, E Amir, D Sobol-Sarag, M Tamir, E Halperin, Using performance art to promote intergroup prosociality by cultivating the belief that empathy is unlimited. *Nat. Commun.* **13**, 1–15 (2022).
19. A Scacco, SS Warren, Can social contact reduce prejudice and discrimination? evidence from a field experiment in nigeria. *Am. Polit. Sci. Rev.* pp. 1–24 (2018).
20. S Mousa, Building social cohesion between christians and muslims through soccer in post-isis iraq. *Science* **369**, 866–870 (2020).

21. YY Zhou, J Lyall, Prolonged contact does not improve locals' relations with migrants in wartime settings. *Available at SSRN 3679746* (2021).
22. D Bar-Tal, Development of social categories and stereotypes in early childhood: The case of "the arab" concept formation, stereotype and attitudes by jewish children in israel. *Int. journal intercultural relations* **20**, 341–370 (1996).
23. NS Kteily, KJ McClanahan, Incorporating insights about intergroup power and dominance to help increase harmony and equality between groups in conflict. *Curr. opinion psychology* **33**, 80–85 (2020).
24. T Saguy, N Tausch, JF Dovidio, F Pratto, The irony of harmony: Intergroup contact can produce false expectations for equality. *Psychol. Sci.* **20**, 114–121 (2009).
25. EL Paluck, Is it better not to talk? group polarization, extended contact, and perspective taking in eastern democratic republic of congo. *Pers. Soc. Psychol. Bull.* **36**, 1170–1185 (2010).
26. W Hsieh, N Faulkner, R Wickes, What reduces prejudice in the real world? a meta-analysis of prejudice reduction field experiments. *Br. J. Soc. Psychol.* (2021).
27. JM Falomir-Pichastor, C Martínez, C Paterna, Gender-role's attitude, perceived similarity, and sexual prejudice against gay men. *The Span. J. Psychol.* **13**, 841–848 (2010).
28. MJ Brandt, Predicting ideological prejudice. *Psychol. Sci.* **28**, 713–722 (2017).
29. Z Liberman, AL Woodward, KD Kinzler, The origins of social categorization. *Trends cognitive sciences* **21**, 556–568 (2017).
30. CL Adida, A Lo, MR Platas, Perspective taking can promote short-term inclusionary behavior toward syrian refugees. *Proc. Natl. Acad. Sci.* **115**, 9521–9526 (2018).
31. EG Bruneau, R Saxe, The power of being heard: The benefits of 'perspective-giving'in the context of intergroup conflict. *J. experimental social psychology* **48**, 855–866 (2012).
32. AC Cameron, JB Gelbach, DL Miller, Bootstrap-based improvements for inference with clustered errors. *The review economics statistics* **90**, 414–427 (2008).
33. A Banerjee, et al., From proof of concept to scalable policies: Challenges and solutions, with an application. *J. Econ. Perspectives* **31**, 73–102 (2017).
34. A Cheema, S Khan, A Liaqat, SK Mohmand, Canvassing the gatekeepers: A field experiment to increase women voters' turnout in pakistan. *Am. Polit. Sci. Rev.* pp. 1–21 (2021).
35. W Lin, Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals Appl. Stat.* **7**, 295–318 (2013).
36. AS Gerber, DP Green, *Field experiments: Design, analysis, and interpretation.* (WW Norton), (2012).
37. D Bar-Tal, Y Rosen, Peace education in societies involved in intractable conflicts: Direct and indirect models. *Rev. Educ. Res.* **79**, 557–575 (2009).
38. M Lowe, Types of contact: A field experiment on collaborative and adversarial caste integration. *Am. Econ. Rev.* **111**, 1807–44 (2021).
39. B Hameiri, R Porat, D Bar-Tal, E Halperin, Moderating attitudes in times of violence through paradoxical thinking intervention. *Proc. Natl. Acad. Sci.* **113**, 12105–12110 (2016).
40. C Samii, Causal empiricism in quantitative research. *The J. Polit.* **78**, 941–955 (2016).
41. M Devaux, N Egami, Quantifying robustness to external validity bias. *Work. Pap.* (2022).

Supporting Information for Online Appendix: Integrating diversity education programs in school curricula can reduce prejudice: evidence from field experiments in Israel

Chagai Weiss* Shira Ran † Eran Halperin ‡

December 29, 2022

Contents

*Conflict and Polarization Lab, Stanford University. Email: cmweiss@stanford.edu.
†aChord Center, Hebrew University of Jerusalem.
‡Department of Psychology and aChord Center, Hebrew University of Jerusalem.

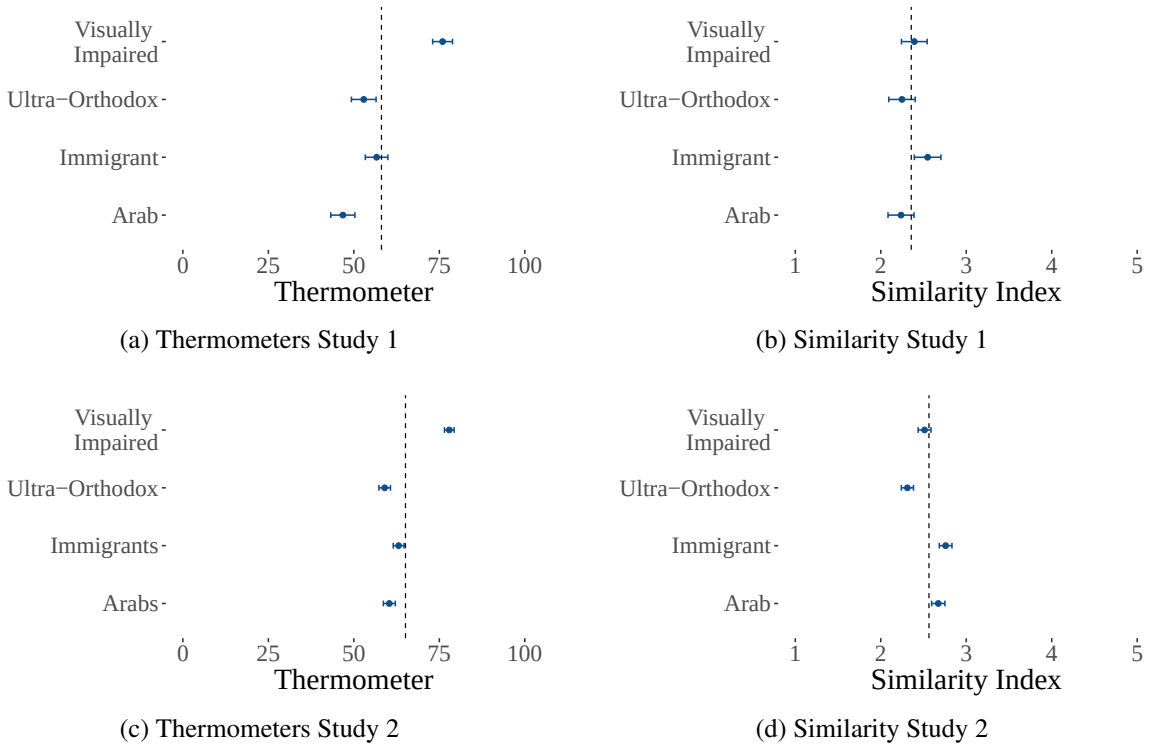
A1 You Can't Ask That

Our intervention was inspired by an Israeli TV series named *Slicha al Ha-Shela, Girsat Ha-Yeladim* which directly translates from Hebrew as "Excuse me for the Question, Kids Version." This TV show was adapted from an Australian TV show called "You Can't Ask That," and was produced by "Kan," Israel's national TV network. All Hebrew version episodes are posted online and can be accessed via the following link: https://testkankids.kan.org.il/program/?catid=1527. Before implementation, we consulted with the producers about using their show, and they expressed enthusiasm about our intervention and noted that there are no copyright issues with using the show since it is publicly available online.

At the time of writing this paper, the Israeli kid's version of "You Can't Ask That" includes three seasons and over 30 episodes, focusing on kids from different backgrounds. In designing our intervention, we chose to focus on three different episodes. These episodes focus on Arab kids, children of immigrant foreign workers from the Philippines (Hebrew: *Ovdim Zarim*), and visually impaired children. We decided to focus on these groups, given that the nature of prejudice and sensitive issues relating to each group are substantively different, albeit very salient for children. As shown in Figure A1, in both Studies 1 and 2, although the overall positive affect toward children from these groups varies, at baseline, all groups considered in the intervention were perceived to be quite different from the ingroup on average.

A1.1 TV Show Content

In this section, we list the questions sent from home audiences to the children filmed as part of the TV show and the associated discussion topics inspired by these questions. In essence, questions from home audiences were presented to children in the studio to inspire and guide



(a) Thermometers Study 1

(b) Similarity Study 1

(c) Thermometers Study 2

(d) Similarity Study 2

Figure A1: **Pre-treatment levels of outgroup affect and perceptions of intergroup similarity in Studies 1 and 2.** This figure demonstrates variation in affect toward different outgroups and perceptions of dissimilarity to outgroups amongst students in the pre-treatment period of Studies 1 and 2.

in-depth discussions about prejudicial taboos and other sensitive and complex topics. While varying in their directness, these questions represent issues that children would be too shy to ask an outgroup to their face. Moreover, many of these questions either directly focused on sensitive topics or sparked discussions directly related to intergroup grievances, disagreements, and dehumanizing stereotypes. Regardless, upon responding to these questions, children raised various additional issues regarding taboos and sensitive topics at the core of intergroup relations.

Below is a list of questions and associated discussion topics presented in the intervention. Each question is marked with a –, and presented alongside a description of discussion topics marked by *. This list aims to give readers a sense of the content described as part of the TV

series “You Can’t Ask That”.

- **Arab Children**
  - Are you Arab, are you Israeli? What are you?
    - Discussion of the complexity of social identities. Different children discuss the importance of different identities (Muslim, Arab, Palestinian, Israeli), and some explicitly state that their Israeli identity is least salient to them.
    - Discussion of the distinction between Arab and Palestinian identities.
    - Many children discuss their varying experiences of being Palestinian and living on territory that was conquered by the Israeli state. Following up on this issue, some children emphasize that they feel like they do not fully belong to a Jewish-Israeli or Palestinian community and are “stuck in the middle” between two groups in conflict.
    - The children discuss the challenges of the Nakba and Israeli Independence day coinciding. On this topic, one student explicitly stated: “*Jews conquered the land, and they got a new country, Arabs, their lives changed in 180 degrees once they were told their homes aren’t their own.*” This quote emphasizes the sensitive nature of being an Arab/Palestinian in Israel.
    - Multiple children emphasize that they cannot relate to the Israeli anthem. They explain that they do not sing the anthem because they cannot relate to the multiple Jewish symbols the anthem celebrates.
  - Why is Arabic a scary language?
    - Children acknowledge that Arabic is a rich, albeit complicated, language.
    - Several students argue that the only reason Arabic is considered a scary language is that Israelis associated it culturally and politically with terrorism. The

children note that Jews call Arabic the language of the terrorist and that when Jews hear someone speaking Arabic, they think they are planning to attack them.

    * Several children describe experiences in which Jewish Israelis blamed them for Palestinian violence or associated them with intergroup violence. In addition, many children shared their experience of suffering from violent slurs (e.g. "Death to Arabs"). Some children emphasize that these generalizations offend them for many reasons, including the fact that as Arabs living in Israel, they and their families also fear and suffer from Palestinian rocket attacks and other forms of violence.

    * Multiple children emphasize their frustration with the fact that most Israelis cannot articulate a single sentence in Arabic.

- Do you listen to Arabic music?

    * Children describe their music preferences. Some emphasize that they listen to both Hebrew and Arabic music; others note that they largely listen to music in English. Overall responses to this question emphasize much variability across different children.

- Did you ever face racism because you are an Arab?

    * All children emphasize that they experienced racism in the past. Many students were confronted with offensive slurs, such as: "dirty terrorist" or "stinky Arab," and some suffered from actual physical violence (stone throwing). One child described an experience of being aggressively interrogated in the airport. Another child described being profiled by security personnel when going to the mall, and all these experiences were linked to underlying racism in Israeli

society.

  - Given previous experiences with racism, children emphasize that they try to avoid speaking Arabic to limit the possibility of awkward or uncomfortable situations.
  - Many children note that their friends have told them insensitive statements, such as "You don't look like an Arab." The children emphasize that the origin of racism and its manifestations relate to the fact that most Jewish Israelis never meet Arabs.

- Do you watch shows in Hebrew or Arabic?
  - Children describe their TV preferences; some emphasize that they watch Arabic language content, and others note that they mainly watch Hebrew and English language content.
  - Several children mention an Israeli TV show named "Fauda," which tells the story of a counter-terrorist unit operating in the West Bank and Gaza and includes a substantial amount of Arabic. The children note that they find the show awkward because the Arabic used in the show does not always sound normal or "correct" to them.

- Are we (i.e., Jews and Arabs) enemies?
  - Many children reject the premise of the question since they do not think Jewish and Arab children are (or should be) enemies. Indeed, many children note that despite differences in nationality and religion, Jewish and Arab children should be friends and get along.
  - Children also acknowledge the sensitive nature of intergroup relations in Israel, stating that there are a lot of tensions between Jews and Arabs. One child notes

that this land (i.e., Israel/Palestine) was initially meant to be a country for Arab people, but clearly, there is now a Jewish state. Even though the presence of a Jewish state is not ideal, everyone has to try and get along together because "we can't change the past."

* Several children acknowledge the merits of diversity, emphasizing that Jewish-Arab cooperation can yield social cohesion and societal strength. Moreover, several children note that most Jewish Israelis and Arabs want to get along together but that some "extreme actors" on both sides try to spoil positive peace and harmonious intergroup relations.

- **Immigrant Children**

  - What does it mean that your parents are foreign workers?

    * Children explain their legal immigration status and how that status relates to their personal history. For example, one child explains that their parents came from the Philippines and have lived in Israel for over 20 years working as care-givers. Another child goes on to explain that when their parent's visas expired, they decided to violate the law and stay in Israel because they had an urgent need to provide for their families abroad.

    * Children emphasize their varying forms of identities (e.g., Israeli, Filipino, etc...). Still, all children emphasize that their Israeli identity is rooted in their experience of growing up in Israel and taking part in the Israeli education system.

    * Some of the children acknowledge that the Israeli government operates according to the law when it classified them as illegal residents, but also emphasize that when the government "follows the law," that has detrimental consequences

for undocumented children. One child notes that sometimes laws can be unethical and that many historical changes that we care about, including the creation of the Israeli state, resulted from violating existing laws and norms.

  - There is a lengthy discussion of the motivations that led children's parents to violate Israeli immigration law and remain in Israel. Several children explain that their parents remained in Israel to ensure that their children would have a better and more stable future.

- Do you think that we (natives) are racist?

  - Several children note that they think some Israelis are racist and allude to instances in which they suffered from racial slurs and hurtful statements. For example, many children experience inappropriate staring, others experience being mixed up to be part of a cleaning staff, and they often receive statements such as "go back to your home."

  - Reflecting on Jewish Israeli racism, one child notes that statements asking him to go home are hurtful because Israel is his home. Another child noted that perhaps Jews are afraid of demographic shifts in which non-Jews will become a growing segment of the Israeli population. Perhaps, this child suggests, that is a motivating factor for racism in Israeli society.

- Are you afraid to be deported?

  - There is a lot of variation in response to this question. Some kids explicitly state their fear of being deported, while others emphasize that they cannot be deported because they luckily obtained Israeli citizenship status.

  - Elaborating about the fear of being deported, one child stated that "*I think every day about what will happen if I will be deported. I rarely leave the house. I*

avoid leaving home. We move every several weeks. I check for cops before I leave the house. It's really stressful." In a similar vein, another child describes her experience of police officers raiding her house and eventually placing her and her family in the Givon jail for ten days. The child went on to describe how her classmates (both Jewish Israelis and non-Jewish immigrants) protested outside the jail where she was placed until she and her family were released.

- Do you eat weird food?
  - Many children initially laugh at this question and emphasize that it is ridiculous.
  - Upon further reflection, some children note that their Jewish Israeli friends often ask them if they eat snakes and mice.
  - More generally, children emphasize that they do not eat "weird" food. They explain that their food might be different from Israeli food, but there is nothing weird about their own food.
  - Children acknowledge the fact that different social and cultural groups might traditionally eat different types of food. They emphasize that when they started bringing their food to school, some kids asked questions about it, but over time they would share their food, and their peers really liked it.
- Are you curious about visiting the Philippines?
  - Children note that they would be excited to visit the Philippines and meet their extended family. However, at the same time, many emphasize that they would not want to live there because Israel is their home, and they are not fluent in Tagalog.
  - This question generates a conversation about relationships with family abroad. Some children note that they talk to family on the phone but have never met

their immediate and extended family in person because they cannot leave Israel without sacrificing their residency status standing. Those children discuss the emotional toll of being away from family and having no way of visiting them or knowing when they might meet in person.

  - Do you feel Israeli?
    - All children emphasize that they feel Israeli. They note that they grew up in Israel and went to Israeli schools. Many of the children explained that they are fully immersed in Israeli culture and that Hebrew is their mother tongue. One child stated in response to this question: "I dream in Hebrew; I speak Hebrew, I sing Hebrew. I am Israeli in my soul."
    - Alongside strong identification as Israelis, some children also point to their complex and layered identities, noting that although they have never visited the Philippines, they still identify as Filipino.

- **Visually Impaired Children**
  - What do you see?
    - There is a lot of variation in response to this question. Some children note that they never saw anything, and it is hard to compare their experience with the experience of other visually abled people. Other children note that they see some colors or blurry scenes. Elaborating on this point, some children explain the physical reason for which they are visually impaired. For example, one child notes that because he cannot control the movement of his eyes, he has trouble with vision. Another child explains how a pigment condition they suffered from has affected their vision.

* In response to the question, it becomes apparent that different children lost their vision in different stages of life as a consequence of different medical conditions. When discussing this process, one child noted that "*It wasn't fun becoming blind. I needed to come to terms with what I was missing out on. The last time I saw a person was two years ago*."

– Do you trip a lot?

  * Many children emphasize that they have trouble navigating space, and that they often trip or bump into different objects. In response to this question, it appears that there is a lot of variation in children's experiences in navigating space. Indeed, different children describe varying challenges of navigating space with impaired vision and how they have learned to overcome such challenges.

  * One child describes how he broke both his hands from tripping. Another child describes a moment in which she bumped into a garbage pale, thought the garbage pale was a human being, and felt very embarrassed during the experience. Many of the children experienced bumping into polls and having peers laugh at them for that. In response to such incidents, one child noted that "*When people laugh at you, rather than with you, it's uncomfortable*." Another child noted that they "*Don't let anything bring [them] down*."

– Do other kids bully you because you are visually impaired?

  * Children elaborate on the different insults they receive in school. Some children note that other kids stick fingers in their eyes or call them by name. In addition, some children note that other kids constantly challenge them in insensitive ways (e.g., guess how many fingers I am raising).

  * One child discusses avoidance. Specifically, they note that "*people do not know*

how to engage with me, and they avoid me because they feel uncomfortable next to me." The child further explains that they think that many people feel uncomfortable discussing and engaging with issues, topics, and people they are not accustomed to.

  * Another child emphasizes that he forgets about the people that insult him but cannot forget about the insults themselves and that the content of these insults poses personal challenges.

- Do you participate in gym class?

  * Many children note that they take an active part in gym class, that they try and participate like everyone else, and that sport is one of their favorite activities.

  * At the same time, several children note that playing with a ball induces much anxiety because it is hard to anticipate balls when being visually impaired.

  * One child notes that he loves running and jumping and elaborates on how he runs with a running partner. He emphasizes that many non-visually impaired children are surprised by the fact that he is very active. Another child from Israel's national goalball team provides an explanation about the sport, which was designed for visually impaired athletes.

- What is the most surprising thing that you taught yourself to do?

  * Different children mention their surprising skills.

  * One child elaborates about how when someone is challenged with regards to a specific sense (e.g., vision), other senses can compensate for that (e.g., hearing). The child goes on to describe the hearing skills that allow them to anticipate and recognize people by the sound of their footsteps, explaining how this skill helps him excel in music and be a good hide-and-seek player.

- If you could choose to see one thing, what would you want to see?
    - One child elaborates on how he wishes he could see stars. He emphasizes that his family always goes stargazing, and he feels left out and wishes he could have the same experience as his siblings.
    - Several children note that they wish they could see their family and closest friends. Other children note that they are really curious to learn about their own looks, and wish they could have seen themselves. They would like to learn about the color of their own eyes, and see how they look.
    - One child that lost their vision at a later age noted that "*I saw everything I wanted to see in life. My eyes left this world satisfied*."

A2 Classroom Curriculum

As mentioned above, our intervention focused on three episodes of the TV series "You can't Ask That" and included four classes. The first three classes centered around the episodes noted above relating to Arab, visually impaired, and immigrant children. The fourth class presented a summary of all episodes and a review of the show's themes. Based on our theoretical framework, which emphasizes the value of parasocial intergroup contact combined with constructive conversations which link sensitive topics with psychological mechanisms of intergroup similarity, within-group heterogeneity, and perspective-taking, we designed our classes to constructively engage with show participants and link the sensitive topics they discuss with theoretical mechanisms of prejudice reduction.

Specifically, our first three classes focused on a particular social group presented in a particular episode. Class number 1 focused on Arab children. In the process of unpacking the Arab children episode and the sensitive topics it discussed, children learned about the concepts of

intergroup similarity and group heterogeneity and applied them to the outgroup discussed in the classroom.

Class number 2 focused on visually impaired children. In the process of exposure to charismatic visually impaired children and discussion of a series ore related sensitive topics, children learned about the concept of intergroup similarity and the value of perspective-taking, applying these concepts to the outgroup discussed in class. Class number 3 focused on children of immigrants. In the process of exposure to these children and discussion of related sensitive topics, children learned about the concept of perspective taking, applying this concept to the outgroup discussed in class. Finally, in class number 4, children watched a brief review of all three episodes and then engaged in summary activities relating to all three psychological mechanisms discussed in classes 1-3.

A2.1 Study 1 Implementation

In our first study, all classes were delivered by an educational practitioner employed by aChord center, our implementation partner. The practitioner was trained to deliver classroom activities ahead of time by the research team and a pedagogy professional and was instructed to deliver content according to carefully curated slides prepared by the research team and the pedagogy practitioner. These slides included instructions for classroom activities to engage students with the core objectives of the intervention.

For example, during the first class, after watching a 15-minute episode regarding Arab children, the students engaged in a classroom activity in which they were required to reflect on the similarities between students in their class and children depicted in the TV series and the differences between various students' portrayed in the TV show. This activity was designed to teach students about concepts of intergroup similarity and within-group heterogeneity.

In the second lesson, children watched an episode regarding visually impaired children. Af-

ter doing so, they played a game where students had their classmates cover their eyes. Students with covered eyes were guided by their friends for a walk around the school to provide them with an opportunity to take the perspective of a visually impaired child. Following the activity, students reflected on their experiences and feelings in a classroom discussion. Finally, in the third class, after watching the episode about children of immigrants, students participated in a classroom discussion in which they were invited to imagine how the kids portrayed in the show felt when engaging in different challenging situations described in the episode. This activity was designed to encourage students' active perspective-taking with their outgroups.

While each class focused on a particular psychological mechanism (e.g., perspective taking, group variability, or intergroup similarity), the curriculum was designed to focus on exposure to charismatic outgroups and discussion of sensitive topics directly linked with all three psychological mechanisms that appear in each episode. Moreover, the educational practitioner was instructed to link the different classes, and indeed, each lesson started with a brief overview of recent class activities relating to the intervention.

A2.2 Study 2 Implementation

In study 2, the content of our intervention remained the same. However, to assess one dimension relating to the scalability of our intervention, we took a "train the trainers" approach and delegated the responsibility of treatment delivery to teachers (rather than an external educational practitioner). This change was implemented in order to assess whether our curriculum can effectively reduce prejudice when implemented by teachers with varying skills, incentives, ideological preferences, and motivations. To familiarize teachers of treated classes with our intervention, we took the following three steps.

First, all teachers were invited to participate in an hour-long information session about our intervention. In this session, our partner organization introduced the intervention, described

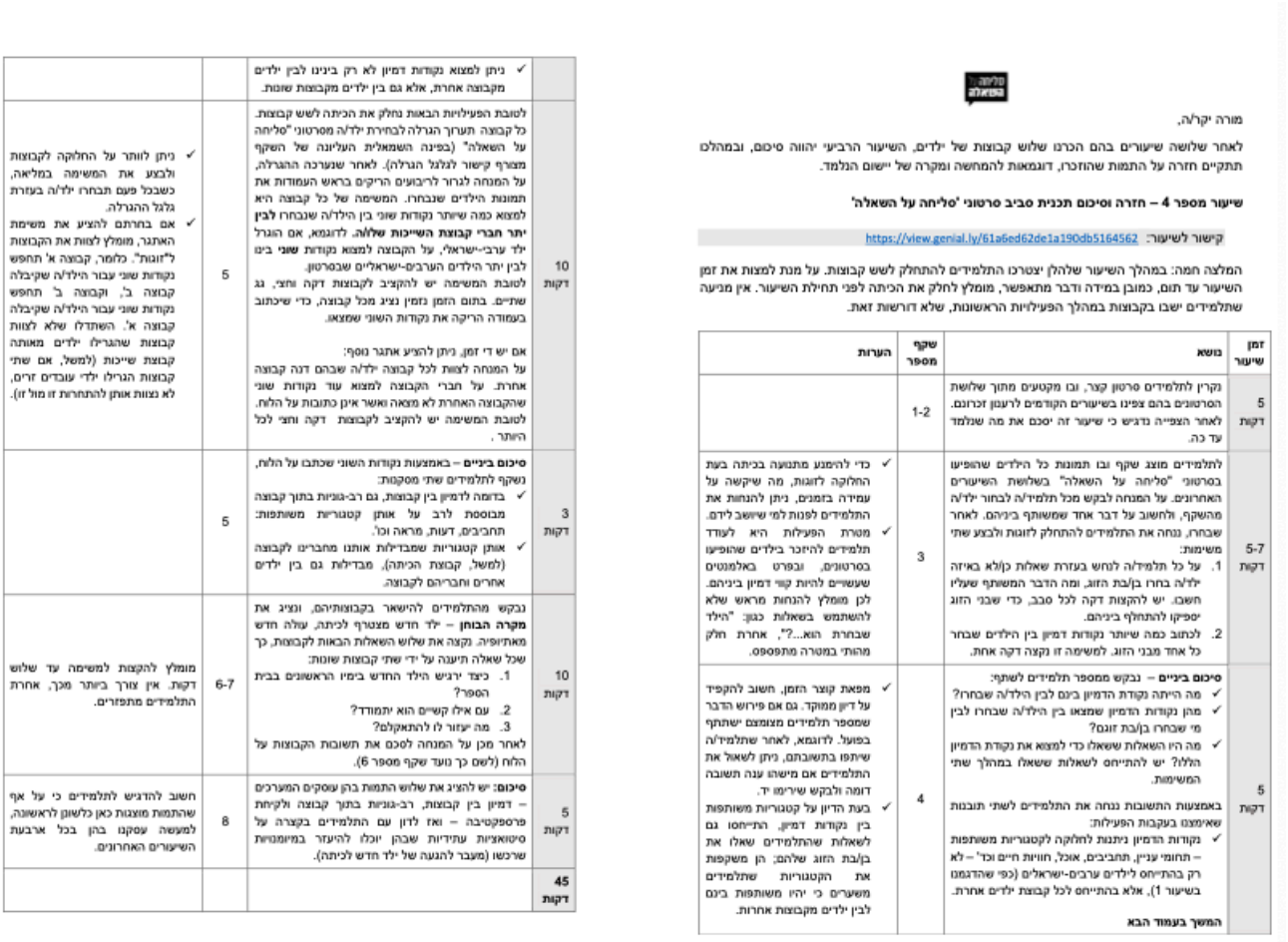
the different lessons, emphasized the psychological mechanisms that inspired the intervention, and opened the room for any clarifying questions. Note that these sessions were only open to teachers whose classes were assigned to the treatment group, and not all teachers attended the session.

Second, each teacher responsible for a treated class received a detailed lesson plan for each one of the four sessions of our intervention. An example of a lesson plan is provided in Figure A2. These lesson plans provided a link to the relevant episode and accompanying class slides. Moreover, the lesson plan included information about the main objectives of the class and a breakdown of all activities to be implemented in a given session. We provided precise instructions about the time that should be allocated to each activity to maximize standardization across teachers and classrooms. Naturally, by virtue of our train-the-trainers approach, the quality and nature of implementation varied by teacher and classroom.

Finally, we designated a point of contact in each school who was in charge of updating our field coordinator about the progress of treatment implementation every week. Our field coordinator worked with each school's point of contact to schedule all activities relating to the field experiment and address any questions arising during the implementation period.

Through our weekly communications with school-specific points of contact, we ensured that teachers from treatment classes implemented all four sessions of the interventions before proceeding to post-treatment surveys. We also ensured that control-class teachers did not have access to intervention materials (at the time of the study). Our initial goal was that each school would roll out the intervention over four successive weeks. However, all schools needed more time for intervention rollout because of unexpected schedule constraints.

To assess treatment implementation and take-up in our second study, we included questions in our post-treatment survey, asking students whether they recall watching specific intervention-related videos in class. We use these survey items in order to create a treatment take-up indicator



מורה יקר/ה,

לאחר שלושה שיעורים בהם הכרנו שלוש קבוצות של ילדים, השיעור הרביעי יהווה סיכום, ובמהלכו תתקיים חזרה על התמות שהוזכרו, דוגמאות להמחשה ומקרה של יישום הנלמד.

שיעור מספר 4 – חזרה וסיכום תכנית סביב סרטוני 'סליחה על השאלה'

קישור לשיעור: https://view.genial.ly/61a6ed62de1a190db5164562

המלצה חמה: במהלך השיעור שלהלן יצטרכו התלמידים להתחלק לשש קבוצות. על מנת למצות את זמן השיעור עד תום, כמובן במידה ודבר מתאפשר, מומלץ לחלק את הכיתה לפני תחילת השיעור. אין מניעה שתלמידים ישבו בקבוצות במהלך הפעילויות הראשונות, שלא דורשות זאת.

זמן שיעור	נושא	שקף מספר	הערות
5 דקות	נקרין לתלמידים סרטון קצר, ובו מקטעים מתוך שלושת הסרטונים בהם צפינו בשיעורים הקודמים לרענון זכרונם. לאחר הצפייה נדגיש כי שיעור זה יסכם את מה שלמדנו עד כה.	1-2	
5-7 דקות	לתלמידים מוצג שקף ובו תמונות כל הילדים שהופיעו בסרטוני "סליחה על השאלה" בשלושת השיעורים האחרונים. על המנחה לבקש מכל תלמיד/ה לבחור ילד/ה מהשקף, ולחשוב על דבר אחד שמשותף ביניהם. לאחר שבחרו, נחה את התלמידים להתחלק לזוגות ולבצע שתי משימות: 1. על כל תלמיד/ה לנחש בעזרת שאלות כן/לא באיזה ילד/ה בחרו בן/בת הזוג, ומה הדבר המשותף שעליו חשבו. יש להקצות דקה לכל סבב, כדי שבני הזוג יספיקו להתחלף ביניהם. 2. לכתוב כמה שיותר נקודות דמיון בין הילדים שבחר כל אחד מבני הזוג. למשימה זו נקצה דקה אחת.	3	✓ כדי להימנע מתסיסה בכיתה בעת החלוקה לזוגות, מה שיקשה על עמידה בזמנים, ניתן להנחות את התלמידים לפנות לתלמי שיושב לידם. ✓ מטרת הפעילות היא לעודד תלמידים להיזכר בילדים שהופיעו בסרטונים, ובפרט באלמנטים שעשויים להיות קווי דמיון ביניהם. לכן מומלץ להנחות מראש שלא להשתמש בשאלות כגון: "הילד שבחרת הוא...?", אחרת חלק מהתוכן במטרה מתפספס.
5 דקות	**סיכום ביניים** – נבקש ממספר תלמידים לשתף: ✓ מה הייתה נקודת הדמיון בינם לבין הילד/ה שבחרו? ✓ מהן נקודות הדמיון שמצאו בין הילד/ה שבחרו לבין מי שבחרו בן/בת זוגם? ✓ מה היו השאלות ששאלו כדי למצוא את נקודת הדמיון הלזו? יש להתייחס לשאלות ששאלו במהלך שתי המשימות. באמצעות התשובות ננחה את התלמידים לשתי תובנות שאימצנו בעקבות הפעילות: ✓ נקודות הדמיון ניתנות לחלוקה לקטגוריות משותפות – תחומי עניין, תחביבים, אוכל, חוויות חיים וכד' – לא רק בהתייחס לילדים ערבים-ישראלים (כפי שהדגמנו בשיעור 1), אלא בהתייחס לכל קבוצת ילדים אחרת. **המשך בעמוד הבא**	4	✓ מפאת קוצר הזמן, חשוב להקפיד על דיון ממוקד. גם אם פירוש הדבר שמספר תלמידים מצומצם ישתתף בפועל. לדוגמא, לאחר שתלמיד/ה שיתפו בתשובתם, ניתן לשאול את התלמידים אם מישהו ענה תשובה דומה ולבקש שירימו יד. ✓ בעת הדיון על קטגוריות משותפות בין נקודות דמיון, התייחסו גם לשאלות שהתלמידים שאלו את בן/בת הזוג שלהם; הן משקפות את הקטגוריות שתלמידים משערים כי יהיו משותפות בינם לבין ילדים מקבוצות אחרות.

	✓ ניתן למצוא נקודות דמיון לא רק בינינו לבין ילדים מקבוצה אחרת, אלא גם בין ילדים מקבוצות שונות.		
10 דקות	לטובת הפעילויות הבאות נחלק את הכיתה לשש קבוצות. כל קבוצה תערוך הגרלה לבחירת ילד/ה מסרטוני "סליחה על השאלה" (בפינה השמאלית העליונה של השקף מצורף קישור לצלגל הגרלה). לאחר שנערכה ההגרלה, על המנחה לגרור לריבועים הריקים בראש העמודים את תמונות הילדים שנבחרו. המשימה של כל קבוצה היא למצוא כמה שיותר נקודות שוני בין הילד/ה שנבחרו **לבין יתר חברי קבוצת השייכות שלו/ה**: לדוגמא, אם הוגרל ילד ערבי-ישראלי, על הקבוצה למצוא נקודות **שוני** בינו לבין יתר הילדים הערבים-ישראליים שבסרטון. לטובת המשימה יש להקציב לקבוצות דקה וחצי, גג שתיים. בתום הזמן נזמין נציג מכל קבוצה, כדי שיכתוב בעמודה הריקה את נקודות השוני שמצאו. אם יש די זמן, ניתן להציע אתגר נוסף: על המנחה לצוות לכל קבוצה ילד/ה שבהם דנה קבוצה אחרת. על חברי הקבוצה למצוא עוד נקודות שוני שהקבוצה האחרת לא מצאה ואשר אינן כתובות על הלוח. לטובת המשימה יש להקציב לקבוצות דקה וחצי לכל היותר.	5	✓ ניתן לוותר על החלוקה לקבוצות ולבצע את המשימה במליאה, כשבכל פעם תבחרו ילד/ה בעזרת גלגל ההגרלה. ✓ אם בחרתם להציע את משימת האתגר, מומלץ לצוות את הקבוצות ל"זוגות". כלומר, קבוצה א' תחפש נקודות שוני עבור הילד/ה שקיבלה קבוצה ב', וקבוצה ב' תחפש נקודות שוני עבור הילד/ה שקיבלה קבוצה א'. השתדלו שלא לצוות קבוצות שהגרילו ילדים מאותה קבוצת שייכות (למשל, אם שתי קבוצות הגרילו ילדי עובדים זרים, לא נצוות אותן להתחרות זו מול זו).
3 דקות	**סיכום ביניים** – באמצעות נקודות השוני שכתבו על הלוח, נשקף לתלמידים שתי מסקנות: ✓ בדומה לדמיון בין קבוצות, גם רב-גוניות בתוך קבוצה מבוססת לרב על אותן קטגוריות משותפות: תחביבים, דעות, מראה וכו'. ✓ אותן קטגוריות שמבדילות אותנו מחבריו לקבוצה (למשל, קבוצת הכיתה), מבדילות גם בין ילדים אחרים וחבריהם לקבוצה.	5	
10 דקות	נבקש מהתלמידים להישאר בקבוצותיהם, ונציג את **מקרה הבוחן** – ילד חדש מצטרף לכיתה. עולה חדש מאתיופיה. נקוב את שלוש השאלות הבאות לקבוצות, כך שכל שאלה תיענה על ידי שתי קבוצות שונות: 1. כיצד ירגיש הילד החדש בימיו הראשונים בבית הספר? 2. עם אילו קשיים הוא יתמודד? 3. מה יעזור לו להתאקלם? לאחר מכן על המנחה לסכם את תשובות הקבוצות על הלוח (לשם כך נועד שקף מספר 6).	6-7	מומלץ להקצות למשימה עד שלוש דקות. אין צורך ביותר מכך, אחרת התלמידים מתפזרים.
5 דקות	**סיכום:** יש להציג את שלוש התמות בהן עוסקים המערכים – דמיון בין קבוצות, רב-גוניות בתוך קבוצה ולקיחת פרספקטיבה – ואז לדון עם התלמידים בקצרה על סיטואציות עתידיות שבהן יוכלו להיעזר במיומנויות שרכשו (מעבר להנגה של ילד חדש לכיתה).	8	חשוב להדגיש לתלמידים כי על אף שהתמות מוצגות כאן כלשונן לראשונה, למעשה עסקנו בהן בכל ארבעת השיעורים האחרונים.
45 דקות			

Figure A2: **Study 2 lesson plan.** This Figure presents an example of the fourth lesson plan provided to teachers in charge of treatment implementation.

ranging from 0 (no recollection of exposure to videos) to 4 (recollection of exposure to all four intervention videos). This indicator is a useful albeit imperfect measure of treatment implementation and take-up.

In Figure A3, we report the distribution of our treatment take-up indicator. Notably, in most treated classes, students recall being exposed to at least 3 intervention-related videos. That said, since this indicator is based on self-reported exposure and since points of contact confirmed implementation in all treatment classes, lower rates of exposure do not imply that a lesson

was not implemented in a given class. Instead, these lower rates are likely driven by students' absence or imperfect recall.

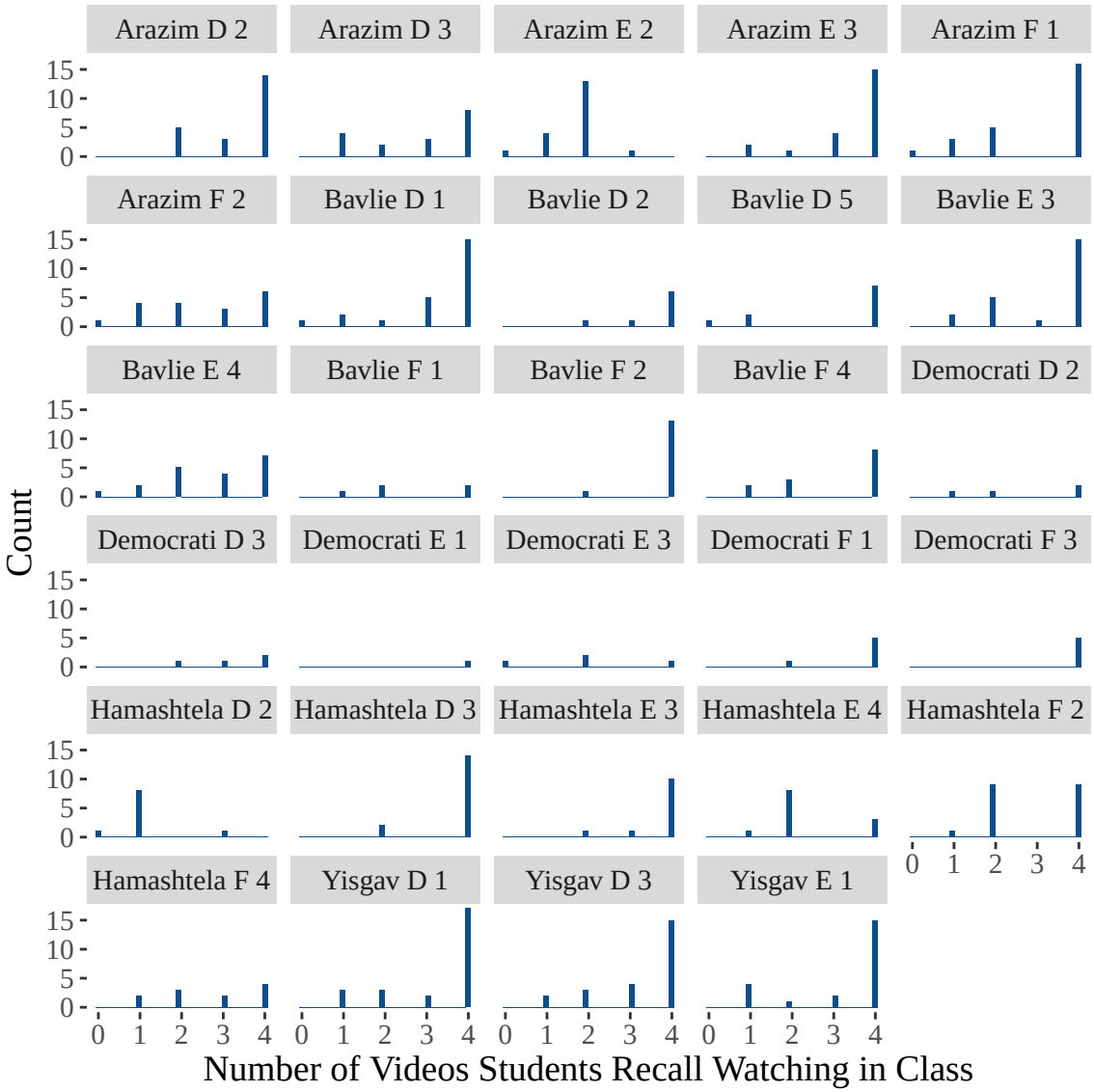


Figure A3: **Study 2 treatment recall.** This figure reports the distribution of students' responses to survey questions regarding treatment take-up. Each panel reports a treatment take-up indicator ranging from 0-4, representing the number of videos a student recalls watching in class.

A3 Survey Methodology

A3.1 Survey Timing

Studies 1 and 2 included a pre-treatment survey implemented before rolling out our intervention and a post-treatment survey implemented 1-2 weeks following the end of the intervention. Study 2 further included a second survey wave implemented 8-13 weeks post-treatment.

Given the challenge of sampling children in schools without interfering with ongoing classes, there was a degree of variation in the exact timing separating between treatment and outcome collection across different subjects in the treatment group. In Figures A4-A5 we plot this variation for students in the treatment condition. Generally, the average number of days buffering treatment implementation and outcome collection for treated students in Study 1 was just above 12 days. Similarly, the average number of days buffering treatment implementation and outcome collection for treated students in Study 2 was just below ten days for the first survey and just above 70 days for the second survey.

A3.2 Survey Implementation and Main Outcomes

Surveys were programmed on Qualtrics and were distributed via tablets to small groups of children by a research assistant. To minimize concerns regarding social desirability bias, children responded to the survey privately, were told that there are no 'right and wrong answers." and were assured multiple times that all their responses would remain fully anonymous.

In both studies' pre and post-treatment surveys, we included measures of all our attitudinal outcomes of interest. Moreover, in Study 1, our endline survey included two behavioral measures asking students to sign up for an intergroup contact initiative and asking students to report social groups that should be covered in future episodes of the TV-Series "You Can't Ask That".[1]

[1] As well as questions that they might want to ask those social groups. We do not report these measures given

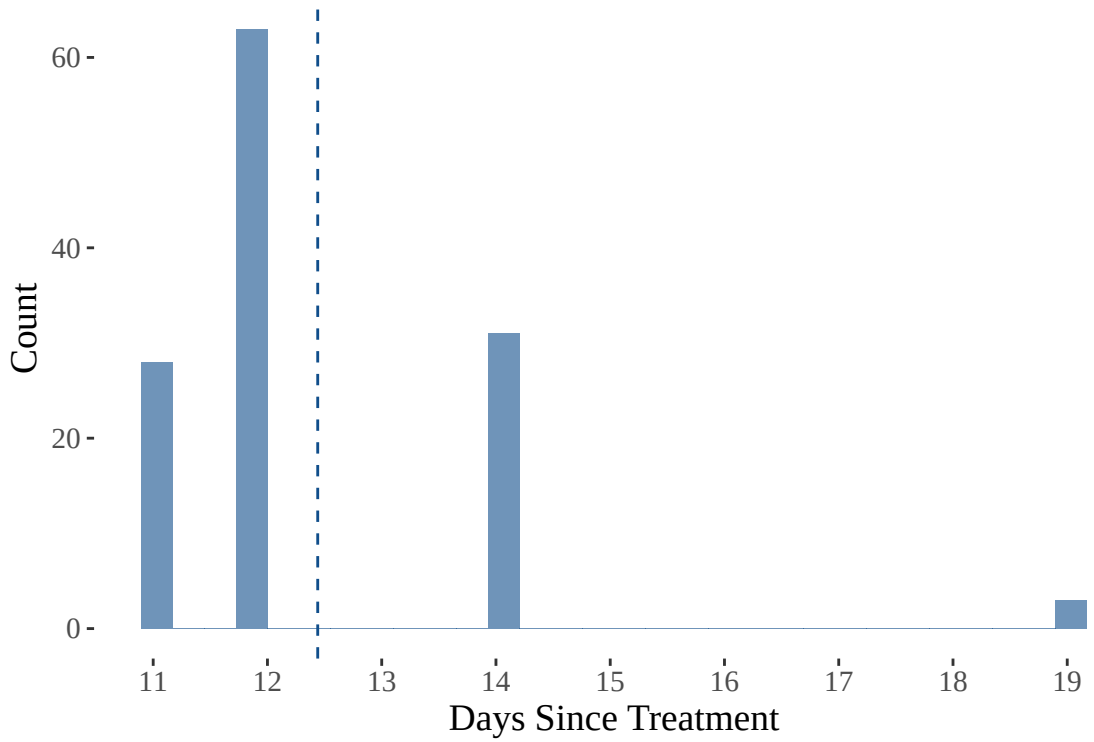


Figure A4: **Distribution of time between treatment and survey implementation in Study 1 for treated students.**

In Study 2, our endline surveys included two behavioral measures asking students to sign up for an intergroup contact initiative (in both waves) and allowing students to select a pro-diversity bracelet as compensation for participation in our surveys (second wave of Study 2). Specifically, all study participants were told that they could choose one of two bracelets as compensation for their participation in our surveys. As reported in Figure A6, one bracelet included a personal reassurance statement, and the other included a pro-diversity statement. We employ the take-up of a pro-diversity bracelet as a behavioral measure of students' support for diversity and their willingness to signal to their peers that they value inclusion.

We describe our key outcome measures in Table A1. Note that in addition to these outcomes,

students' confusion regarding this outcome which arose in the implementation of Study 1, and motivated us to omit this measure in Study 2.

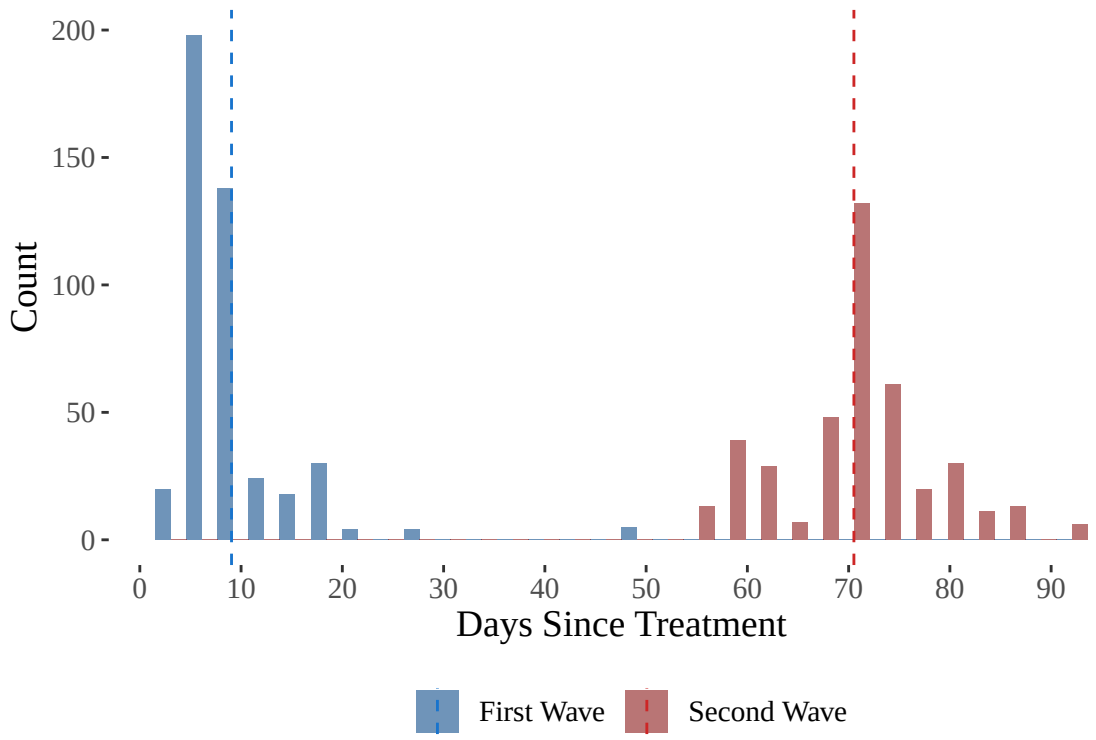


Figure A5: **Distribution of time between treatment and survey implementation in Study 2 for treated students.**

in Study 2 we further include questions about the psychological mechanisms underlying the program (intergroup similarity, group heterogeneity, and perspective-taking). We provide the wording of these questions, as well as the wording of all other survey questions, in the following section.

Table A1: Main Outcomes

	Question	*Range*	*Measured Pre and Post*
Thermometer Index	What are your feelings towards blind/immigrant/Ultra-Orthodox children?	0-100	Yes
Diversity Index	To what extent do you agree with the following statements: 1) I can learn a lot from different children. 2) It is important to learn from other children even when their ideas are different then mine. 3) I enjoy studying with children who are different from me. 4) I enjoy playing with children who are different from me. 5) In team work... it helps the team a lot when there are children who are different from one another.	1-5	Yes
Contact Intention Index	Think about a blind/Arab/Immigrant/Ultra-Orthodox child you do not know. To what extent would you like to: 1) Play with this child. 2) Invite this child to your birthday. 3) Help this child with their difficulties in homework. 4) Help this child if they were lost.	1-5	Yes
Register for Contact	There may be a project bringing together children from different backgrounds. Would you like to sign up for this project?	Yes/No	Only Post
Pro-Diversity Bracelet (Study 2 Wave 2)	After submitting the survey, students were asked what type of bracelet (if any) they would like to receive as compensation.	Yes/No Diversity	Only Post

(a) I am good exactly the way I am (b) In this school everyone belongs

Figure A6: **Description of the bracelets we employed to measure pro-diversity behavior in Study 2.** Panel (a) portrays the bracelet with a personal reassurance message. The bracelet text notes: "I am good exactly the way I am," Panel (b) portrays the bracelet with a pro-diversity message. The bracelet text notes: "In our school, everyone belongs." We consider take-up of the pro-diversity bracelet depicted in Panel (b) as a behavioral measure of support for diversity.

A3.3 Survey Instruments

All surveys employed in Study 1 and Study 2 included common demographic and social questions and questions relating to intergroup relations. When needed, we modified survey wording to ensure that questions were clear to students in grades 4-6. Below we report the English translation of our survey. We mark with † behavioral measures that were only included in all post-treatment waves. We mark with $\wedge$ all survey measures that were only included in Study 1. Finally, we mark with $\Cap$ all survey measures that were only included in Study 2.

- Demographics
  - Boy Or Girl?
  - Age?
  - What grade are you in?
  - What is your class name?
- Main Attitudes
  - In this question we are going to ask you to report how many bad and cold feelings, or good and warm feelings you feel towards kids from specific groups. If you feel positive feelings towards kids from a specific group move your pointer towards the warmer and higher portion of the scale. If you feel negative feelings towards kids from a specific group move your pointer towards the colder and lower portion of the scale (Two practice rounds, Arab Kids, Children of Immigrants, Blind Kids, Ultra-Orthodox Kids). *0-100 point scale*.
  - To what extent do you agree with the following statements:
    * I can learn a lot from kids who are different from me *five point scale*
    * It is important to hear other kids' opinions even when their opinions are different than mine *five point scale*
    * I enjoy learning with kids who are different from me *five point scale*
    * I enjoy playing with kids who are different from me *five point scale*
    * In group activities (for example in gym class) it helps when a group includes kids who are different from one another *five point scale*
  - Please take a moment to think about a (Arab/blind/Immigrant/Ultra-Orthodox) child that you do not know, to what extent would you like to:

    - Play with this kid *five point scale*
    - Invite this kid to your birthday *five point scale*
    - Help this kid with their homework *five point scale*
    - Assist this kid if they were lost *five point scale*

- Psychological Mechanisms
  - People can be similar in some ways, and different in other ways (for example in their personality traits, hobbies, interests, or looks). How similar or different are you from the kids listed below (Arab Kids, Children of Immigrants, Blind Kids, Ultra-Orthodox Kids)? *Five point scale*.
  - People can be similar in some ways, and different in other ways (for example in their personality traits, hobbies, interests, or looks). How similar or different are Arab Kids / Children of Immigrants/Blind Kids/Ultra-Orthodox Kids from one another? *Five point scale*.⋒
  - Different people from different social groups might experience different types of challenges and hardships. At times we might want to try and understand those challenges and hardships. In order to do so we can think about how those people feel, what they think about, and put ourselves in their shoes. Sometimes we do this, and other times we do not. To what extent do you think it is it important to do this towards each one of the following groups (Arab Kids / Children of Immigrants/Blind Kids/Ultra-Orthodox Kids)? *Five point scale*.⋒

- Behavioral Measures†
  - "You Can't Ask That" is a TV show that collects questions from children to ask children from other groups. Currently, there is a new season that is filming new

episodes about children from different social groups. Are there social groups that you would be interested to learn about? Please list any groups that you would like to see included in future episodes so that we could share this information with the broadcasting team. (We gave children 6 open spaces to mention social groups to be included in future episodes. For each mentioned group, respondents were given space to include questions of interest).∧

  - There may be an activity in the near future, that will bring together children from different backgrounds (secular children, religious children, Arab children, blind children, ultra-orthodox children, and children of immigrants) to meet each other. If you would like to be included in this activity please check this box.
  - After completing the survey, the person overseeing survey implementation showed each student the personal reassurance and diversity bracelets and asked them what bracelet they want (if any). ⋓ (only in second post-treatment wave).

- Miscellaneous †
  - Did you ever watch the TV series "You Can't Ask That" at home?
  - Did you watch the following episodes recently in class (list of all three episodes, and an overview of all episodes)? Only for the treatment group.
  - In the past month, some classes engaged in some activities as part of a research project. Do you know the topic of this project or its objective? (open-ended question). Only for respondents who answered no in above question.

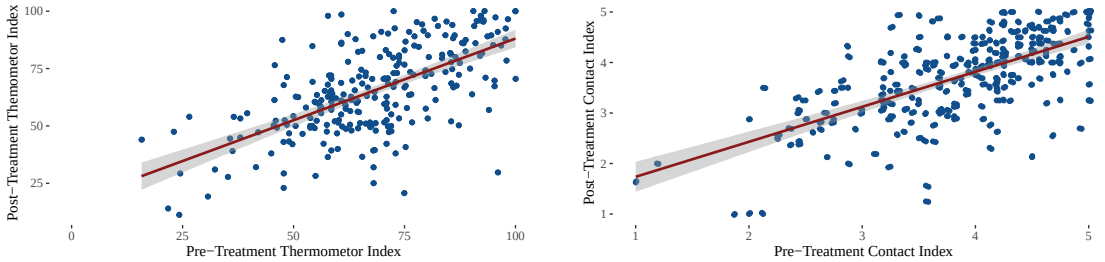
A3.4 Measuring Prejudicial Attitudes and Behaviors

As part of our field experiments, we collected a range of attitudinal and behavioral measures of prejudice. As depicted in Table A1, our primary measures of prejudicial attitudes include a feeling thermometer index, a diversity index, and an index of contact intentions based on students' responses to our surveys. Our primary behavioral outcomes include registration for an intergroup contact event and take-up of a pro-diversity (rather than self-affirming) wristband as compensation for participating in our survey. Prejudicial attitudes were measured in both our pre- and post-treatment surveys, and behavioral measures of prejudice were collected only in the post-treatment period.

Importantly, our attitudinal and behavioral measures are closely linked. Specifically, our behavioral outcome measuring registration for contact was designed to complement our attitudinal measure of intention for contact. Similarly, our behavioral outcome measuring take-up of a pro-diversity bracelet was designed to complement our attitudinal support for diversity index.

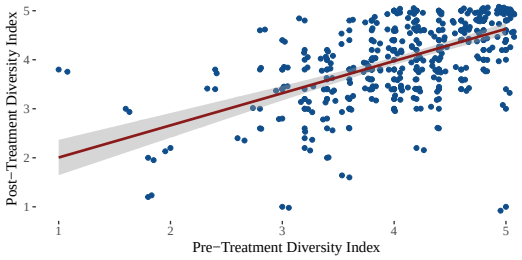
In this section, we take a closer descriptive look at our primary outcomes in order to contextualize the key results of our paper. First, in Figure A7, we provide evidence in support of the test re-test reliability of our attitudinal prejudice measures. Specifically, focusing on control group students (who were not directly influenced by our intervention), we show that there is a strong positive correlation in our attitudinal measures of prejudice over time. This suggests that our main attitudinal measures of prejudice have strong test-retest reliability.

Second, in Figure A8, we report the correlation of our key attitudinal measures with our behavioral measures of prejudice measured during the same survey wave. We find that our contact intention index, as well as our other main attitudinal measures of prejudice, have a positive and statistically significant correlation with the behavioral measure of registration for an intergroup contact event. Similarly, we find that our attitudinal pro-diversity index, as well as our thermometer index, have a positive and statistically significant correlation with the take-



(a) Thermometers Index

(b) Contact Intention Index



(c) Diversity Attitudes Index

Figure A7: **Attitudinal measures of prejudice are strongly correlated over time for control group students.** Each plot reports the bivariate correlation of our main outcome measures for the control group in the pre-and (first) post-treatment period. The strong correlation in outcomes over time suggests that our measures of prejudice have a high degree of reliability.

up of a pro-diversity bracelet. While the correlation of contact intentions with take-up of a pro-diversity bracelet is positive, this correlation is imprecisely estimated.

The results reported in Figure A8 emphasize the construct validity of our key behavioral and attitudinal measures, suggesting that they are capturing interrelated dimensions of intergroup prejudice. However, one might wonder, if our attitudinal and behavioral measures of prejudice are strongly correlated, why does our intervention yield mixed results in terms of behavioral outcomes? In other words, what explains the positive effects of our intervention on our main attitudinal outcomes and pro-diversity wristband take-up, and the null effects on registration for an intergroup contact event? We explore one potential explanation relating to ceiling effects.

In Figure A9, we plot the distribution of our key behavioral measures. Since we did not

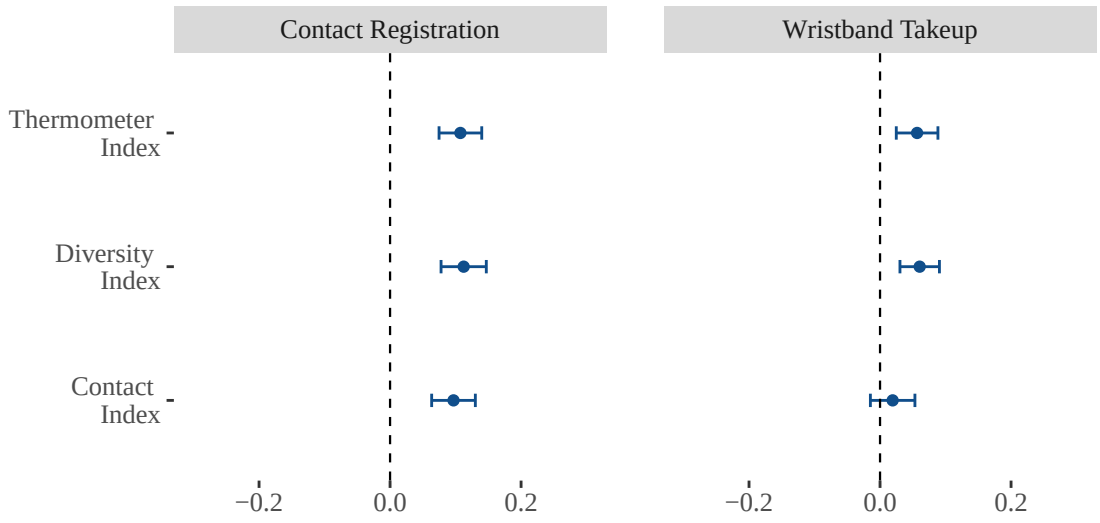
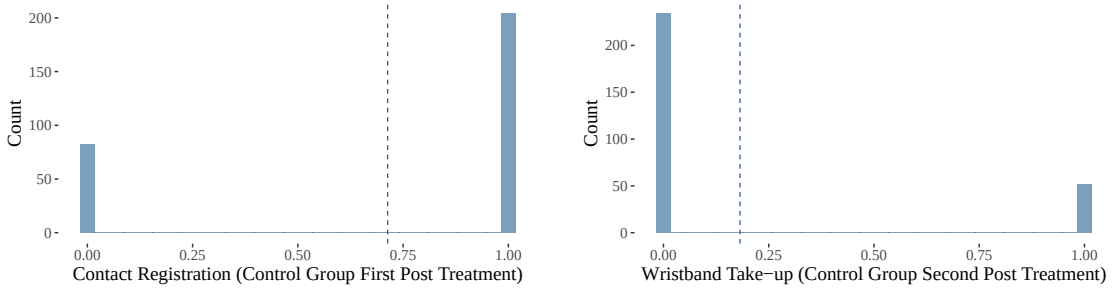


Figure A8: **Correlation of attitudinal and behavioral measures of prejudice**. In this Figure, we plot the bivariate correlation between students' self-reported attitudinal prejudice and their revealed preference prejudice.

collect pre-treatment behavioral measures, we report the control group's post-treatment behavioral measures of registration for contact and take-up of pro-diversity bracelets. Doing so, an interesting pattern emerges with regard to registration for contact: even without exposure to treatment, a large majority of students register for our proposed future intergroup contact event. Possibly, the high rates of registration could relate to the fact that registration for a potential future event is a rather costless behavior.

Regardless, we cautiously interpret the distribution in panel a) of Figure A9, as suggestive evidence of a ceiling effect, implying that most students are interested in registering for a contact intervention at baseline (i.e., without direct exposure to our intervention). Thus the potential effect of our intervention on this outcome of interest is limited from the start because a majority of students are already interested in registering for contact even without intervention. In contrast, our second behavioral outcome, which measures the take-up of a pro-diversity wristband, has substantially lower take-up rates at baseline and yields positive and precisely

estimated effects.



(a) Contact Registration Study 2

(b) Wristband Take-up Study 2

Figure A9: **Distribution of behavioral measures in Study 2.** Panel a) reports the distribution of study 2 control group students who registered for contact in the first post-treatment period. Panel b) reports the distribution of study 2 control group students who decided to take a pro-diversity bracelet as compensation for participation in our surveys in the second post-treatment period.

A3.5 Spillovers and General Awareness to Intervention

One design concern with school-based interventions relates to spillovers that can result from children in treatment classes sharing the content of the intervention with children in control classes. Such a dynamic could introduce downward bias to our estimates since control-group students would also be influenced by the content of the intervention and potentially report lower levels of prejudice in post-treatment surveys. Though our study is not set up to detect such spillover effects, we make use of an open-ended post-treatment survey item to explore whether and how informed students' are about our interventions.

Specifically, in Study 2, any student who noted that they did not watch episodes of the show "You Can't Ask That" in their classroom over the past 3 months was asked the following question: "*In the past weeks, several classes participated in activities as part of a social research project. Do you know what these activities entailed? If so, please describe the activities below.*" Most students did not respond to this question, implicitly implying that they do not know

about research-related activities. However, we coded the small number of open-ended responses provided by students into several general categories presented in Figure A10.

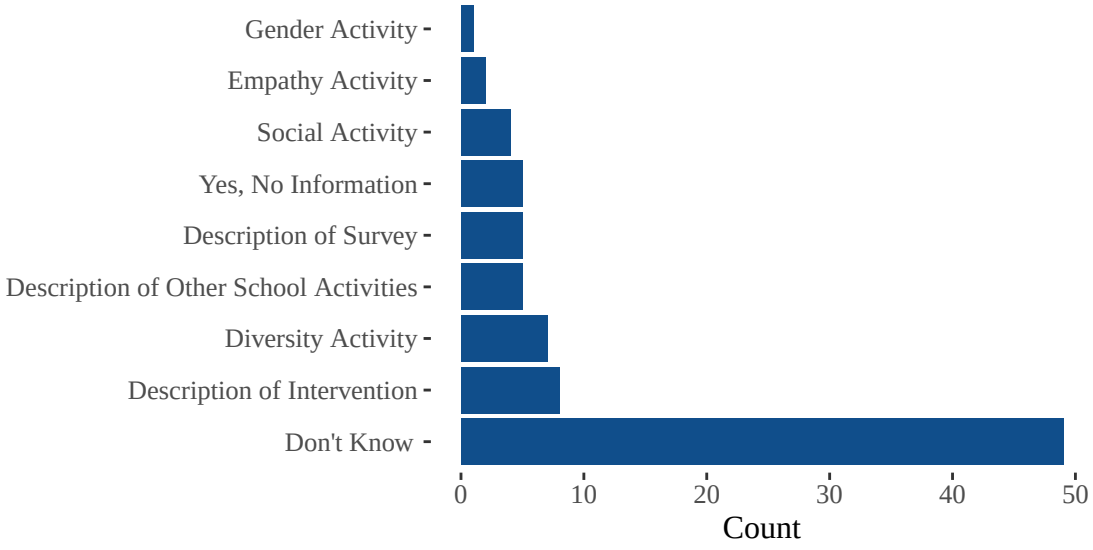


Figure A10: **Responses to a post-treatment open-ended survey item eliciting information about research-related activities in school in Study 2.** At the end of our post-treatment surveys, we asked students whether they watched the show "You Can't Ask That" in class during the past 3 months. Students who did not report watching the show were then informed that in recent months some classes participated in activities as part of a social research project and were asked to share anything they know about these activities. We coded students' open-ended responses into different categories and report category frequencies in our sample. Note that in Study 2, only 81 students provided an actual text response to this question.

While suggestive in nature, the pattern of responses reported in Figure A10 suggests that spillovers were unlikely prevalent as our intervention rolled out in the field. A majority of students did not share information about research-related activities in the post-treatment survey, and to the extent they did, very few (less than 10 students) described our intervention as a research-related activity in the post-treatment survey. These descriptive patterns serve to reduce concerns regarding spillovers that would likely attenuate our main average treatment effects reported in the paper. However, we encourage future researchers who evaluate school-based interventions to elicit students' pre-treatment social network structure and employ suitable spillover

designs (e.g. (*1*)) that would allow for testing the direct and indirect effects of prejudice reduction interventions.

A4 Study 1 Additional Analyses

A4.1 Study 1: Descriptive Statistics

In Table A2, we report descriptive statistics relating to students gender, age, and grade. In Table A3, we further consider a balance check on pre-treatment demographics and key attitudes. As depicted in Table A3, we can not reject the null hypothesis for any pre-treatment variable. More importantly, for the overall balance test reported in Table A3, we can not reject the null hypothesis of similarity, providing further assurance that our treatment and control groups are similar on observables and unobservables.

Table A2: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Boy	270	0.493	0.501	0	1
Age	270	10.389	0.941	9	12
Grade 4	270	0.341	0.475	0	1
Grade 5	270	0.359	0.481	0	1
Grade 6	270	0.300	0.459	0	1

A4.2 Study 1: Robustness Checks

A4.2.1 Study 1: Attrition

In study 1, 17 students participated in our baseline survey but were unavailable to participate in our post-treatment survey. Moreover, as further discussed in Section A4.4, due to a technical

Table A3: Balance Table (Pre-Treatment Measures) - Study 1

	Control	Treatment	Std. Diff.	Adj. Diff.	Pooled SD	z
Boy	0.44	0.54	0.21	0.10	0.50	1.82
Age	10.32	10.46	0.14	0.13	0.94	-0.02
Thermometer Scale	58.77	57.37	-0.07	-1.41	21.63	-0.29
Thermometer Arab	46.25	47.33	0.04	1.08	29.56	0.05
Thermometer Immigrant	56.03	57.31	0.05	1.28	27.60	0.05
Thermometer UO	56.24	49.50	-0.22	-6.74	29.99	-0.94
Thermometer Blind	76.57	75.34	-0.05	-1.24	24.26	-0.27
Similarity Scale	2.42	2.29	-0.14	-0.14	1.00	-0.56
Similarity UO	2.40	2.10	-0.23	-0.30	1.28	-1.01
Similarity Arab	2.33	2.14	-0.15	-0.19	1.27	-0.63
Similarity Immigrant	2.51	2.59	0.06	0.08	1.30	0.13
Similarity Blind	2.46	2.32	-0.11	-0.14	1.25	-0.66
Diversity Scale	3.95	3.84	-0.15	-0.12	0.78	-0.42
Contact Scale	3.49	3.43	-0.08	-0.06	0.73	-0.26

error, all students were randomly exposed to four out of five batteries of questions relating to intentions for intergroup contact. In other words, all students had one battery of intention for contact with a specific social group they did not get a chance to report. Finally, though encouraged to complete the full surveys, some students skipped specific items.

To reduce concerns regarding bias in our estimates as a result of attrition, in Figure A11, we report point estimates from a series of regressions diagnosing the correlates of attrition. Specifically, we regress a binary indicator taking the value of 1 if a respondent is missing a given post-treatment outcome over two key indicators: A treatment indicator and a pre-treatment measure of the missing outcome. Across all models on the right-hand panel of Figure A11, we do not find any evidence that attrition correlates with treatment. Similarly, attrition does not correlate with pre-treatment measures of outcomes, with the exception of the non-response to the Ultra-Orthodox contact intention item, which is negatively related to pre-treatment intentions for contact with Ultra-Orthodox children. On the whole, the finding in Figure A11 reduces concerns that attrition threatens the internal validity and unbiasedness of our main estimates in

Study 1.

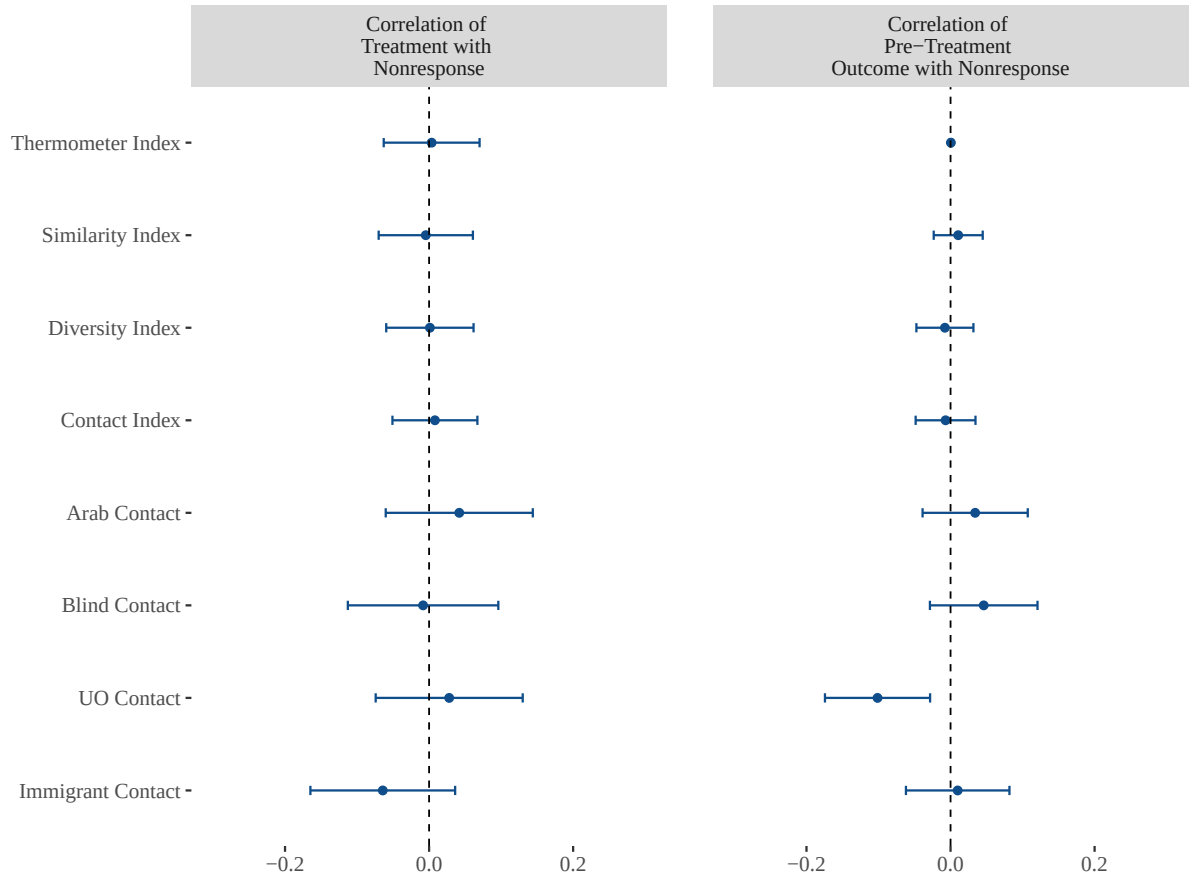


Figure A11: **Non-Response survey items in study 1 do not consistently correlate with treatment or pre-treatment attitudes.** This figure reports point estimates and 95% confidence intervals representing the correlation of non-response to a given survey item with respondents' treatment assignment status and pre-treatment measure of outcome.

A4.2.2 Study 1: Alternative Specifications

In the main text, we address the modest number of clusters in our data from Study 1 by employing a wild cluster bootstrap procedure to cluster our standard errors (*2*). However, in this section, we implement additional analyses that employ randomization inference to address concerns regarding the modest number of clusters in Study 1 (*3*).

We report randomization inference results for our main outcomes from Study 1 in Table A4. The average treatment effect on our thermometer and similarity outcomes are robust to this specification. However, when employing randomization inference, the average treatment effect on our contact intention and support for diversity scales are imprecisely estimated. We interpret this pattern as providing somewhat mixed results regarding the robustness of our findings to alternative specifications and emphasize the importance of further testing the robustness of these patterns on large samples with more clusters, as we do in Study 2.

Table A4: Randomization Inference - Main Outcomes

	Term	Estimate	p.value
1	Thermometers	0.37	0.01
2	Contact Intensions	0.21	0.21
3	Diversity Attitudes	0.22	0.18
4	Similarity Perceptions	0.35	0.08
5	Register Contact	0.13	0.10

In the main text, we employ index outcomes and report the average treatment effects of our intervention on general attitudes toward multiple outgroups rather than particular attitudes toward specific social groups. In Figure A12 we consider the effects of our intervention on group-specific measures relating to intergroup affect and perceptions of intergroup similarity. These additional analyses provide some interesting insights.[2]

When focusing on intergroup affect, it appears that the treatment had substantial effects on attitudes towards Arabs and immigrants, more moderate and precisely estimated effects on attitudes towards Ultra-Orthodox children who were not discussed in the intervention, and no effect on students' affect toward visually impaired students. Somewhat similar patterns emerge with regard to our measure of intergroup similarity. However, in this case, the measure for Ultra-Orthodox children is imprecisely estimated, but the effects of treatment on perceptions of

[2]Note that we do not consider group-specific effects relating to intentions for intergroup contact due to the technical limitation in our contact measures described in Section A4.4.

intergroup similarity with visually impaired children are large and precisely estimated.

We argue that the substantively small and imprecisely estimated treatment effect of the intervention on attitudes towards visually impaired children is driven by ceiling effects. In other words, as we show in Figure A1, in both our studies, students have very high levels of affect towards visually impaired students. However, they still report high levels of dissimilarity perceptions. We suggest that given the high levels of pre-treatment affect toward visually impaired children, treated students have less room to move with regards to how warm they feel towards the group in question. We argue that this could plausibly explain the null effect we identify in Figure A12.

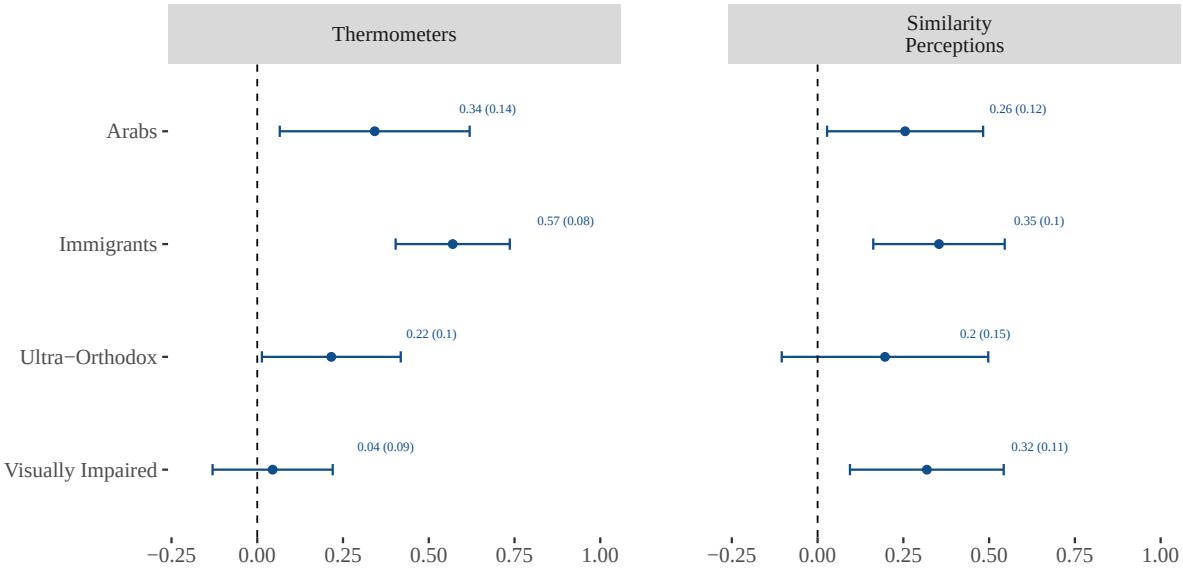


Figure A12: **Exposure to the intervention in Study 1 improved children's attitudes towards most social groups.** This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on students' attitudes towards specific social groups 1-2 weeks post-treatment. Point estimates and standard errors (in parentheses) are reported along each estimate.

Our experimental design, in which we collected data about students' pre-treatment prejudice, allows us to consider whether our identified average treatment effects are driven by stu-

dents with low (or high) levels of prejudice. To do so, we consider how our pre-treatment thermometer index moderates the average treatment effect of our intervention on the post-treatment thermometer index. In Figure A13, we first diagnose our data to ensure that we are set up for credibly estimating an interaction model. After doing so, we report a marginal effect plot with a binning estimator proposed by Hainmueller et al. (*4*) in Figure A14.

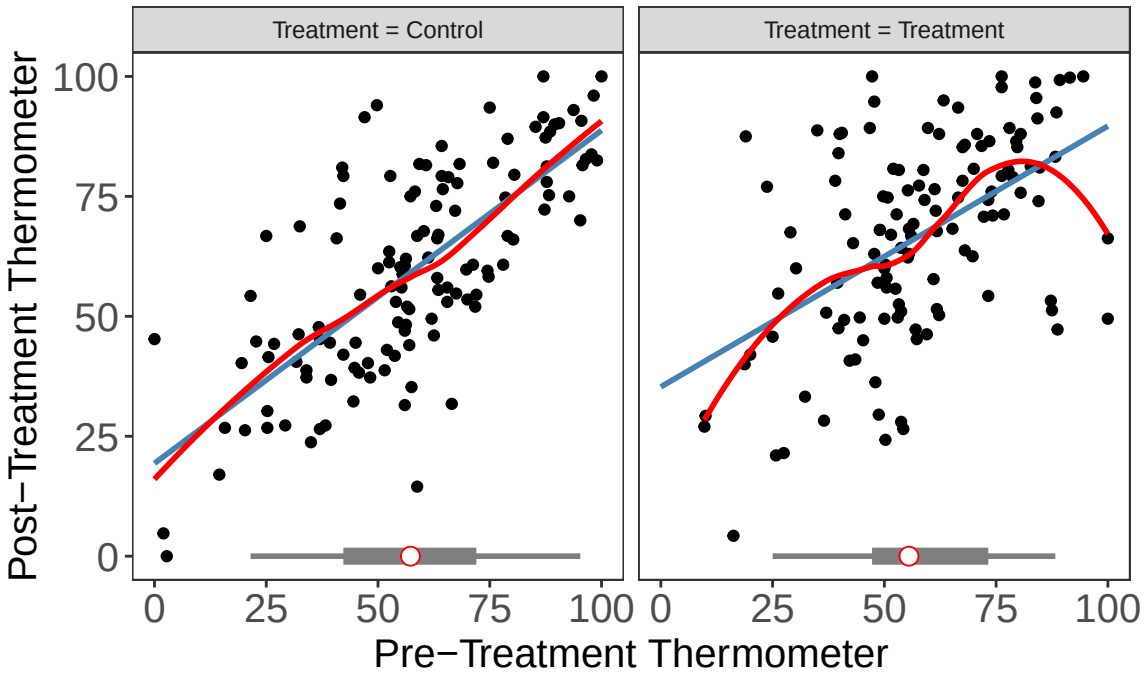


Figure A13: **Interaction term diagnostics Study 1.** This Figure plots the post-treatment thermometer index over the pre-treatment thermometer index by treatment condition.

While the pattern in Figure A14 shows that the moderating effect of treatment is smaller for less-prejudicial respondents, this pattern is imprecisely estimated. Indeed, there are no statistically distinguishable differences between respondents with low, medium, or high levels of intergroup affect (as categorized by the binning procedure proposed by Hainmueller et al. (*4*)). We thus interpret the evidence in Figure A14 to suggest that both prejudicial and non-

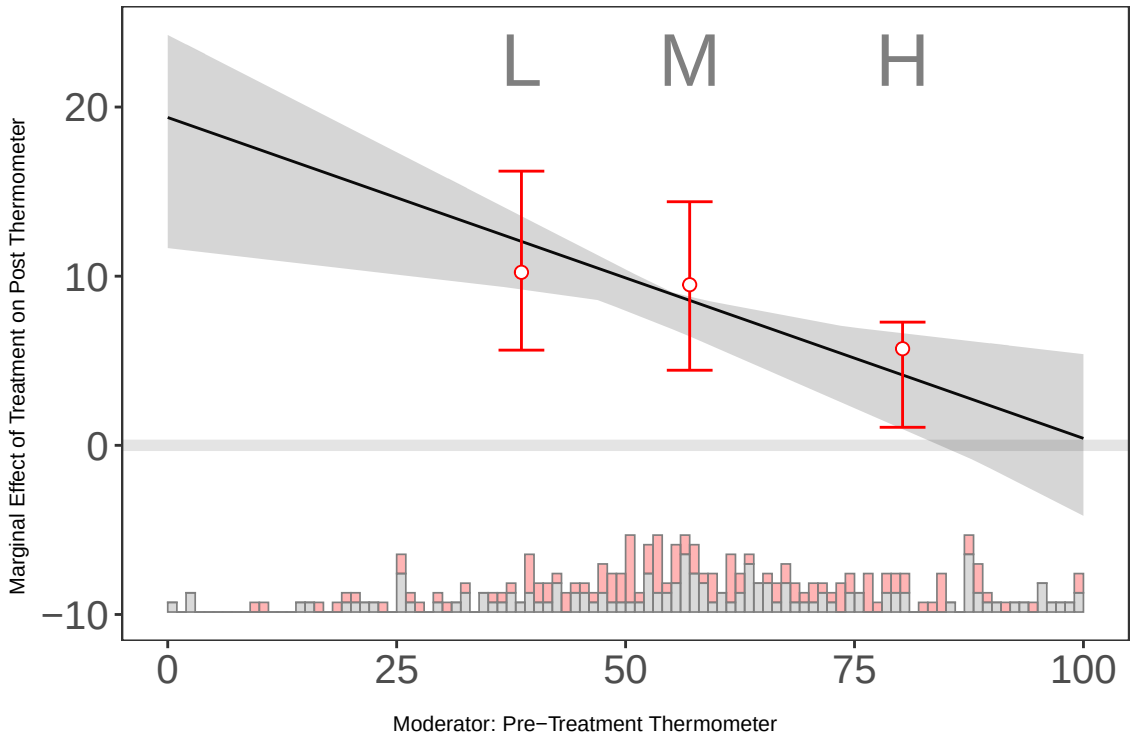


Figure A14: **In Study 1, pre-treatment attitudes do not moderate average treatment effects.** This plot reports that the average treatment effect of our intervention conditional on levels of our pre-treatment thermometer index using the binning estimator proposed by Hainmueller et al. (*4*).

prejudicial students in Study 1 react similarly to treatment. To the extent that moderation exists, we are likely underpowered to detect it. However, similar patterns of non-moderation arise in Study 2 when we focus on a substantively larger sample.

A4.3 Attitudes towards Arabs in the Shadow of Conflict

As indicated in the main text, intense missile fires and inter-communal clashes between May 10-21, 2021, disrupted life in many cities across Israel, including our intervention site. One might expect that such events that unfolded during our intervention may have shaped students' prejudice, and specifically their attitudes towards Arab children. In this section, we provide

suggestive evidence to assess this possibility.

To examine patterns of prejudice towards Arab children and their sensitivity to conflict dynamics, we created a scale based on our Arab thermometer, similarity, and contact intention questions ($\mu = 0$ and $\sigma^2 = 1$), which were all measured pre-and post-treatment. Higher values on the scale indicate more positive attitudes toward Arabs. In Figure A15, we plot the pre-and post-treatment means for treated and control students.

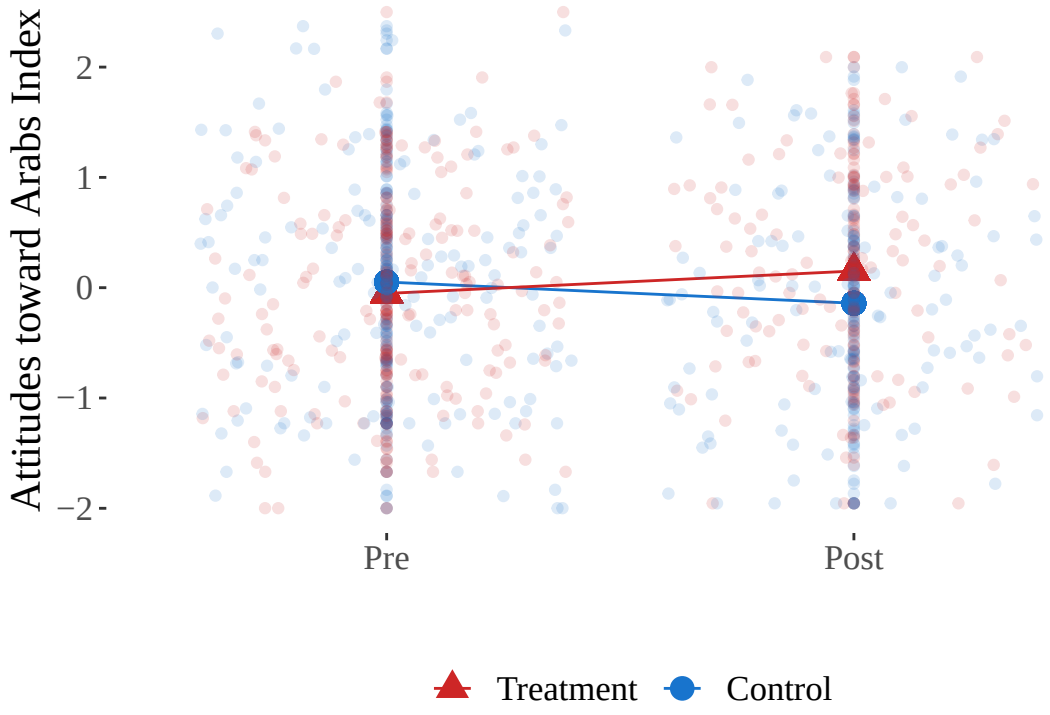


Figure A15: **Attitudes toward Arabs improve (deteriorate) among treated (control) students over time in Study 1.** This figure reports students' overall attitudes toward Arabs before and after the intervention by treatment status.

Interestingly, we find that in the pre-treatment period, both groups have similar average attitudes towards Arabs. However, in the post-treatment period, these attitudes diverge for treatment and control students. Indeed, attitudes towards Arabs become more positive amongst students in the treatment group. In contrast, attitudes towards Arabs become more negative

amongst students in the control group. We interpret the patterns in Figure A15 as suggestive evidence informing us about the potential promise of prejudice reduction interventions in buffering the deterioration of intergroup attitudes during cycles of violence. Indeed, it appears that education programs that facilitate vicarious contact and constructively address sensitive topics at the heart of intergroup relations can increase positive attitudes towards outgroups even at times of intensifying conflict. and hinder the deterioration of intergroup attitudes.

A4.4 Deviation from Pre-Analysis Plan

In analyzing study 1, we take four deviations from our pre-analysis plan. First, though technically identical, we control for pre-treatment covariates on the right-hand side of the regression instead of estimating treatment effects on a first difference of the pre-post outcome. Second, due to a technical error in our Qualtrics surveys, each student responded to three of four contact intention batteries. Thus, students reported their intention to interact with three of our four key outgroups. For that reason, rather than considering intention for contact with specific social groups, we consider an overall index of intention for contact with social groups throughout our analyses. For all students, this index is comprised of responses to three batteries relating to three randomly selected social groups. As reported in Figure A11, since the presentation of batteries to each student was randomly assigned by Qualtrics, missingness is not correlated with treatment or pre-treatment levels of the outcome (with the exception of a single measure where missingness is not correlated with treatment but is correlated with the pre-treatment outcome).

Third, we intended to consider the effects of our intervention on attitudes towards a made-up minimal group (which we described as the "Supza group" in our surveys), and we measured a behavioral measure by asking students to list groups they might want to see in future iterations of the show. However, while fielding our surveys in Study 1, we realized that the concept of a minimal group and the behavioral question confused children. For that reason, we do

not consider this outcome in our intervention. Finally, given the high α Cronbach of our five support for diversity measures, and to maintain consistency with the analyses in Study 2, we aggregate all our diversity measures into a single index rather than considering the effects of our intervention on two different outcomes relating to students' appreciation for diversity (H4a in the pre-analysis plan) and support for diversity (H4b in the pre-analysis plan).

A5 Study 2 Additional Analyses

A5.1 Study 2: Descriptive Statistics

Our second study focused on 767 students in 5 schools located in central Israel. We provide descriptive statistics of our sample in Table A5 focusing on students' gender, age, and grade. In Table A6, we further report balance tests considering respondents' demographics and pre-treatment levels of prejudice. Our balance tests suggest that treatment and control groups are similar on observables.

Table A5: Descriptive Statistics - Study II

Statistic	N	Mean	St. Dev.	Min	Max
Boy	767	0.503	0.500	0	1
Age	767	10.126	0.912	9	12
Grade 4	767	0.026	0.159	0	1
Grade 5	767	0.035	0.184	0	1
Grade 6	767	0.018	0.134	0	1

Table A6: Balance Table (Pre-Treatment Measures) - Study 2

	Control	Treatment	Std. Diff.	Adj. Diff.	Pooled SD	z
Boy	0.51	0.50	-0.01	-0.01	0.50	-0.16
Age	10.20	10.08	-0.13	-0.12	0.91	-0.19
Thermometer Index	67.39	63.74	-0.21	-3.65	17.56	-0.48
Thermometer Arab	63.06	58.78	-0.17	-4.27	24.43	-0.57
Thermometer Immigrant	65.42	61.73	-0.17	-3.69	22.30	-0.50
Thermometers UO	60.13	58.36	-0.07	-1.77	23.92	-0.29
Thermometer Blind	80.98	76.10	-0.25	-4.87	19.26	-0.53
Contact Intentions Index	3.83	3.81	-0.03	-0.02	0.77	-0.14
Contact Arab	3.70	3.63	-0.08	-0.08	0.95	-0.24
Contact Immigrant	3.83	3.76	-0.09	-0.07	0.86	-0.24
Contact UO	3.68	3.68	0.00	0.00	0.91	-0.09
Contact Blind	4.11	4.16	0.06	0.05	0.79	-0.01
Group Similarity Index	2.56	2.56	0.00	0.00	0.80	-0.09
Similarity Arabs	2.67	2.67	0.00	0.00	1.10	-0.09
Similarity UO	2.28	2.33	0.06	0.06	1.00	0.07
Similarity Blind	2.57	2.48	-0.09	-0.10	1.06	-0.35
Similarity Immigrant	2.73	2.78	0.04	0.04	1.05	0.01
Group Heterogeneity Index	2.99	2.99	0.00	0.00	0.73	-0.10
Heterogeneity Arabs	3.00	2.98	-0.02	-0.02	0.92	-0.14
Heterogeneity Immigrants	3.12	3.23	0.13	0.12	0.90	0.16
Heterogeneity Blind	3.03	2.96	-0.08	-0.08	0.96	-0.28
Heterogeneity UO	2.81	2.80	-0.02	-0.02	0.97	-0.14
Perspective Taking Index	3.54	3.47	-0.09	-0.07	0.81	-0.23
Perspective Taking Arabs	3.43	3.35	-0.07	-0.08	1.12	-0.25
Perspective Taking Immigrants	3.44	3.34	-0.10	-0.10	1.02	-0.30
Perspective Taking UO	3.29	3.17	-0.11	-0.13	1.11	-0.36
Perspective Taking Blind	3.99	4.01	0.03	0.03	0.86	-0.05
Diversity Attitudes Index	4.11	4.03	-0.11	-0.08	0.69	-0.23

A5.2 Study 2: Robustness Checks

A5.2.1 Study 2: Attrition

In Study 2, one of our five schools did not participate in the second post-treatment survey because their treatment rollout was severely delayed, and thus the second data collection wave coincided with the summer break. Notably, since classes were blocked to treatment and control conditions at the school-grade level, the omission of a given school from the final survey wave resulted in attrition amongst treated and controlled students and does not pose a threat to internal validity. Beyond this component of attrition, in any given post-treatment wave, we have a minority of survey respondents who were not sampled or did not respond to specific survey items.

To minimize concerns regarding selective attrition that could bias our estimates, in Figure A19, we show that both treatment and pre-treatment levels of prejudice do not predict attrition. To do so, we employ our main specification from study 2 and set the outcome as an indicator taking the value of 1 if a respondent is missing a response to a given item (0 otherwise). For each of our main outcomes, we regress this attrition measure over our treatment and pre-treatment measure of the outcome under investigation.

The results in Figure A19 reduce concerns regarding selective attrition. In all but one model on the left panel of Figure A19, treatment is not associated with attrition, reducing concerns regarding the internal validity of our estimates. Moreover, in all but one model on the right-hand side of Figure A19, pre-treatment measures of an outcome do not predict attrition. We interpret the small and insignificant point estimates reported in Figure A19 to suggest that attrition does not pose a threat to inference in our case and that attrition is not correlated with important pre-treatment measures.

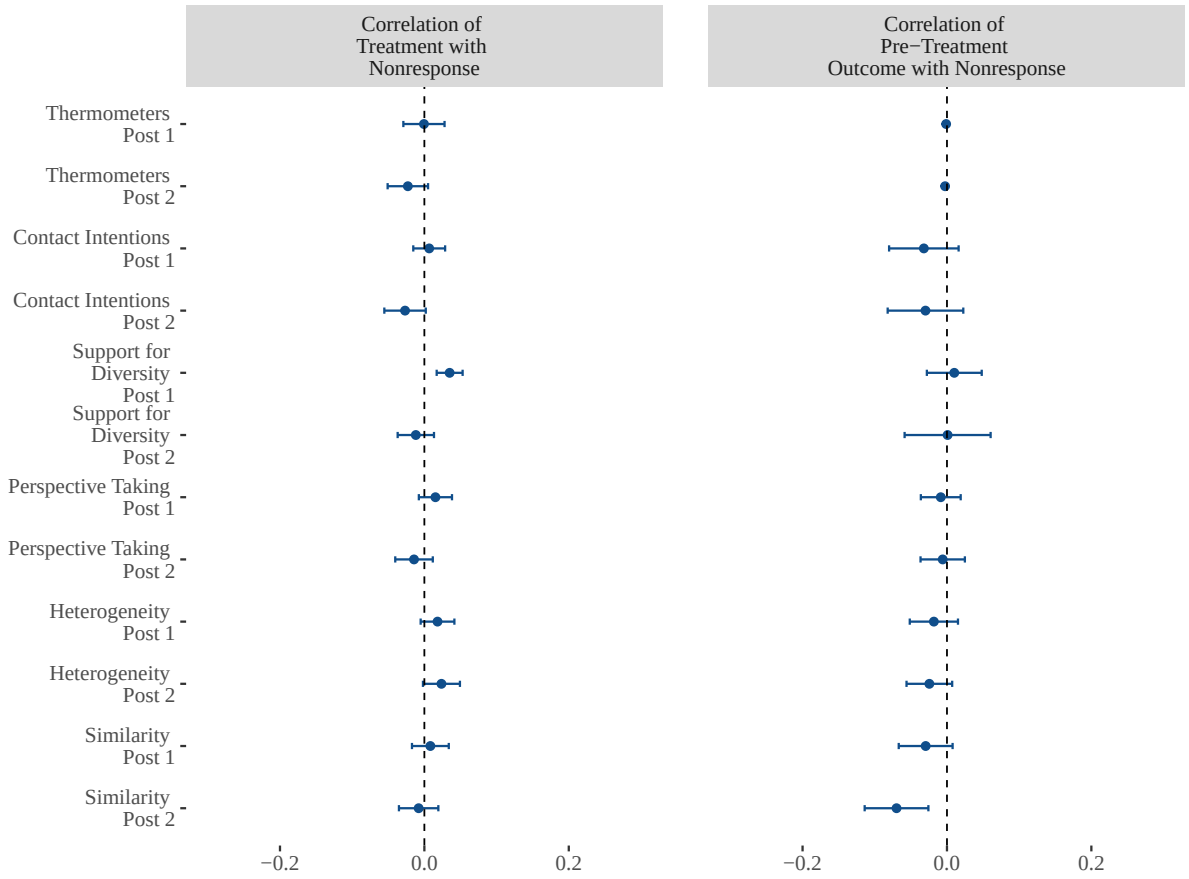


Figure A16: **Non-Response to survey items does not consistently correlate with treatment or pre-treatment attitudes in Study 2.** This figure reports point estimates and 95% confidence intervals representing the correlation of non-response for a given survey item with respondents' treatment assignment status and pre-treatment measure of outcome.

A5.2.2 Study 2: Alternative Specifications

In the main text, we report changes in intergroup affect and intentions for contact on aggregate scales. Both scales we employ are highly consistent ($\alpha_{thermometer} = 0.79$ and $\alpha_{contact} = 0.9$). Moreover, we focus on aggregate scales because our intervention was designed to shape students' attitudes towards outgroups as a whole, including outgroups not mentioned in the intervention rather than a specific social group.

However, in Figure A17, we report additional models based on our main specification, in which we consider students' group-specific changes in prejudice with regards to our contact intention and thermometer indices. We find that, for the most part, our intervention affected students' affect and contact intentions with varying social groups in a consistent fashion. The one exception to this pattern relates to students' attitudes towards visually impaired children. Treatment effects on affect and intention for contact with visually impaired children are small and imprecisely estimated in the first post-treatment wave. However, these effects are larger and precisely estimated in the second wave post-treatment. As we argue above, shaping attitudes towards visually impaired children in this context faces a challenge of ceiling effects. This might explain why we recover small point estimates in Figure A17, which are harder to estimate with precision.

In Figure A18 we further report disaggregated effects on the components of our support for diversity scale. We find that with the exception of a single post-treatment effect that is estimated with very high uncertainty, all disaggregated effects are positive and for the most part, precisely estimated. However, as emphasized in our pre-analyses plan, we combine these measures due to their high degree of consistency ($\alpha = 0.76$) to reduce measurement error.

In Figure A19 we further examine the average treatment effects of our intervention on disaggregated measures of our potential psychological mechanisms: perspective getting, perceptions of intergroup similarity, and perceptions of within-group heterogeneity. We find that our treatment increased students' willingness to take the perspective of Arab, immigrant, and Ultra-Orthodox children. However, the effects on the visually impaired perspective-taking measure are very small and imprecisely estimated. We suggest that the null effect on the visually impaired perspective-taking item may be driven by ceiling effects. Indeed, as we show in Figure A20 at the pre-treatment period, students' agreement that it is important to take their outgroup perspective is substantially higher when the target of perspective taking is visually

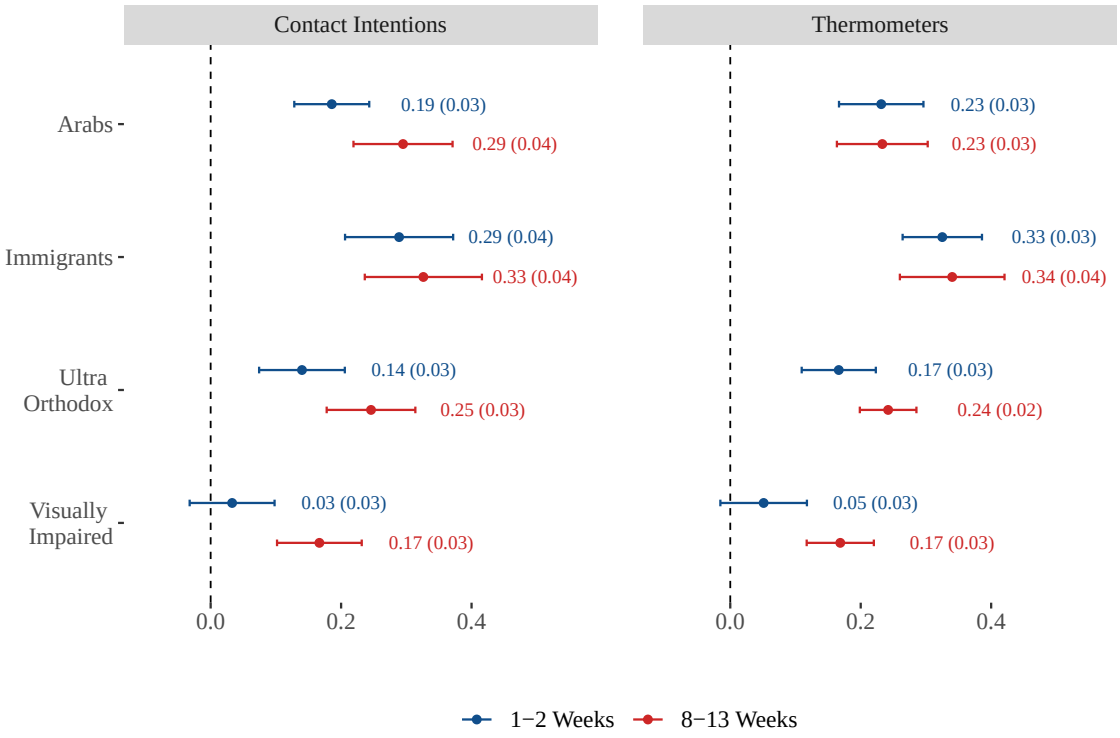


Figure A17: **The Intervention had significant effects on attitudes towards all social groups with the exception of visually impaired children.** This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on disaggregated components of the contact and thermometer indices. Point estimates and standard errors (in parentheses) are reported along each estimate.

impaired children when compared with other children.

With regards to group similarity, we find that, for the most part, the intervention had small positive and imprecisely estimated positive effects on students' belief that they are similar to different outgroups. In contrast, to the rather consistent pattern of perceptions of intergroup similarity, our measure of within-group heterogeneity appears to be far less consistent. Indeed, in the first post-treatment survey, it appears that the intervention increased students' perceptions of within-group heterogeneity with regard to Arab and Ultra-Orthodox children. However, in the second post-treatment survey, the effect with regard to Arab children flips and appears to be negative, implying that treatment reduced perceptions about Arab outgroup heterogeneity.

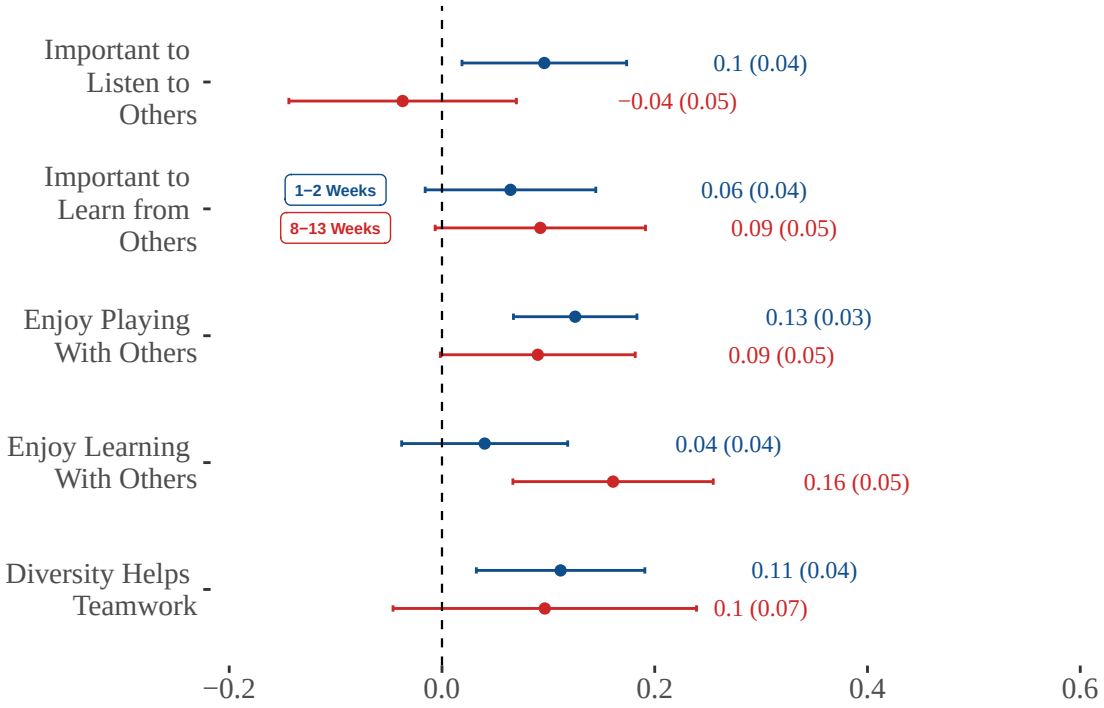


Figure A18: **The Intervention affected a majority of disaggregated support for diversity measures.** This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on disaggregated components of the diversity scales. Point estimates and standard errors (in parentheses) are reported along each estimate.

Moreover, it appears that our treatment reduced perceptions regarding within-group heterogeneity of immigrant children and had a small positive effect on students' perceptions of visually impaired outgroup heterogeneity in the second survey wave but not in the first wave. Taken together, these mixed patterns explain the null result we identify in the main text and emphasize that it is very unlikely that perceptions of group heterogeneity are the central mechanism underlying the main effect of our intervention.

Like in our analyses of Study 1, we consider whether our treatment effects are driven by larger shifts among students with higher (or lower) levels of pre-treatment prejudice. To do so, we consider how the pre-treatment thermometer index moderates the effects of our intervention

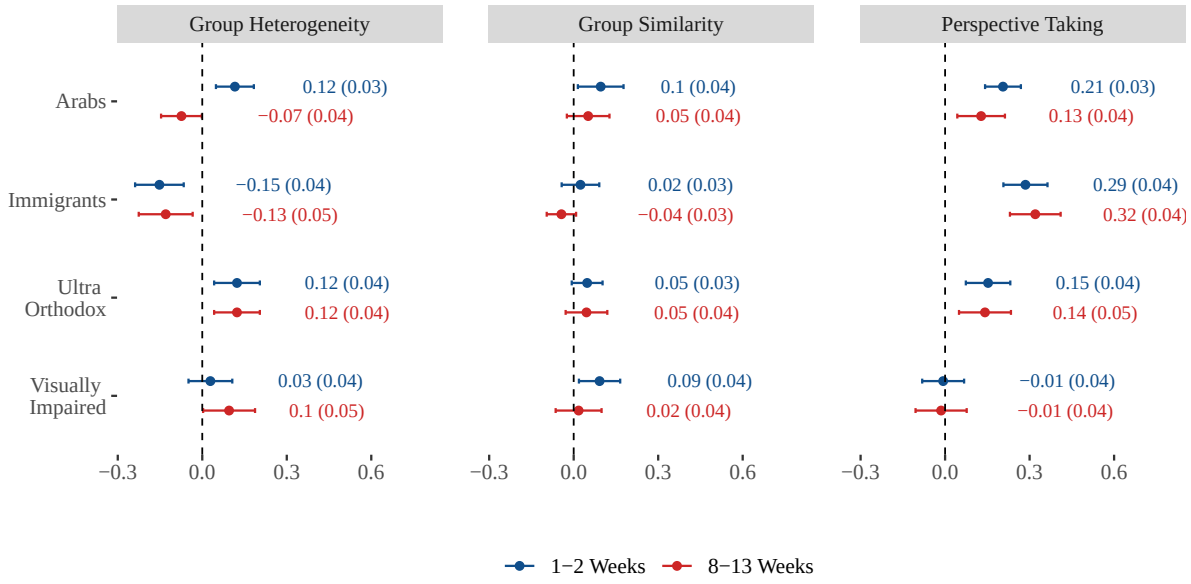


Figure A19: **The intervention had mixed effects on disaggregated components of the group similarity and group heterogeneity index, and consistent effects ofon the disaggregated components of the perspective-taking index.** This figure reports point estimates and 95% confidence intervals representing the effect of our intervention on disaggregated components of the mechanism indices. Point estimates and standard errors (in parentheses) are reported along each estimate.

on the post-treatment thermometer index. Before doing so, in Figure A21 we implement a diagnostic test proposed by Hainmueller et al. (*4*) and plot our post-treatment thermometer index over our pre-treatment thermometer index for both treated and controlled students. After doing so, we estimate the moderating effect of the pre-treatment thermometer index using the binning estimator proposed by Hainmueller et al. (*4*). We report these moderation analyses in Figure A22.

While treatment size seems to be negatively related to pre-treatment levels of prejudice, this relationship is not statistically significant. Indeed, the differences between low, medium, and high bins in Figure A22 are statistically indistinguishable. We thus interpret our evidence to suggest that treatment was similarly effective on students with varying levels of pre-treatment

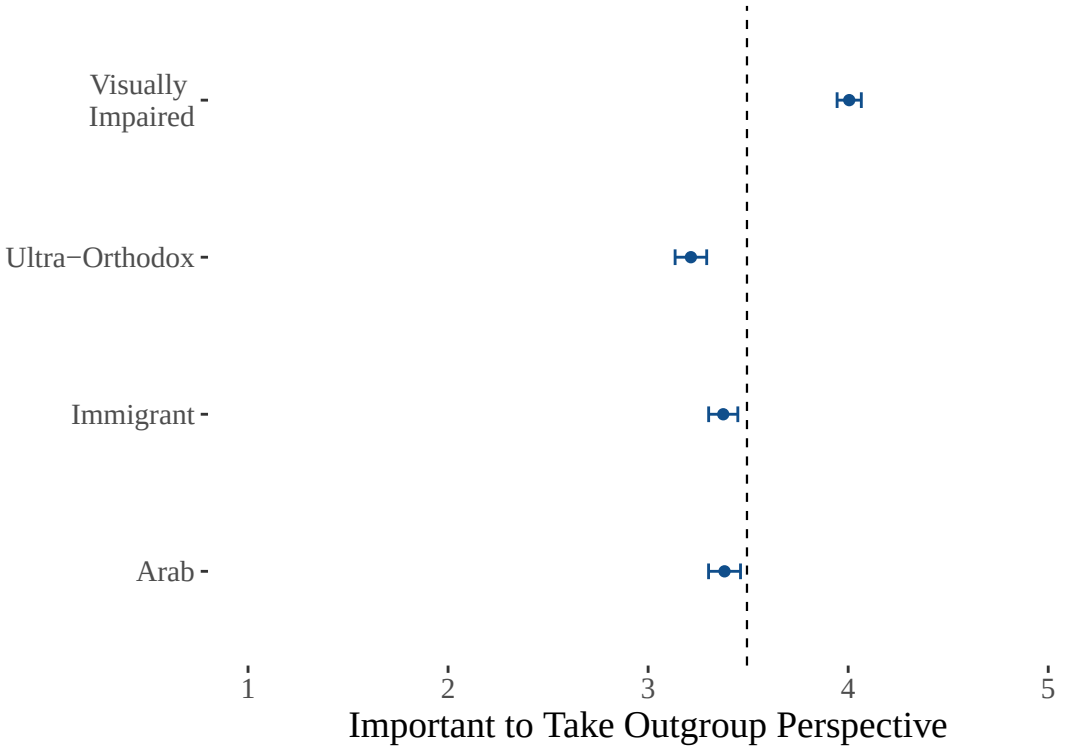


Figure A20: **Students' pre-treatment perspective-taking measures are substantively higher for visually impaired measures when compared with all other social group measures.** This figure reports means and 95% confidence intervals representing students' pre-treatment average agreement that it is important to take a specific outgroup's perspective. The dotted line represents the average value for all groups combined.

prejudice. To the extent to which heterogeneity in response to treatment exists, it is small and hard to estimate precisely with our current sample.

As we indicate in the main text, one of the five schools participating in our intervention was unable to join the final post-treatment wave. We emphasize above in Section A5.2.1 that this does not pose a threat to internal validity. However, one concern with our main results relates to the fact that it is hard to compare the first and second post-treatment waves because they focus on somewhat different samples. In other words, it might be that larger point estimates for a given outcome in the second wave are an artifact of variations in sample properties rather than overtime increases in the average treatment effects of our interventions.

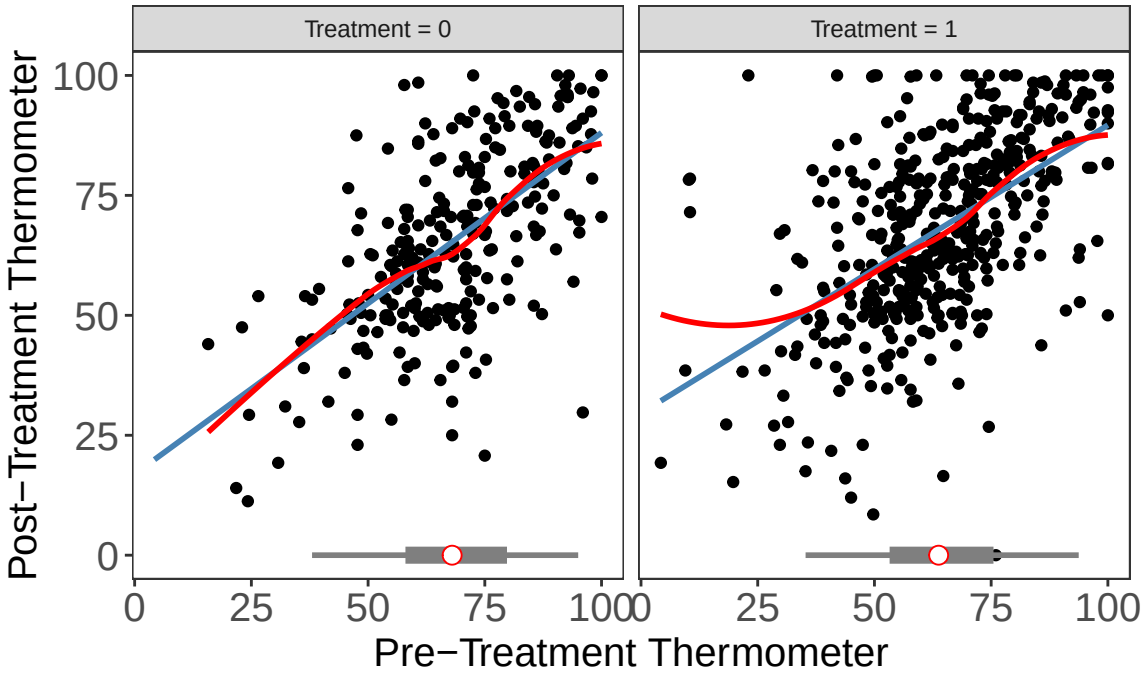


Figure A21: **Interaction term diagnostics for Study 2**. This Figure plots the post-treatment thermometer index over the pre-treatment thermometer index for treatment and control groups.

We address this concern in Figure A23-A24 by reestimating our main analyses focusing only on respondents who reported outcomes of interest in both our post-treatment waves. This allows us to hold the sample constant and further investigate differences in the magnitude of treatment effects in the first and second post-treatment surveys. The pattern of results remains similar in this analysis. Indeed, like in our main analyses, the point estimates for the effect of our intervention on the thermometer and contact indices are larger in the second post-treatment wave. Moreover, the effect on the diversity index is subtly smaller in the second post-treatment wave. Despite these consistent patterns, it is important to emphasize that point estimates are not statistically distinguishable from one another. Thus, we cannot point to any substantial changes in average treatment effects over time.

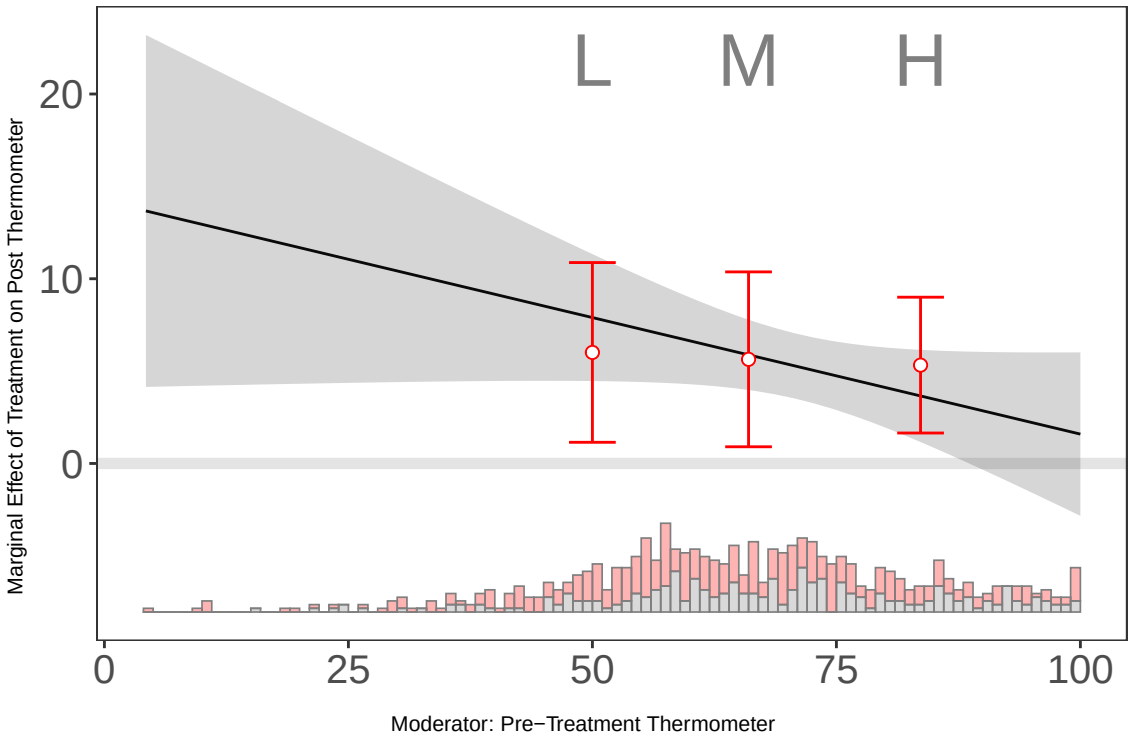


Figure A22: **In Study 2 pre-treatment attitudes do not moderate our interventions' average treatment effects.** This plot reports that the average treatment effect of our intervention conditional on levels of our pre-treatment thermometer index using the binning estimator proposed by Hainmueller et al. (*4*).

Finally, the analyses we pre-registered for Study 1 and Study 2 are subtly different. Specifically, in Study 2, we do not only control for pre-treatment covariates but also interact them with our treatment to increase precision (*5*). Moreover, in our main specification in Study 2, we do not employ a wild cluster bootstrap procedure because of our larger sample (and cluster) size.

However, to examine the robustness of our main pre-registered specification in Study 2 to alternative specifications employed in the paper, we estimate several additional models. First, in Table A7, we report randomization inference tests for our main findings from Study 2, using the same specification employed in Study 1 and reported in Table A4. The pattern of results reported in Table A7, based on randomization inference, is largely consistent with the main findings we

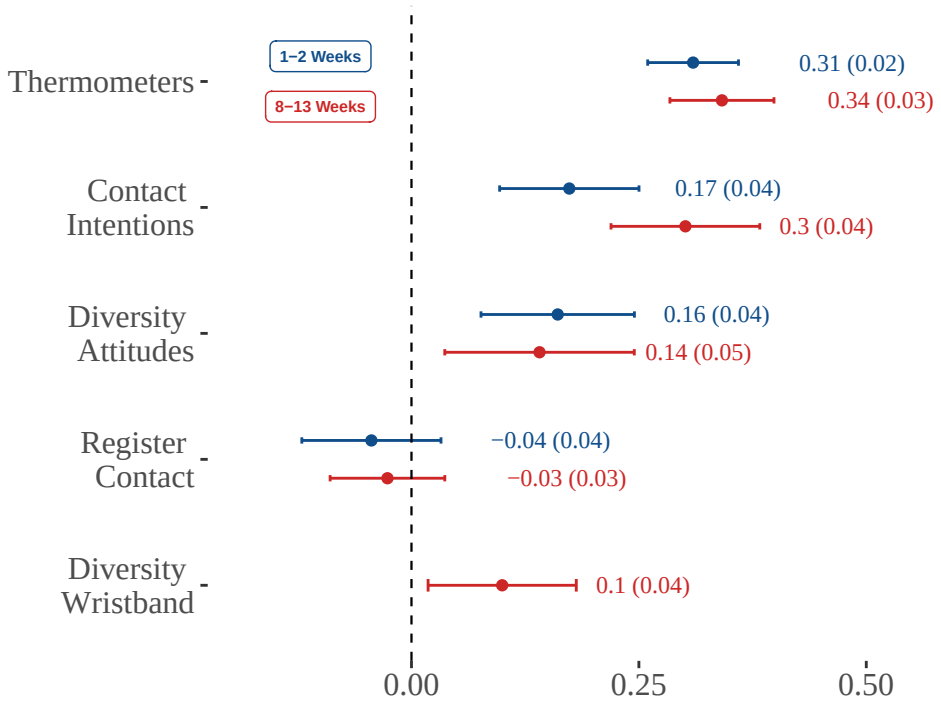


Figure A23: **The main pattern of results is consistent when focusing on outcomes only among respondents that participated in both post-treatment survey waves.** This figure reports point estimates and 95% confidence intervals representing the effect of our intervention on main outcomes amongst students participating in both post-treatment survey waves. Point estimates and standard errors (in parentheses) are reported along each estimate.

report in the paper using our pre-registered specification. One important difference relates to the result regarding the intergroup similarity mechanism in the first post-treatment wave, which is imprecisely estimated when employing randomization inference (but is precisely estimated in our main pre-registered specification).

We further consider the robustness of our main findings from Study 2 when employing Study 1's main pre-registered specification reported in the main text. This specification controls for pre-treatment covariates (without interacting them with the treatment), and employs a wild-cluster bootstrap procedure to account for clustered treatment assignment. As we show in Figure A25, our main pattern of results from the pre-registered specification reported in the

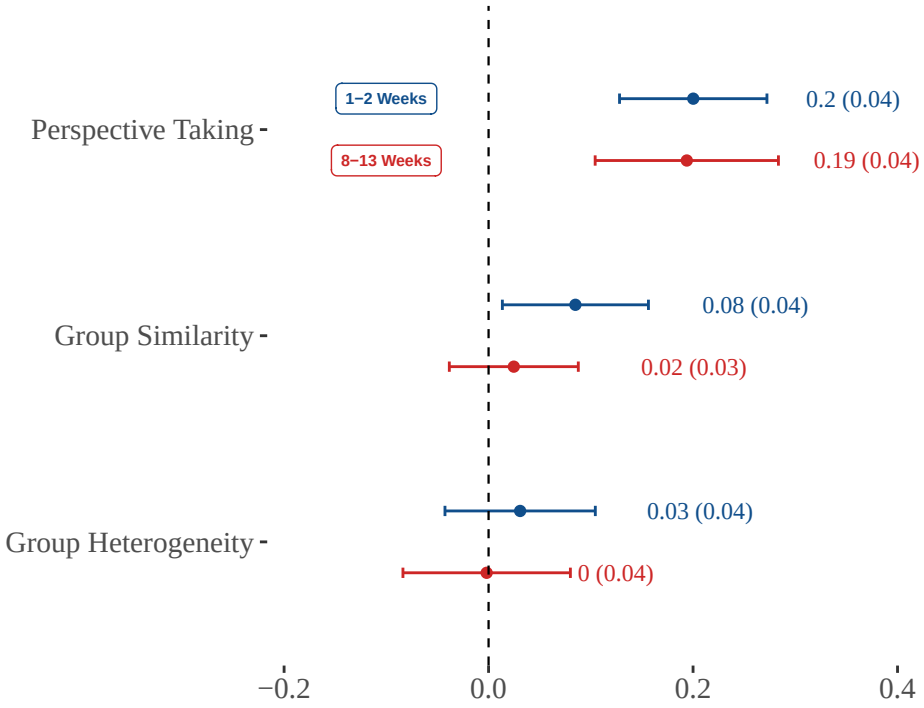


Figure A24: **The main pattern of mechanism results is consistent when focusing on measures only among respondents participating in both post-treatment survey waves.** This figure reports point estimates and 95% confidence intervals representing the effect of our intervention on mechanism outcomes amongst students participating in both post-treatment survey waves. Point estimates and standard errors (in parentheses) are reported along each estimate.

paper is largely consistent with this alternative specification. Similarly, when focusing on our results of potential mechanisms in Figure A26, we find a similar pattern with the exception of the group similarity mechanism, which is imprecisely estimated in both post-treatment waves in this alternative specification (but precisely estimated in our first post-treatment wave employing our pre-registered specification).

Table A7: Randomization Inference - Main Outcomes (Study 2)

	Term	Estimate	p.value
1	Thermometers W1	0.29	0.00
2	Thermometers W2	0.33	0.00
3	Contact W1	0.18	0.02
4	Contact W2	0.28	0.00
5	Diversity W1	0.15	0.04
6	Diversity W2	0.14	0.09
7	Register Contact W1	-0.06	0.44
8	Register Contact W2	-0.02	0.74
9	Bracelet W2	0.12	0.08
10	Perspective Taking W1	0.21	0.01
11	Perspective Taking W2	0.21	0.02
12	Similarity W1	0.09	0.35
13	Similarity W2	0.03	0.66
14	Heterogeneity W1	-0.00	0.96
15	Heterogeneity W2	-0.02	0.84

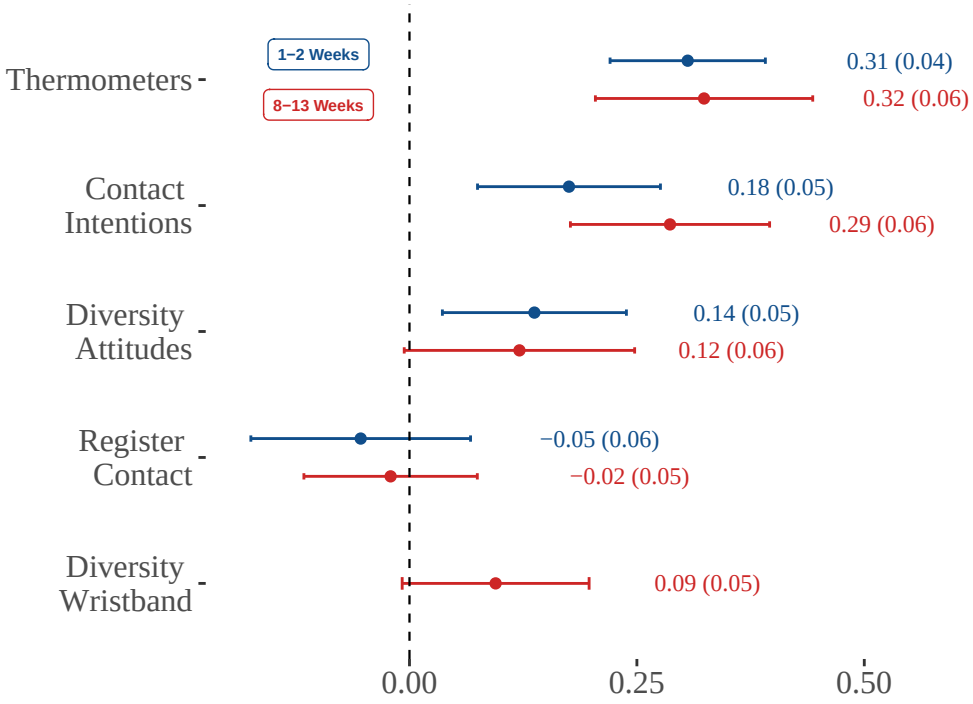


Figure A25: **The main pattern of results is consistent when employing the main specification from Study 1.** This figure reports point estimates and 95% confidence intervals representing the effect of our intervention on the main outcomes, employing the Study 1 specification in which we condition on pre-treatment covariates and employ a wild cluster bootstrap. Point estimates and standard errors (in parentheses) are reported along each estimate.

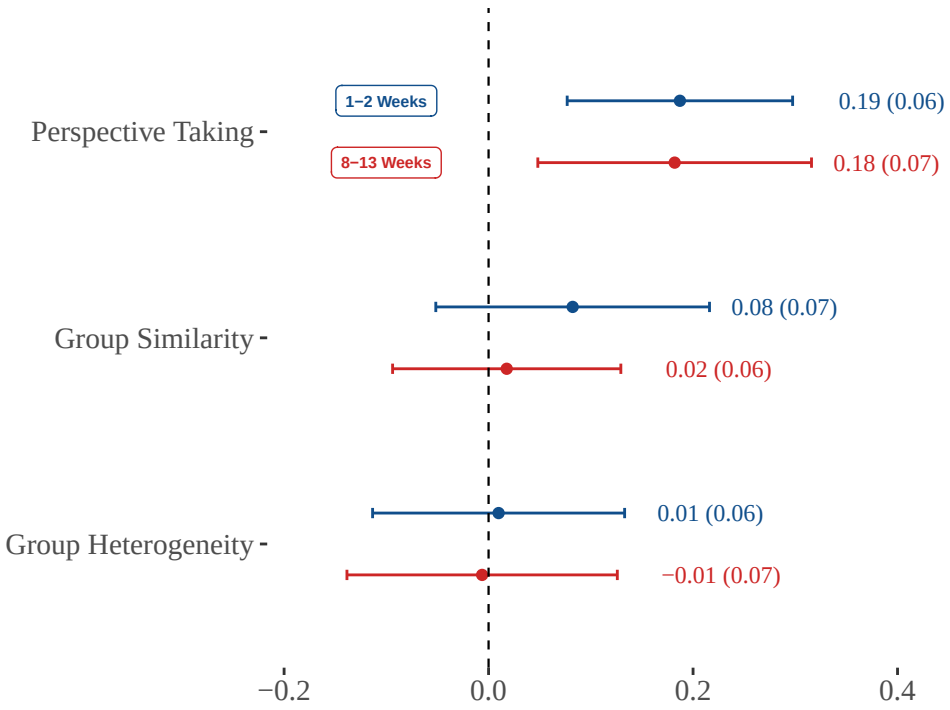


Figure A26: **The main pattern of mechanisms is consistent when employing the main specification from Study 1.** This figure reports point estimates and 95% confidence intervals representing the effect of our intervention on potential mechanisms employing the Study 1 specification in which we condition on pre-treatment covariates and employ a wild cluster bootstrap. Point estimates and standard errors (in parentheses) are reported along each estimate.

A5.3 Study 2: External Validity

Naturally, like in any empirical investigation, one might wonder whether the effects we identify generalize beyond our current sample. Addressing generalizability is largely an empirical task that can be addressed by a scholarly community that replicates similar findings in multiple contexts (*6*). We take the first step in this direction by showing that our main results replicate across two different studies in six different Israeli schools.

However, to further consider the generalizability of our evidence, we follow a procedure recommended by Deveaux and Egami (*7*), that provides a measure of an experiment's robustness to external validity bias. This measure ranges from 0-1, where 0 implies high sensitivity to external validity bias, and 1 implies low sensitivity to external validity bias. The premise of this measure is to quantify how much a sample would need to be different to explain an average treatment effect.

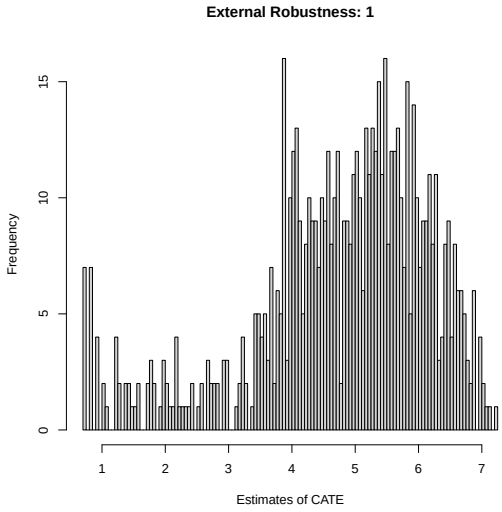
We follow two main steps to estimate Deveaux and Egami's measure and consider the sensitivity of our results to external validity bias. First, based on a set of pre-treatment covariates, we use the R package `exr` to identify the CATE (conditional average treatment effects) of our main estimate, given a set of covariates, using a causal forest machine learning approach. We then use the CATE, which essentially provides a measure of possible heterogeneity in response to treatment, to evaluate how much reweighting we would need to introduce into our sample given the estimated heterogeneity to explain away our main average treatment effects. This measure of sensitivity to external validity bias ranges between 0-1. Low rates on the 0-1 scale imply high sensitivity to external validity bias — in other words, even minimal reweighting could explain away the average treatment effect. In contrast, how rates on the scale imply low sensitivity to external validity bias — in other words, even substantial reweighting will not explain away the average treatment effect.

It is important to note that the virtue of this measure depends on the theoretical and practical

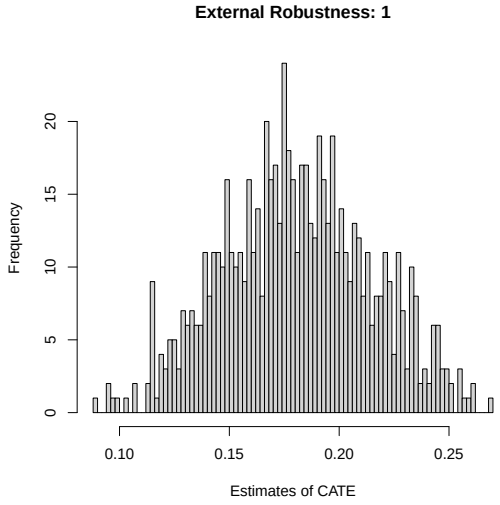
relevance of covariates used to estimate the CATE. In our case, we employ measures for a student's block, gender, age, and pre-treatment outcome measures. In that sense, as we see it, the main virtue of this exercise is in informing us about the following questions: would we reach a similar substantive conclusion if our experiments were to focus on samples that are substantively more (or less) prejudicial to out-groups? Would we reach a similar substantive conclusion if our experiments focus on younger or older students? Would we reach a similar substantive conclusion if our experiments focused primarily on male or female students? That said, since our CATE cannot speak directly to differences in average treatment effects between students and adults, this exercise cannot directly inform us about whether our results generalize to adult populations. Similarly, since our CATE cannot speak directly to differences in average treatment effects between students from Jewish and Arab backgrounds (because we don't have variation in ethnicity), this exercise cannot directly inform us about whether our results generalize to other subgroups in Israeli society.

With these caveats in mind, we attempt to address questions of external validity bias in Figure A27. To do so, we report estimated CATEs and our measure of external validity bias for each of our main findings in Study 2. Using the R package `exr` we specify our main models, as well as a set of pre-treatment covariates that might generate a degree of heterogeneity in response to treatment. The covariates include respondents' gender, age, experimental block, and all pre-treatment measures of prejudice. The four plots reported in Figure A27 provide reassuring evidence regarding external validity.

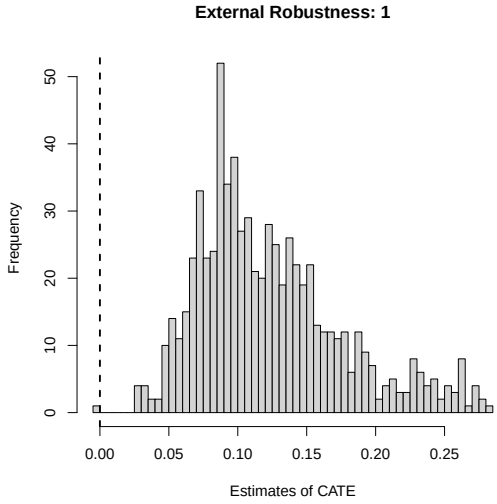
Indeed, it appears that our key results yield high levels of robustness to external validity bias. In other words, on a scale of 0-1, three of our four results receive a score of 1, and our behavioral measure's robustness score is 0.93. Substantively, this suggests that despite some heterogeneity of average treatment effects in our data, even substantial amounts of re-weighting to our sample would not explain away our average treatment effects. Ultimately, the encouraging results in



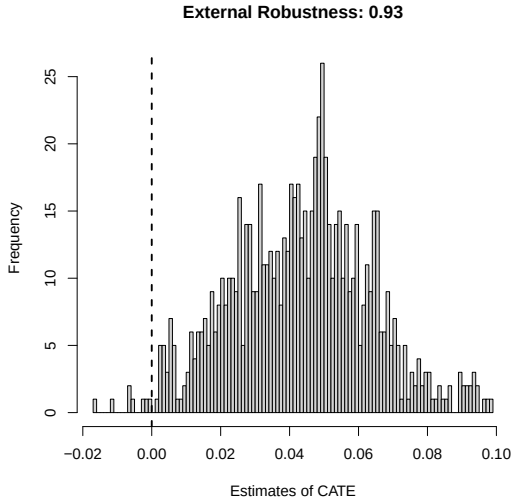
(a) Thermometers Index



(b) Contact Intention Index



(c) Diversity Attitudes Index



(d) Diversity Wrist Band

Figure A27: **Sensitivity analysis considering the robustness of Study 2 main estimates to external validity bias.** Each plot reports the CATE for a given main outcome in Study 2, as well as an associated measure of robustness to external validity bias ranging between 0 (high sensitivity) and 1 (low sensitivity).

Figure A27 are explained by the fact that while there is a degree of heterogeneity in our CATE, the CATE appears to almost always remain positive. In other words, almost no students respond

negatively to our intervention.

Thus, given our CATEs, re-weighting our sample to resemble some new target population of interest that is more (or less prejudicial), might reduce the size of our estimates, but is unlikely to explain away our average treatment effects. We construe results from this exercise as encouraging with regards to the potential external validity of our results. However, we encourage scholars to further replicate our findings in new contexts. Doing so, could inform us about variation in the magnitude of effects across different samples and populations.

References

1. E. L. Paluck, H. Shepherd, P. M. Aronow, *Proceedings of the National Academy of Sciences* **113**, 566 (2016).

2. A. C. Cameron, J. B. Gelbach, D. L. Miller, *The review of economics and statistics* **90**, 414 (2008).

3. A. S. Gerber, D. P. Green, *Field experiments: Design, analysis, and interpretation* (WW Norton, 2012).

4. J. Hainmueller, J. Mummolo, Y. Xu, *Political Analysis* **27**, 163 (2019).

5. W. Lin, *The Annals of Applied Statistics* **7**, 295 (2013).

6. C. Samii, *The Journal of Politics* **78**, 941 (2016).

7. M. Devaux, N. Egami, *Working Paper* (2022).