Addressing the Elephant in the Room: Field Experiments in Israel Show that Education Programs that Broach Sensitive Topics Can Reduce Prejudice

Chagai M. Weiss,[1*] Shira Ran,[2] Eran Halperin[3,2]

[1]Department of Political Science, University of Wisconsin – Madison.
[2]aChord Center, Hebrew University of Jerusalem.
[3]Department of Psychology, Hebrew University of Jerusalem.

[*]To whom correspondence should be addressed; E-mail: cmweiss3@wisc.edu.

Prejudice reduction interventions often avoid broaching sensitive topics, fearing that addressing grievances, disagreements, or dehumanizing beliefs may backfire and increase prejudice. We argue that interventions that overlook sensitive issues do not engage with the foundations of intergroup tensions and suggest that constructively broaching sensitive topics during prejudice reduction interventions can effectively improve intergroup relations. Building on this argument, we develop and test a scalable educational program inspired by the Israeli TV series "You Can't Ask That," which depicts children from minority groups discussing sensitive topics at the core of intergroup relations. Through two field experiments, we show that the intervention improved Jewish students' attitudes towards minorities (0.32 SD) and pro-diversity behavior (0.1 SD) up to 13 weeks post-treatment and provide suggestive evidence that the intervention was effective by encouraging students to take their outgroups' perspectives.

Introduction

Sensitive topics relating to intergroup grievances, disagreements, and dehumanizing beliefs are central components of intergroup tensions. For example, beliefs about outgroup brutality and lack of morality motivate support for conflict (*1*). Despite the salience of sensitive topics in conflict-ridden societies, they often remain unspoken in intergroup interactions (*2, 3*), but still influence intergroup perceptions and preferences (*4*). Moreover, many existing interventions for prejudice reduction that focus on direct and indirect intergroup contact and exposure, actively encourage discussions of intergroup commonalities and relationship building (*5, 6, 7, 8, 9*), and discourage conversations about group grievances, disagreements, and dehumanizing beliefs (*10, 11, 12, 3*).

The emphasis of most existing interventions on group similarities and avoidance of group differences is rooted in theory and evidence. Theories of direct, extended, vicarious, and parasocial intergroup contact all emphasize that exposure to outgroups can reduce prejudice if it is "positive" — entailing cooperation rather than competition and emphasizing intergroup similarities rather than differences (*13, 14, 15, 5*). In line with these theories, studies show that positive exposure to outgroups can reduce prejudice (*5, 6, 9, 16*), and exposure that entails competition or discussions of grievances and disagreements does not affect prejudice (*17*), and at times impairs intergroup relations (*18, 19*). Accordingly, avoiding sensitive topics in prejudice reduction interventions makes sense because initiating sensitive and taboo conversations might impose burdens on minorities (*20*) while failing to reduce majority group members' prejudice.

We argue that despite some encouraging evidence, prejudice reduction interventions that shy away from sensitive topics face two significant limitations. First, such interventions often fail to raise majority group members' awareness of the challenges and concerns that minority group members face (*12*), leading to increased intergroup harmony at the cost of hindering intergroup

equality (*10, 12, 3*). Second, interventions that avoid sensitive topics overlook some of the most salient dimensions of intergroup tensions. This might explain why recent meta-analyses find that existing prejudice reduction interventions have modest effects on average (*7, 21*).

The importance of broaching sensitive topics during prejudice reduction interventions, combined with the potential risk of backlash, and the ethical concern of burdening disadvantaged groups, raises a challenging question for scholars and practitioners alike: How can one reduce prejudice while directly addressing sensitive topics without risking backlash? Following recent calls to develop and test novel theoretically-informed interventions to improve intergroup relations (*7, 21*), we lay out a prejudice reduction approach that combines parasocial intergroup contact with constructive discussions regarding sensitive and taboo topics at the core of intergroup relations.

We designed a month-long educational intervention based on the Israeli TV series "You Can't Ask That."[1] Each episode of this show depicts charismatic children from different social groups responding with sophistication and humor to sensitive questions from home audiences that children would never consider asking outgroups directly (see Figure 1). Our educational intervention focused on three episodes depicting Arab, visually impaired, and immigrant children,[2] and included four sessions in which Jewish elementary school children watched the TV series and engaged in guided follow-up classroom discussions. The core objective of the intervention was to expose students to outgroups, constructively discuss issues relating to intergroup grievances, disagreements, and dehumanizing beliefs, and in doing so, improve intergroup attitudes and behaviors.

[1]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.

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

(a) Arab Children

(b) Immigrant Children

(c) Visually Impaired Children

Figure 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.

Broaching Sensitive Topics and Reducing Prejudice

We argue that linking the discussion of sensitive topics with common psychological mechanisms of prejudice reduction can ensure that "addressing the elephant in the room" does not come at the cost of reducing prejudice. Additionally, we suggest combining psychologically informed discussions with parasocial contact interventions in which majority group members are exposed to highly charismatic and persuasive minorities in mass media (*5, 6, 9*), can ensure that the process of addressing sensitive topics does not overburden disadvantaged groups.

But how exactly can one link between discussions of grievances, disagreements, or differences and popular mechanisms of prejudice reduction relating to within-group heterogeneity, intergroup similarity, and perspective taking? Constructive discussions of cultural differences

between social groups can emphasize outgroup heterogeneity. Doing so, such discussions inform majority group members' that outgroups members vary in their commitment to cultural practices and moral obligations. The realization that outgroups are not homogenous regarding their practices and values can emphasize that intergroup differences are not so stark and in turn lead to prejudice reduction (*22*).

Similarly, constructive discussions of intergroup disagreements regarding conflict resolution can emphasize how ingroups and outgroups share similar preferences, motivations, and feelings. In turn realizations about intergroup similarity can improve majority group members' attitudes toward minorities (*23,24,25*). Finally, discussions of group grievances can emphasize how often minorities suffer from state repression or discrimination and how harmful such experiences can be. In turn, constructive discussions of grievances provide a unique opportunity for perspective taking that can effectively reduce prejudice (*26, 27*).

In light of the proposed premises, we designed our intervention to focus on three episodes of the TV series "You Can't Ask That." The show's main objective is to provide home audiences with an opportunity to ask members of different social groups forthright questions regarding sensitive topics and to generate a constructive discussion (in the studio) about intergroup grievances, disagreements, and dehumanizing beliefs. 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 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.

In designing our intervention, our primary goal was to facilitate parasocial intergroup contact in which consenting minority children constructively discuss sensitive topics, which are

then linked with conducive psychological mechanisms during follow-up classroom activities. We selected episodes focusing on three outgroups: Arab, visually impaired, and immigrant children. Within these episodes, children discussed multiple topics and broached conversations about experiences that are rarely addressed in interpersonal exchanges. Such experiences include being targeted by hate speech and violence, excluded based on identity, detained by immigration police, and bullied for physical disabilities.

Our intervention included four meetings. In the first three meetings, students watched a group-specific episode (15 minutes) and engaged in follow-up classroom 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). We designed classroom discussions to focus on the TV series' central theme—discussions of sensitive topics—and to connect this theme with the core theoretical mechanisms noted above, relating to information about outgroup heterogeneity, intergroup similarities, and perspective taking. An Elaborate description of the TV series is provided in Appendix A1, and an overview of our educational program is described 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 survey and behavioral measures.

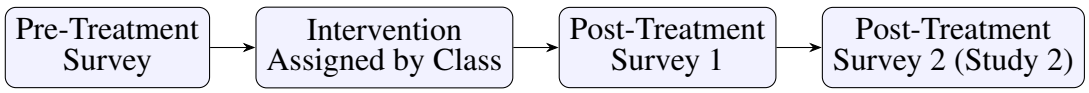


Figure 2: **Study Procedure.**

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.[3] 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.

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 largely operated, 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 from 253 students. 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 in-

[3] 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.

tentions 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 preregistered OLS regressions in which we regress standardized outcomes ($\mu = 0$, $\sigma^2 = 1$) over our treatment, controlling for respondents' gender, assignment block, and pre-treatment outcome measures.[4] Given the modest number of clusters in our data, we employ a wild-cluster bootstrap procedure to cluster our errors at the classroom level (*28*). Our main estimating equation is:

$$y_{ic} = \beta Z_c + \phi \boldsymbol{X}_{ic} + \varepsilon_{ic} \tag{1}$$

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. 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 feeling thermometer. This result is almost double the magnitude

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

of the average effect of well-powered interventions report in a recent meta-analysis of prejudice reduction experiments (*21*).

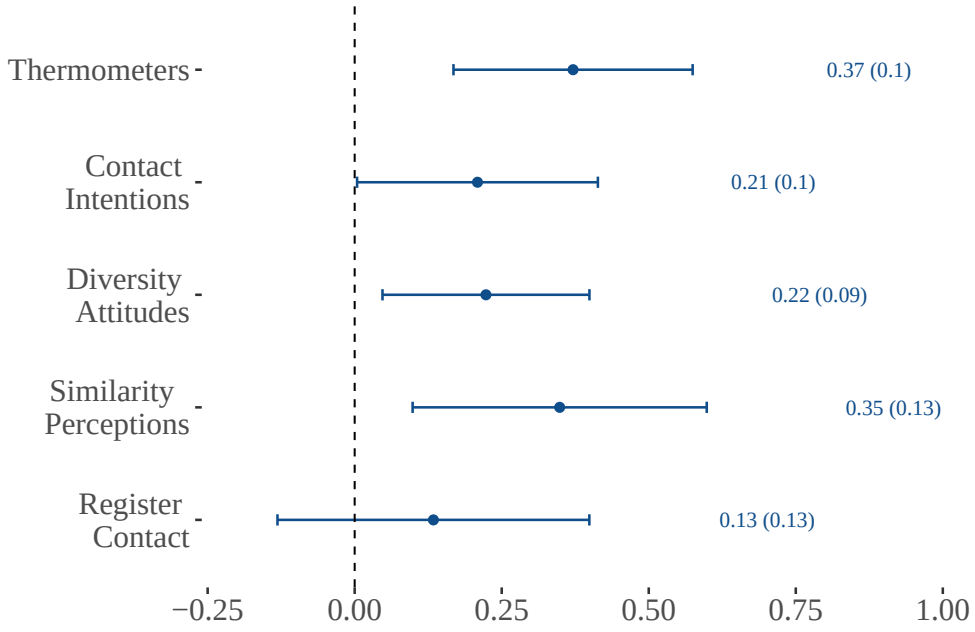


Figure 3: **Exposure to the intervention improved children's attitudes in Study 1.** This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on students' attitudes and behaviors 1-2 weeks post-treatment. Point estimates and standard errors (in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

We 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 shaped 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 study.

We subject our results to several diagnostic and robustness checks. In the Appendix, we report balance checks (Appendix A4.1), examinations of attrition (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 emphasize how 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,[5] 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 an essential dimension of scalability relating to treatment implementation (*29*). Thus, teachers from treated classes were provided with all necessary materials to implement the intervention and were instructed to deliver four sessions of the intervention over four

[5]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.

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 completed all four sessions of the intervention, we returned to each school 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 treated classes completed all 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 problems 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 our intervention. 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. We measure whether each student selected the pro-diversity bracelet and interpret this selection as an act of

signaling support for diversity. 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 the weighted least square regression depicted in equation 2. Following our pre-analysis plan, we interact mean-centered covariates with our main treatment indicator to increase precision (*30*), cluster errors by class and employ weights that account for varying treatment assignment probabilities across blocks (*39*). Covariates include gender, assignment block, and pre-treatment outcome measures,[6] and all outcomes are standardized ($\mu = 0$, $\sigma^2 = 1$).

$$y_{ics} = \beta Z_{cs} + \phi \boldsymbol{X}_{ics} + \gamma(Z_{cs} * \boldsymbol{X}_{ics}) + \varepsilon_{ics} \tag{2}$$

In our analyses, we focus on β, representing the average 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 pro-diversity behavior but did not increase 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.

Indeed, exposure to our month-long intervention increased students' intergroup positive affect toward outgroups by almost a third of a standard deviation in the short and longer term.

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

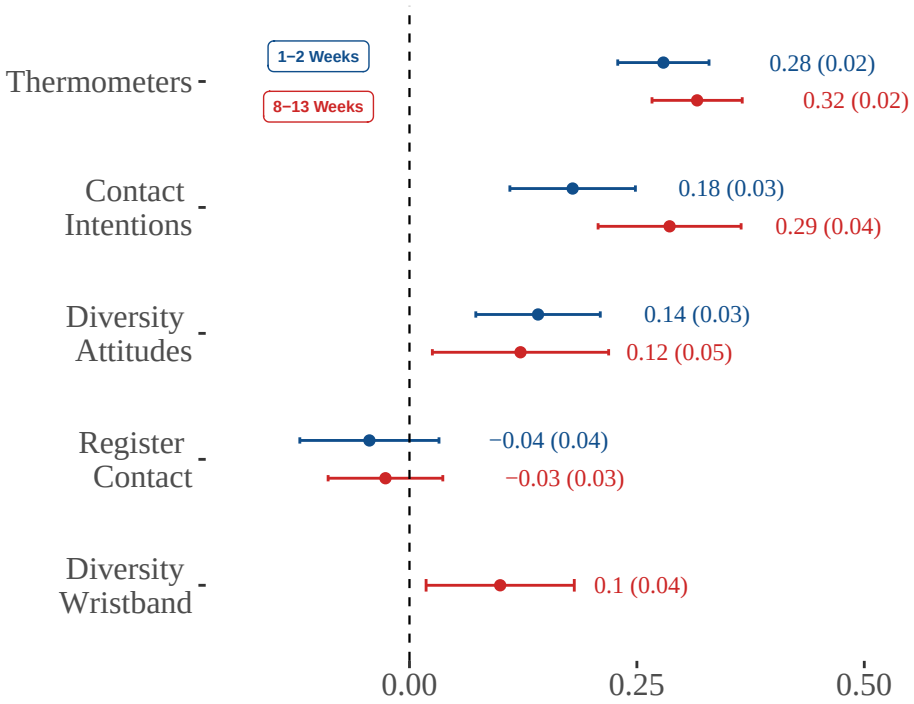


Figure 4: **Exposure to the intervention in Study 2 improved children's attitudes and behaviors up to 13 weeks post-treatment.** This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on students' attitudes and behaviors 1-2 (8-13) weeks post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

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. 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 this 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 constructively broaching sensitive topics during parasocial intergroup contact intervention? To answer this question, we turn to survey items measuring the three psychological mechanisms emphasized in our program: within-group 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 outgroups as a whole are heterogeneous, that students are similar to outgroups, and that it is important to try and take the outgroup's perspective.

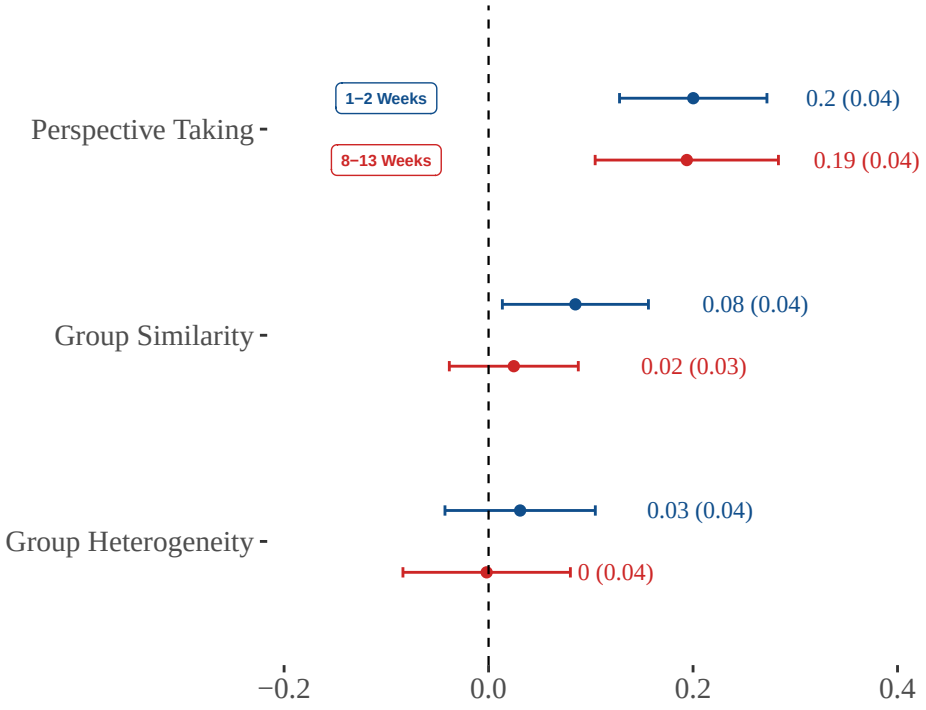


Figure 5: **Exposure to the intervention in Study 2 increased students' ability and willingness to take outgroup perspectives, had a limited impact on perceptions of cross-group similarities and had no effect on perceptions of within-group heterogeneity.** This figure reports point estimates and 95% 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 the outgroups' perspective, smaller and inconsistent effects on students' perceptions of intergroup similarity, and no noticeable effect on perceptions of within-group heterogeneity. We cautiously interpret these additional results to suggest that our intervention, which exposed students' to parasocial contact and broached a variety of sensitive topics, was effective because it helped students appreciate 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 engagement with sensitive topics facilitated increased appreciation for understanding other groups' circumstances and in turn, improved intergroup attitudes and behaviors.

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 Section A5.2 that a similar pattern of results emerges when focusing only on respondents participating in both surveys.

Discussion

In this paper, we attempted to develop a theoretically driven approach for prejudice reduction that directly engages with sensitive topics at the heart of intergroup relations without generating backlash or imposing burdens on disadvantaged groups. Through multiple field experiments in Israel, we show that our intervention had substantial long-term effects on Jewish students' attitudes and behaviors. Our findings contribute to several theoretical and applied questions.

First, we contribute to the literature on prejudice reduction. Our results emphasize that successful prejudice reduction interventions do not need to avoid discussions regarding intergroup grievances, disagreements, and dehumanizing beliefs and that critical yet constructive confrontations of these topics can reduce prejudice among children. Our findings from Israeli

schools are especially notable given extensive literature documenting the acquisition of social and political attitudes at a young age (*32*), and the ability of education systems to facilitate effective socialization (*33*).

Second, we join a growing body of research that employs natural and field experiments to gain insight on prejudice reduction and conflict resolution (*34, 8, 9, 19, 16*). We develop and test a theoretically motivated, intensive, and scalable intervention in a naturalistic setting, measuring attitudes and behaviors amongst our population of interest. Moreover, by implementing multiple studies with several design modifications, we address central concerns regarding intervention scalability (*29*). Through our intensive field-based research, we depart from ongoing trends in the prejudice reduction literature, which employ nudge-like interventions and examine effects immediately post-treatment (*21*). In turn, our findings emphasize how intensive and carefully curated educational programs can reduce prejudice and serve to buffer the deterioration of intergroup attitudes even in times of escalated conflict.

Despite these contributions, our findings are not without limitations. First, like many field-experiments (*35, 8, 9*), 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 (*36*). Moreover, in Appendix A5.3 we quantify the sensitivity of our analyses to external validity bias (*37*), and find encouraging evidence regarding the likelihood that our evidence would generalize to substantially diverse target populations.

Second, our intervention is a bundle of multiple components, including parasocial contact 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 em-

pirical focus is motivated by the understanding that effective prejudice reduction interventions may very well require the "mixing of ingredients from multiple theoretical perspectives" (*21, p. 555*). Like recent landmark studies (*34*), 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 prejudice reduction interventions need not be exclusively harmonious in nature (*10*). That is, programs that broach sensitive topics constructively can improve intergroup attitudes in divided societies.

References

1. E. Bruneau, N. Kteily, *PloS one* **12**, e0181422 (2017).
2. L. R. Tropp, Ö. M. Uluğ, M. S. Uysal, *International journal of intercultural relations* **80**, 7 (2021).
3. N. K. Reimer, N. K. Sengupta, *Journal of Personality and Social Psychology* (2022).
4. A. Fischer, E. Halperin, D. Canetti, A. Jasini, *Emotion Review* **10**, 309 (2018).
5. E. Schiappa, P. B. Gregg, D. E. Hewes, *Communication monographs* **72**, 92 (2005).
6. S. Murrar, M. Brauer, *Group Processes & Intergroup Relations* **21**, 1053 (2018).
7. E. L. Paluck, S. A. Green, D. P. Green, *Behavioural Public Policy* **3**, 129 (2019).
8. A. Scacco, S. S. Warren, *American Political Science Review* pp. 1–24 (2018).
9. S. Mousa, *Science* **369**, 866 (2020).
10. T. Saguy, N. Tausch, J. F. Dovidio, F. Pratto, *Psychological Science* **20**, 114 (2009).

11. T. Saguy, *Policy Insights from the Behavioral and Brain Sciences* **5**, 75 (2018).

12. N. S. Kteily, K. J. McClanahan, *Current opinion in psychology* **33**, 80 (2020).

13. G. W. Allport, *The nature of prejudice* (Addison-Wesley Pub. Co., Reading, Mass., 1954).

14. Y. M. Herrera, A. Kydd, *APSA 2013 Annual Meeting Paper, American Political Science Association 2013 Annual Meeting* (2013).

15. A. Mazziotta, A. Mummendey, S. C. Wright, *Group Processes & Intergroup Relations* **14**, 255 (2011).

16. C. M. Weiss, *Proceedings of the National Academy of Sciences* **118** (2021).

17. E. Santoro, D. E. Broockman, *Science advances* **8**, eabn5515 (2022).

18. E. L. Paluck, *Personality and Social Psychology Bulletin* **36**, 1170 (2010).

19. M. Lowe, *American Economic Review* **111**, 1807 (2021).

20. J. A. Richeson, J. N. Shelton, *Current Directions in Psychological Science* **16**, 316 (2007).

21. E. L. Paluck, R. Porat, C. S. Clark, D. P. Green, *Annual review of psychology* **72**, 533 (2021).

22. W. Hsieh, N. Faulkner, R. Wickes, *British Journal of Social Psychology* (2021).

23. J. M. Falomir-Pichastor, C. Martínez, C. Paterna, *The Spanish Journal of Psychology* **13**, 841 (2010).

24. M. J. Brandt, *Psychological Science* **28**, 713 (2017).

25. Z. Liberman, A. L. Woodward, K. D. Kinzler, *Trends in cognitive sciences* **21**, 556 (2017).

26. E. G. Bruneau, R. Saxe, *Journal of experimental social psychology* **48**, 855 (2012).

27. J. Kalla, D. Broockman, *American Journal of Political Science* (2021).

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

29. A. Banerjee, *et al.*, *Journal of Economic Perspectives* **31**, 73 (2017).

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

31. A. Abadie, S. Athey, G. W. Imbens, J. Wooldridge, When should you adjust standard errors for clustering?, *Tech. rep.*, National Bureau of Economic Research (2017).

32. D. O. Sears, S. Levy (2003).

33. D. Bar-Tal, Y. Rosen, *Review of Educational Research* **79**, 557 (2009).

34. D. Broockman, J. Kalla, *Science* **352**, 220 (2016).

35. B. Hameiri, R. Porat, D. Bar-Tal, E. Halperin, *Proceedings of the National Academy of Sciences* **113**, 12105 (2016).

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

37. M. Devaux, N. Egami .

38. R. Gomila, C. S. Clark, *Psychological Methods* (2020).

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

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

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 Ryan Enos, Josh Kertzer, Nicholas Sambanis, Alex Scacco, Macartan Humphreys, Nahomi Ichinao, Anna Wilke, Jonathan Homola, and workshop participants at American University, Harvard MEI, Harvard WoGPoP, Harvard Department of Psychology, the University of Pennsylvania's Conflict and Identity Lab, WZB, and POLMETH XXXIX for helpful comments and suggestions.

Addressing the Elephant in the Room: Field Experiments in Israel Show that Education Programs that Broach Sensitive Topics Can Reduce Prejudice

Supporting Information for Online Appendix

Contents

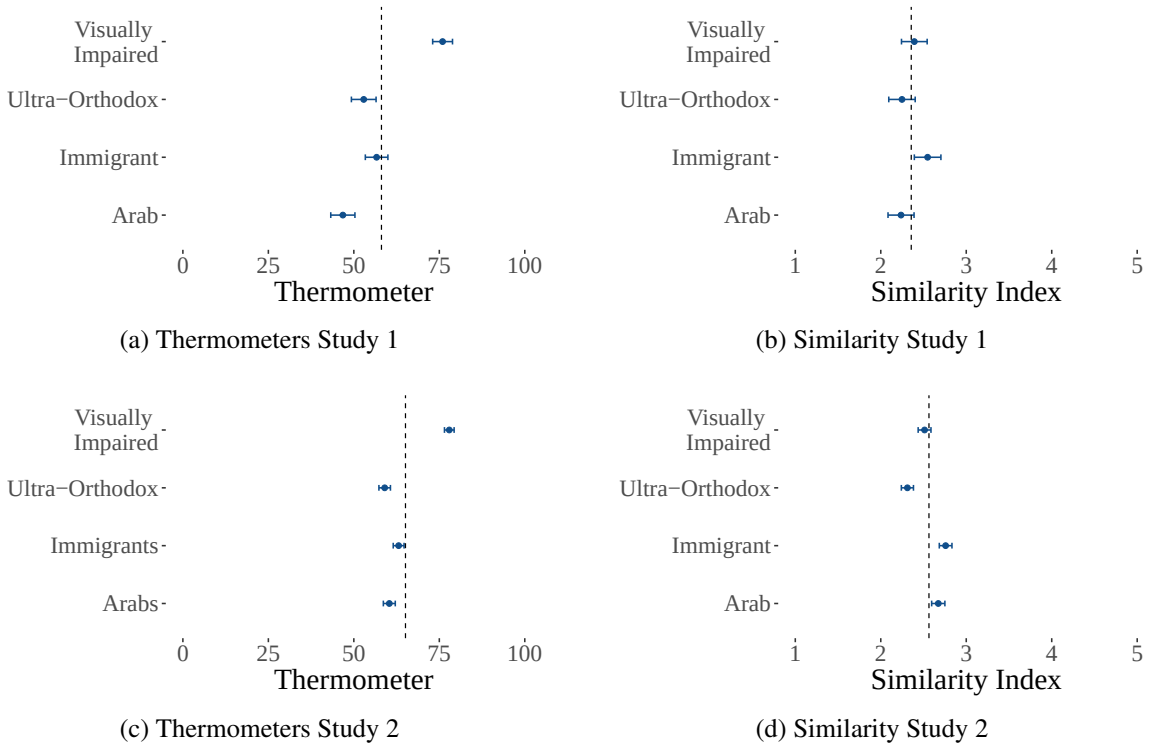
A1 You Can't Ask That

Our intervention was inspired by an Israeli TV series named *Slicha al Ha-Shela, Girsat Ha-Yeladim* which directly translates from Hebrew to mean "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.

To date, 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 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 intergroup similarity in Studies 1 and 2.** This figure demonstrates variation in affect to different outgroups and a general sense 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 prejudicial taboos.

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 every year. 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 because 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 an Arab speaking Arabic, they think they are planning a terror attack.

  * 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 the English language. 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 met 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 clearly 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 an aid. 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 uneth-

ical and that many historical changes that we care about, including the creation of the Israeli state, result 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?

  - 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 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 that they are afraid to be deported, while others emphasize that they cannot be deported because they luckily obtained Israeli citizenship.

  - 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 think 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, and another child describes a moment in which she bumped into a garbage pale, mistakened the garbage pale for 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 don't 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 sports 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 athletes with vision impairment.

- 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 them 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 star watching, 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 linking sensitive topics with psychological mechanisms of intergroup similarity, within-group heterogeneity, and perspective taking, we designed our classes to constructively engage with sensitive topics and link them with our theorized mechanisms.

Specifically, our first three classes focused on a particular social group presented in a particular episode. Class number 1 focused on sensitive topics relating to Arab children. In the process of unpacking these sensitive topics, 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 sensitive topics relating to visually impaired children. In the process of unpacking these 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 sensitive topics relating to children of immigrants. In the process of unpacking these 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 professional. 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. After 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 with 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 theme (e.g., perspective taking, group variability, or intergroup similarity), the curriculum was designed to focus on sensitive topics and link these topics with all three psychological mechanisms that appear in each episode. Moreover, the educational practitioner was instructed to link between the different classes, and indeed, each lesson started with a brief overview of recent class activities relating to the intervention.

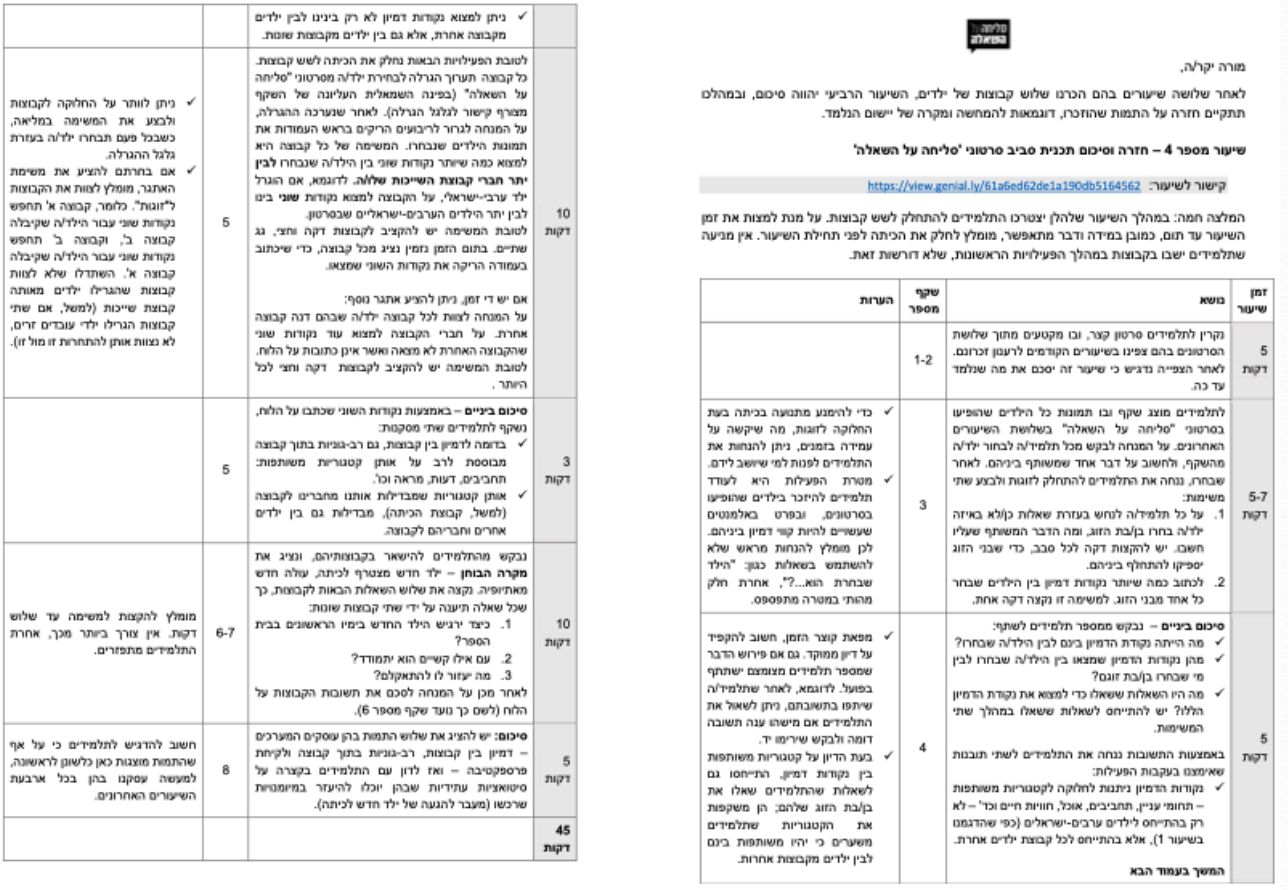
A2.2 Study 2 Implementation

In study 2, the content of our intervention remained the same. However, to assess 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). 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 open to teachers assigned to treatment, 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.



	✓ ניתן למצוא נקודות דמיון לא רק בינינו לבין ילדים מקבוצה אחרת, אלא גם בין ילדים מקבוצות שונות.		
10 דקות	לסוגת 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 העמודות 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. לסוגת המשימה יש להקציב לקבוצת דקה וחצי, גג שתיים. בתום הזמן נמין נציג מכל קבוצה, כדי שיכתוב בעמודה הריקה את נקודות השוני שמצאו. אם יש די זמן, ניתן להציע אתגר נוסף: על המנחה לצוות לכל קבוצת ילד/ה שבהם דנה קבוצה אחרת. על חברי הקבוצה למצוא עוד נקודות שוני שהקבוצה האחרת לא מצאה ואשר אינן כתובות על הלוח. לסוגת המשימה יש להקציב לקבוצת 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 מחברינו 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. נקצה 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 להגעה 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			

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 לרענן זכרונם. לאחר הצפייה נדגיש כי שיעור זה יסכם את מה שנלמד עד כה.	1-2	
5-7 דקות	לתלמידים מוצג שקף ובו תמונת 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, ננחה 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 מתסיגע בכיתה בעת החלוקה לזוגות, מה שיקשה על עמידה בזמנים, ניתן להנחות את התלמידים לפנות לתי 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 מהתי 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 וכו' – 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.

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

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.

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 A3-A4 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 respondent to the survey privately 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".[7]

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

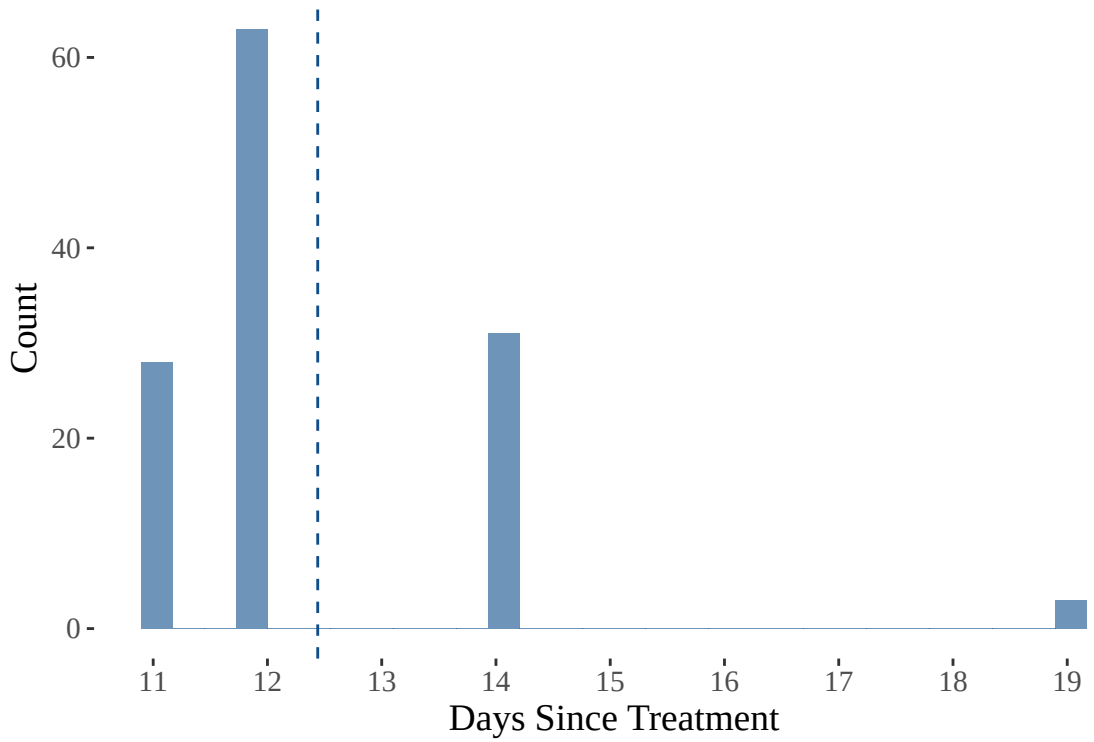


Figure A3: **Distribution of time between treatment and survey implementation in Study 1 for treated students.**

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 A5, 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.

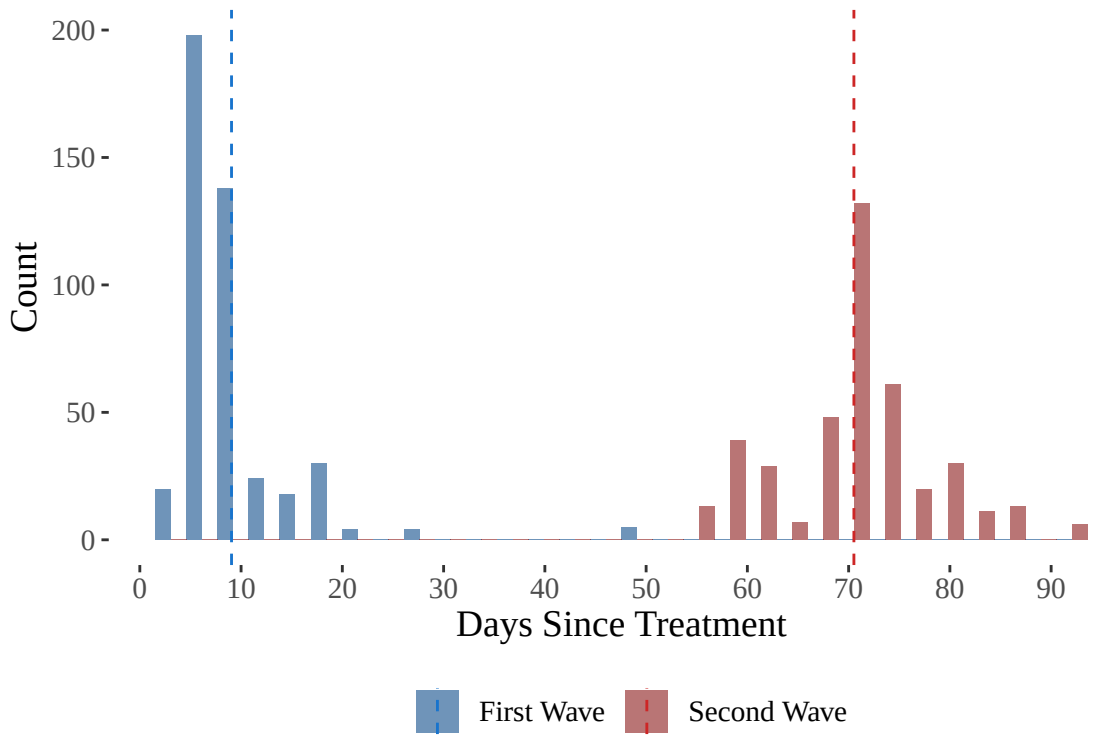


Figure A4: **Distribution of time between treatment and survey implementation in Study 2 for treated students.**

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 A5: **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 this 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 the control group.

A4 Study 1 Additional Analyes

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 attitudes. As depicted in Table A3, we can not reject the null hypothesis for any pre-treatment variable. Some pre-treatment variables, including attitudes towards Ultra-Orthodox and Arabs, are slightly unbalanced, but these differences are small and accounted for in our estimation strategy. More importantly, for the overall balance test reported in Table A3, we can not reject the null hypothesis of similarity ($p = .379$), 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 endline survey. Moreover, as further discussed in Section A4.4, due to a technical error, all students were randomly exposed to four out of five batteries of questions relating to intentions

Table A3: Balance Table (Pre-Treatment Measures)

	Adjusted Difference	Signifcance
Boy	0.10	
Age	-0.01	
Arab Thermometer	-0.12	
Immigrant Thermometer	-0.27	
UO Thermometer	-7.69	*
Blind Thermometer	-2.17	
Thermometer Scale	-2.56	
UO Similarity	-0.36	*
Arab Similarity	-0.26	.
Immigrant Similarity	0.03	
Blind Similarity	-0.18	
Similarity Scale	-0.19	
Diversity Scale	-0.13	
Contact Scale	-0.08	

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.

To reduce concerns regarding bias in our estimates as a result of attrition, in Table A4 we report a series of regressions diagnosing the correlates of attrition. Specifically, we regress a binary indicator taking the value of 1 if a respondent did not participate in the endline survey (column 1), or did not report answers to a group's specific contact intention item (columns 2-4) over four key variables: A treatment indicator, a gender indicator, an age variable, and our pre-treatment thermometer index. We do not find any evidence that attrition correlates with treatment or other pre-treatment measures. This finding reduces concerns that attrition threatens the internal validity and unbiasedness of our main estimates in Study 1.

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 (*28*). However, in

Table A4: Correlates of Missing Values

	Full Post	Vis Impaired Contact	Arab Contact	Immigrant Contact	UO Contact
	(1)	(2)	(3)	(4)	(5)
Treatment	0.008	−0.009	0.043	−0.064	0.026
	(0.030)	(0.053)	(0.052)	(0.051)	(0.052)
Boy	−0.025	0.007	−0.010	0.044	−0.080
	(0.030)	(0.054)	(0.052)	(0.051)	(0.052)
Age	0.020	−0.012	0.020	−0.042	0.054
	(0.016)	(0.030)	(0.029)	(0.029)	(0.029)
Therm Index	−0.00005	0.001	0.002	0.001	−0.004
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
NAs	17	56	53	51	55
N	270	253	253	253	253

this section, we implement additional analyses that employ randomization inference to address concerns regarding the modest number of clusters in Study 1 (*39*).

We report randomization inference results for our main outcomes from Study 1 in Table A5. 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 A5: 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 report average treatment effects on general attitudes towards multiple outgroups rather than particular attitudes towards specific social groups. In Figure A6 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.

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 towards visually impaired students. Somewhat similar patterns emerge with regards 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 intergroup similarity with visually impaired children are large and precisely estimated.

We argue that the substantively small and imprecisely estimated treatment effect of treatment 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. 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 A6.

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 students 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 A7 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

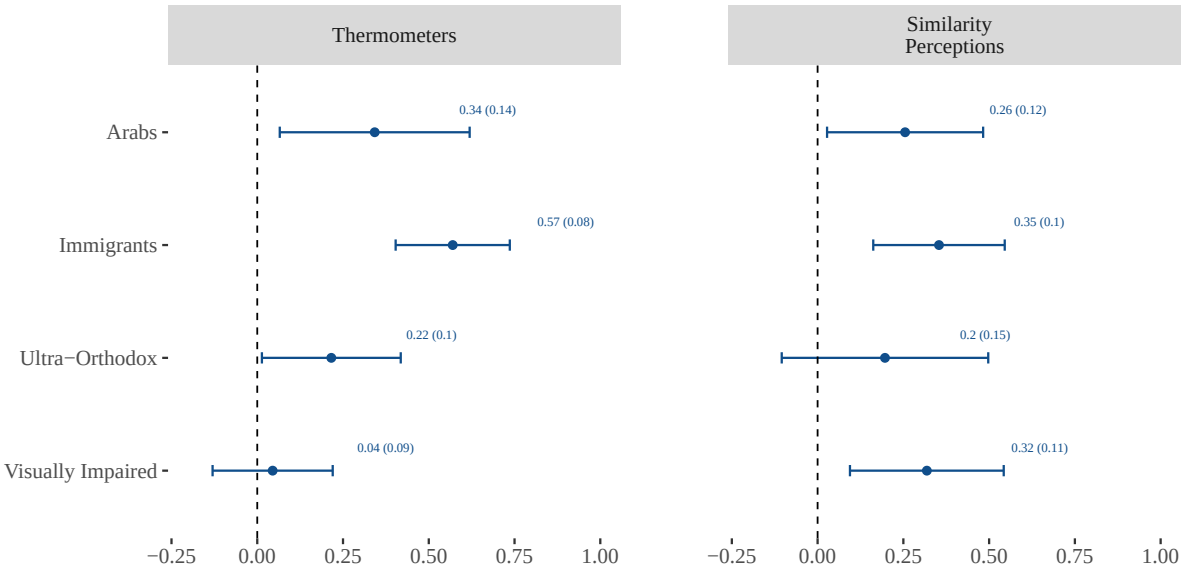


Figure A6: **Exposure to the intervention in Study 1 improved children's attitudes towards most social groups.** This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on students' attitudes towards specific social groups 1-2 weeks post-treatment. Point estimates and standard errors (in parentheses) are reported along each estimate.

a binning estimator proposed by Hainmueller et al. (*40*) in Figure A8.

While the pattern in Figure A8 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. (*40*)). We thus interpret the evidence in Figure A8 to suggest that both prejudicial and non-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.

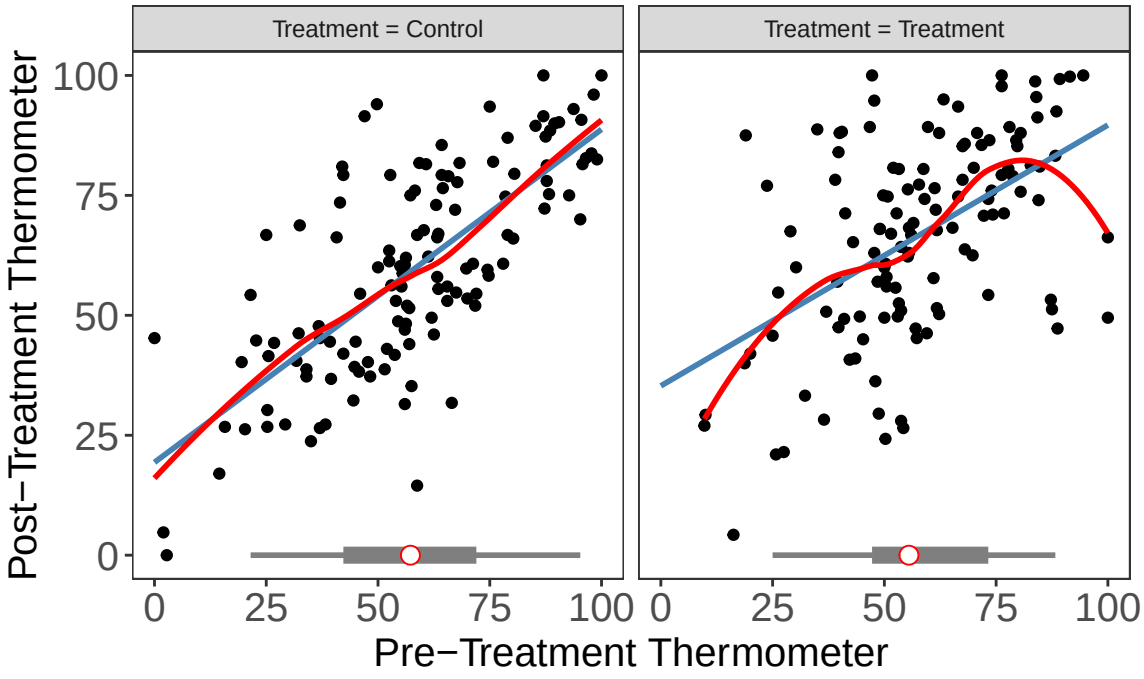


Figure A7: **Interaction term diagnostics Study 1.** This Figure plots the post-treatment thermometer index over the pre-treatment thermometer index by treatment condition.

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' attitudes and specific 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 towards Arabs. In Figure A9 we plot the pre-and

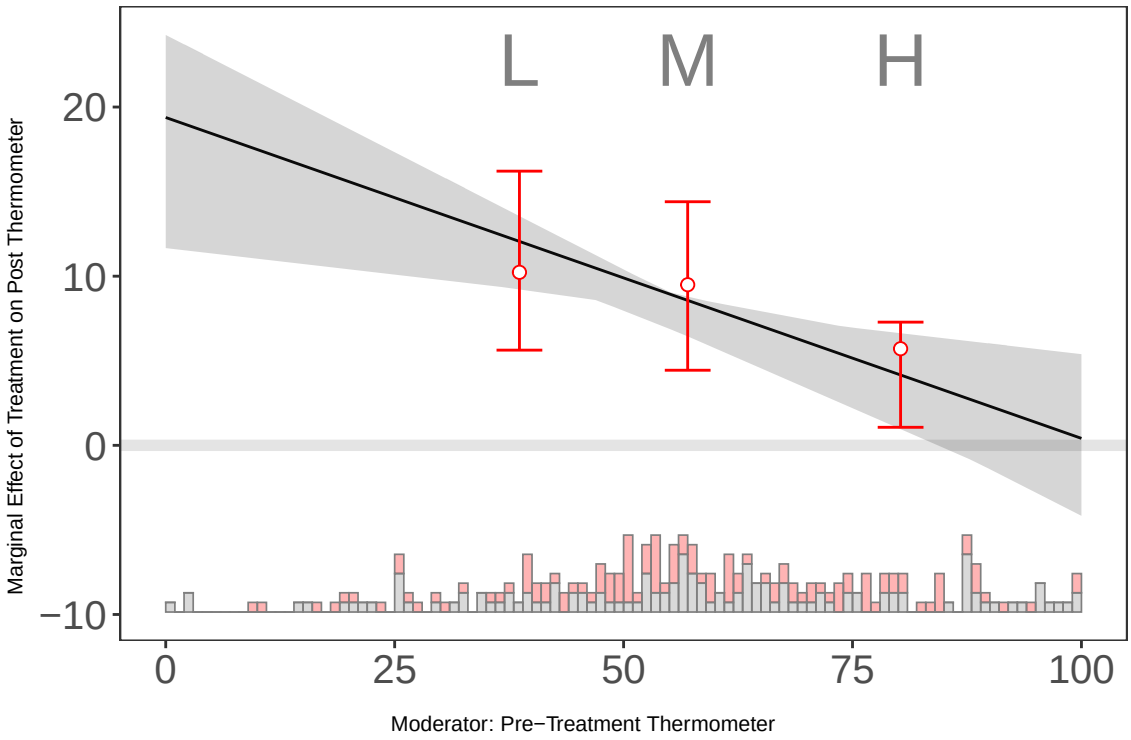


Figure A8: **In Study 1, pre-treatment attitudes do not moderate average treatment effects.** This plot reports that the average treatment effect of our intervention conditional on levels of our pre-treatment thermometer index using the binning estimator proposed by Hainmueller et al. (*40*).

post-treatment means for treated and control students.

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 A9 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 address sensitive topics at the heart of intergroup

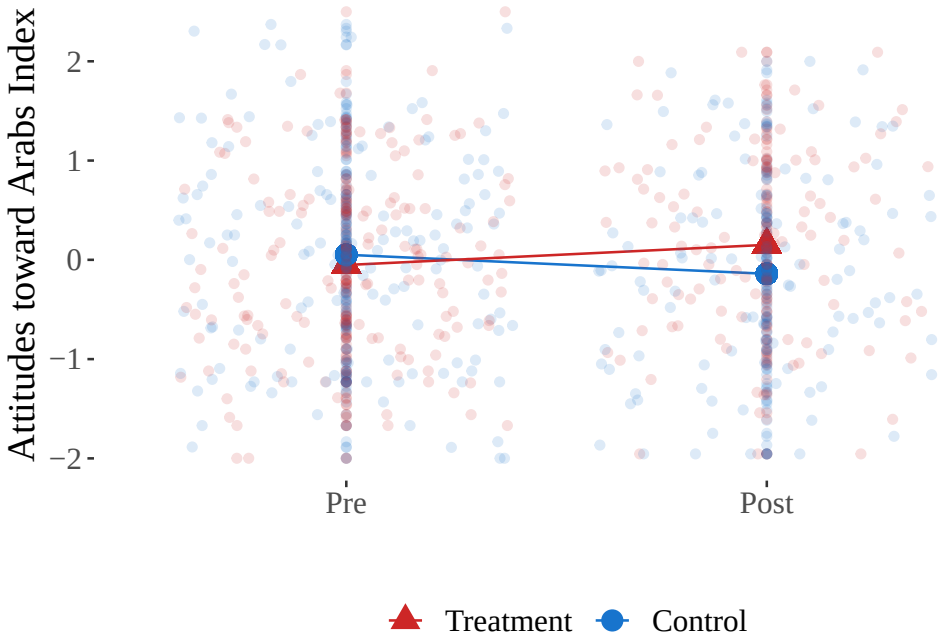


Figure A9: **Attitudes toward Arabs improve (deteriorate) among treated (control) students over time in Study 1.** This figure reports students' overall attitudes toward Arabs before and after the intervention by treatment status.

relations can increase positive attitudes towards outgroups event at times of intensifying conflict.

A4.4 Deviation from Pre-Analysis Plan

In analyzing study 1, we make 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 Table A4, since the presentation of batteries to each student was randomly assigned by Qualtrics, missingness is not correlated with treatment, pre-treatment attitudes, or demographics.

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 A6 focusing on student survey respondents' gender, age, and grade. In Table A7 we further report balance tests considering respondents' demographics and pre-treatment levels of prejudice. The thermometer index appears to be slightly unbalanced. However, this unbalance is small and accounted for in our estimation strategy. More importantly, for the overall balance test reported in Table A7 for all covariates, we can not reject the null hypothesis of similarity ($p = .142$), providing further assurance that our treatment

and control groups are similar on observables and unobservables.

Table A6: 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 A7: Balance Table (Pre-Treatment Measures) - Study 2

	Adjusted Difference	Signifcance
Boy	0.00	
Age	-0.02	
Contact Intentions Index	0.00	
Thermometers Index	-2.95	*
Diversity Attitudes Index	-0.08	
Group Heterogeneity Index	0.01	
Group Similarity Index	0.04	
Perspective Taking Index	-0.03	

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 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 external 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 A13, 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.

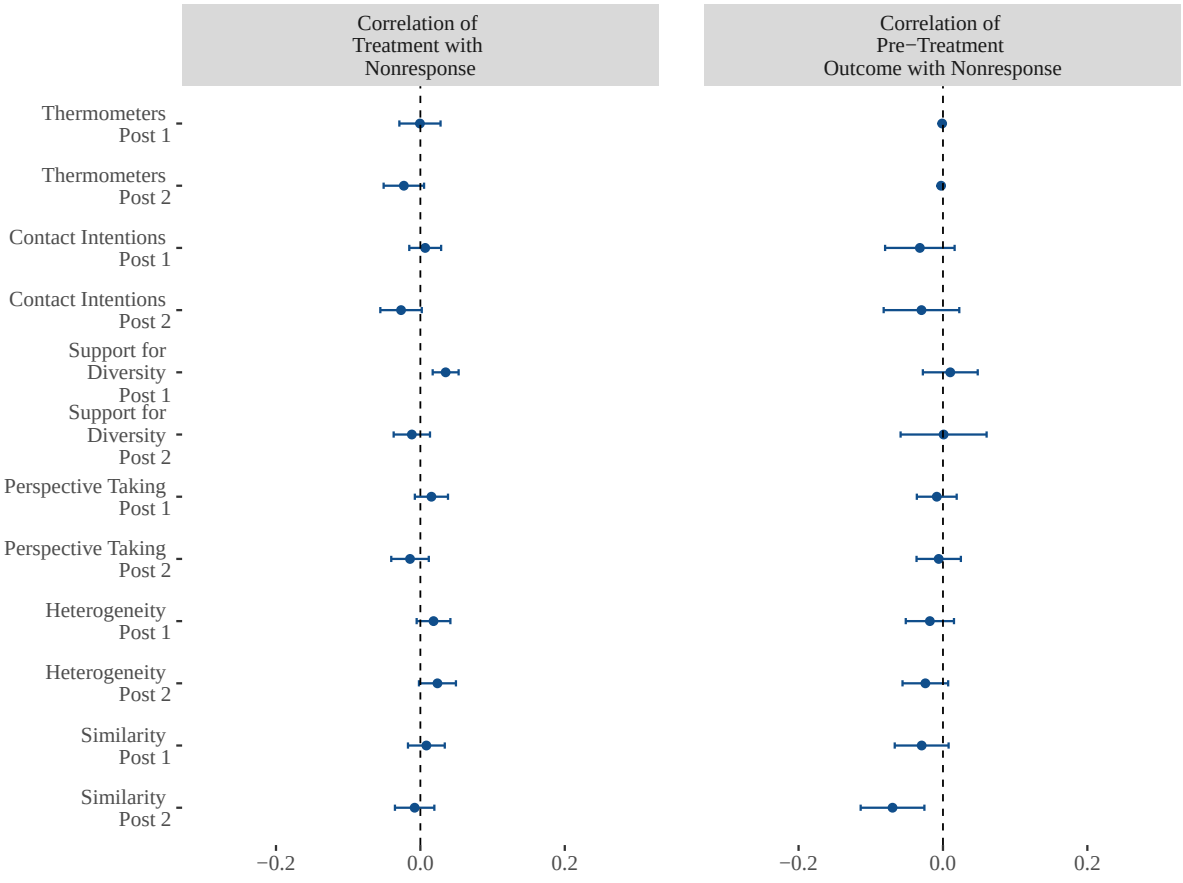


Figure A10: **Non-Response to survey items does not consistently correlate with treatment or pre-treatment attitudes in Study 2.** This figure reports point estimates and 95% confidence intervals representing the correlation of non-response for a given survey item with respondents' treatment assignment status and pre-treatment measure of outcome.

The results in Figure A13 reduce concerns regarding selective attrition. In all but one model on the left-hand side of Figure A13, 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 A13 pre-treatment measures of an outcome do not predict attrition. We interpret the small and insignificant point estimates reported in Figure A13 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.

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 A11 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 A11, which are harder to estimate with precision.

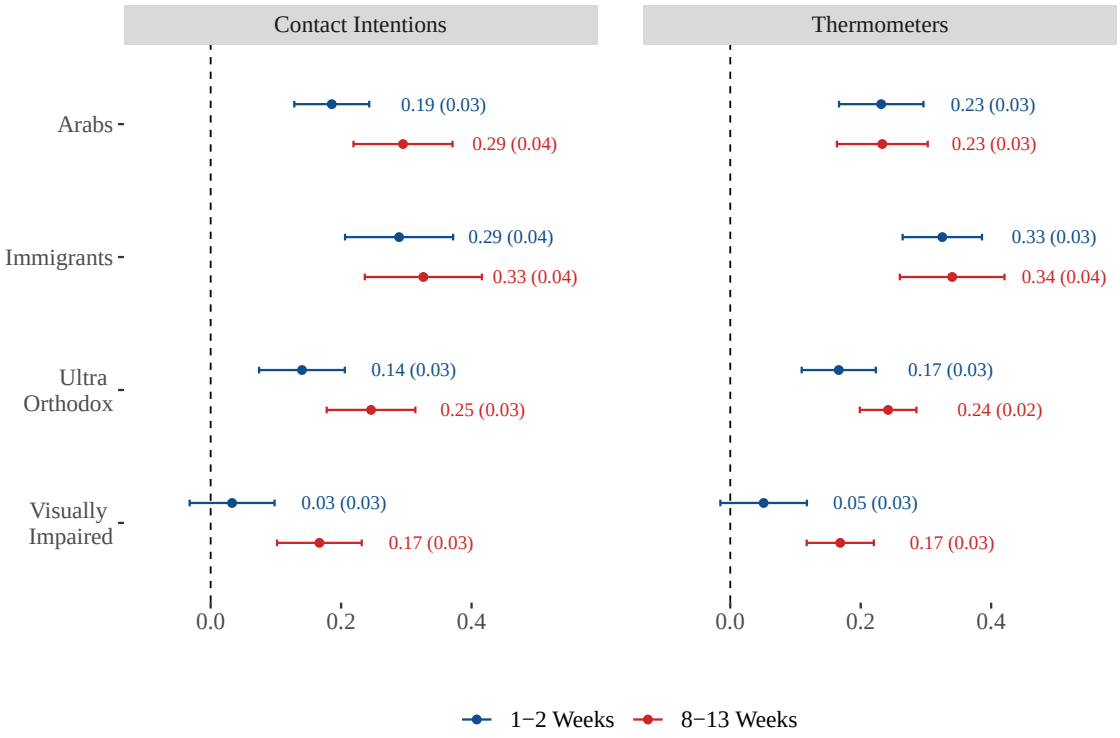


Figure A11: **The Intervention had significant effects on attitudes towards all social groups with the exception of visually impaired children.** This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on disaggregated components of the contact and thermometer indices. Point estimates and standard errors (in parentheses) are reported along each estimate.

In Figure A12 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 A13 we further examine the average treatment effects of our intervention on disaggregated measures of our central outcomes: 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.

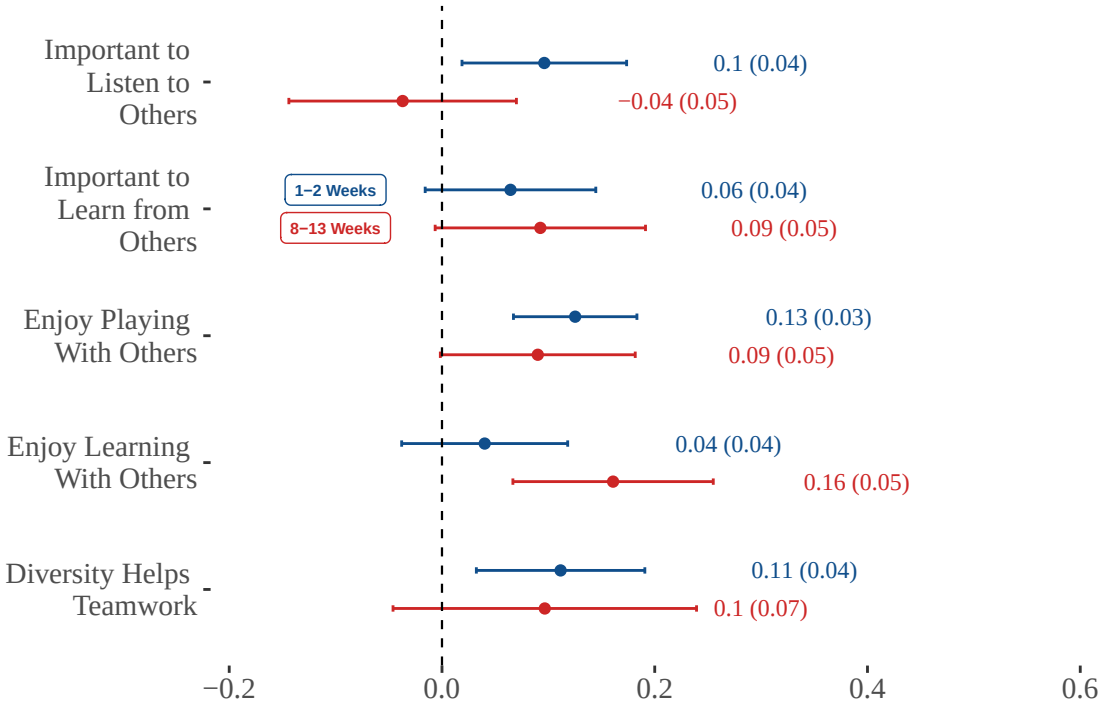


Figure A12: **The Intervention affected a majority of disaggregated support for diversity measures.** This figure reports point estimates and 95% confidence intervals representing the main effect of our intervention on disaggregated components of the diversity scales. Point estimates and standard errors (in parentheses) are reported along each estimate.

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 A14 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 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

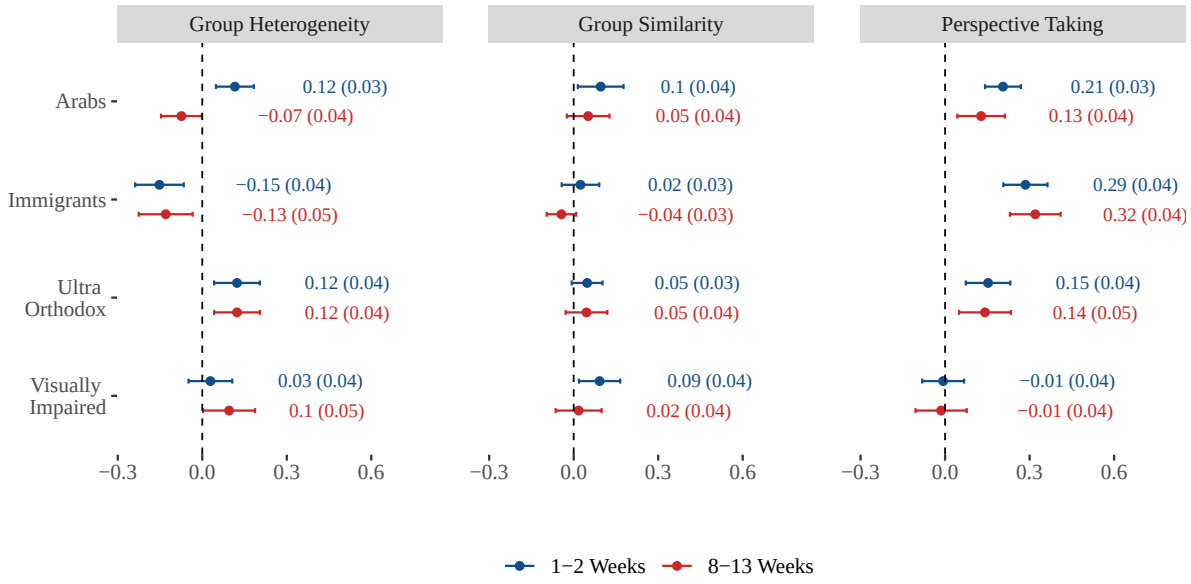


Figure A13: **The intervention had mixed effects on disaggregated components of the group similarity and group heterogeneity index, and consistent effects ofon the disaggregated components of the perspective-taking index.** This figure reports point estimates and 95% confidence intervals representing the effect of our intervention on disaggregated components of the mechanism indices. Point estimates and standard errors (in parentheses) are reported along each estimate.

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.

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 mech-

anism underlying the main effect of our intervention.

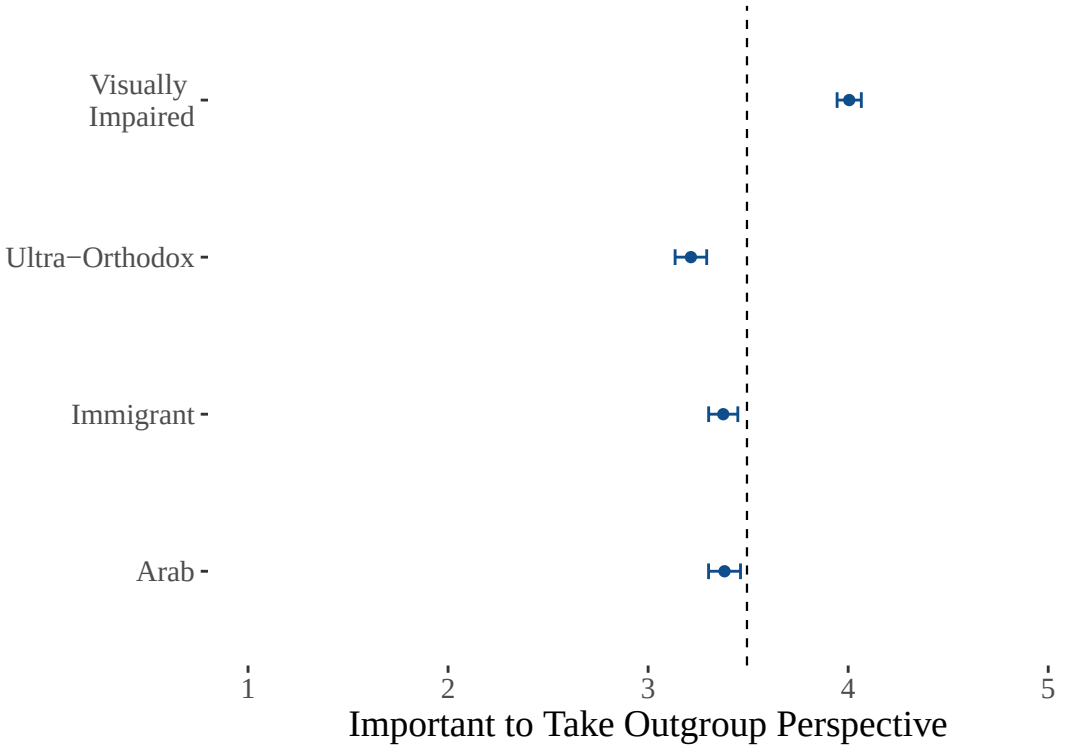


Figure A14: **Students' pre-treatment perspective-taking measures are substantively higher for visually impaired measures when compared with all other social group measures.** This figure reports means and 95% confidence intervals representing students' pre-treatment average agreement that it is important to take a specific outgroup's perspective. The dotted line represents the average value for all groups combined.

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 on the post-treatment thermometer index. Before doing so, in Figure A15 we implement a diagnostic test proposed by Hainmueller et al. (*40*) 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. (*40*). We report these moderation analyses in

Figure A16.

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 A16 are statistically indistinguishable. We thus interpret our evidence to suggest that treatment was similarly effective on students with varying levels of pre-treatment prejudice. To the extent to which heterogeneity in response to treatment exists, it is small and hard to estimate precisely with our current sample.

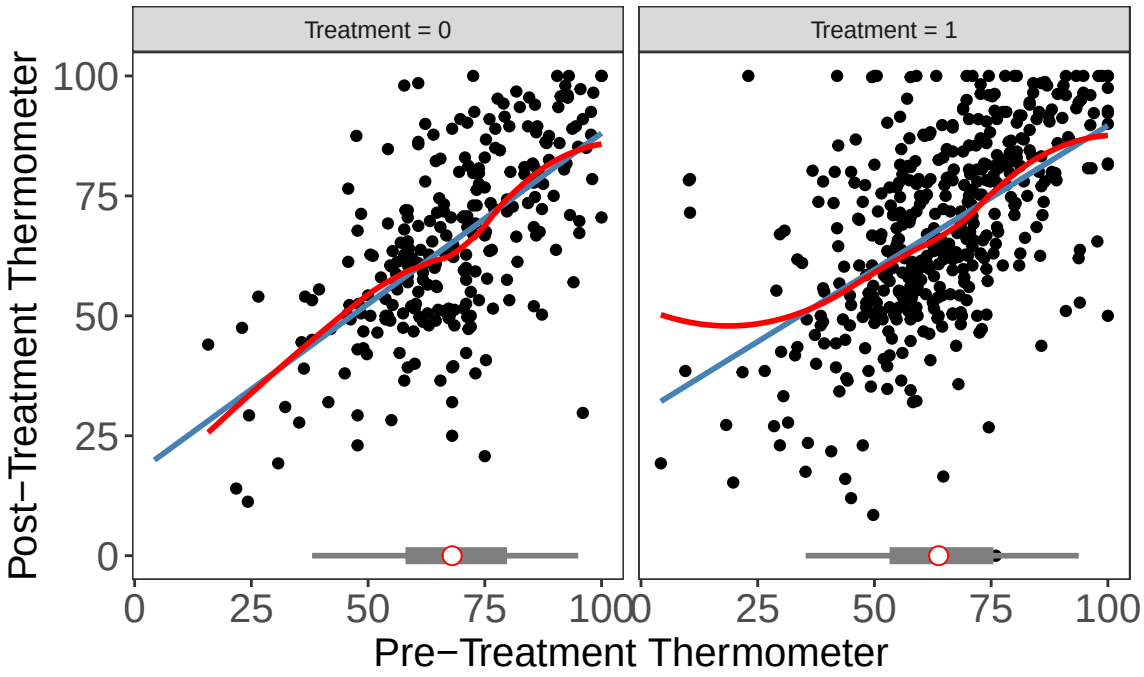


Figure A15: **Interaction term diagnostics for Study 2**. This Figure plots the post-treatment thermometer index over the pre-treatment thermometer index for treatment and control groups.

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

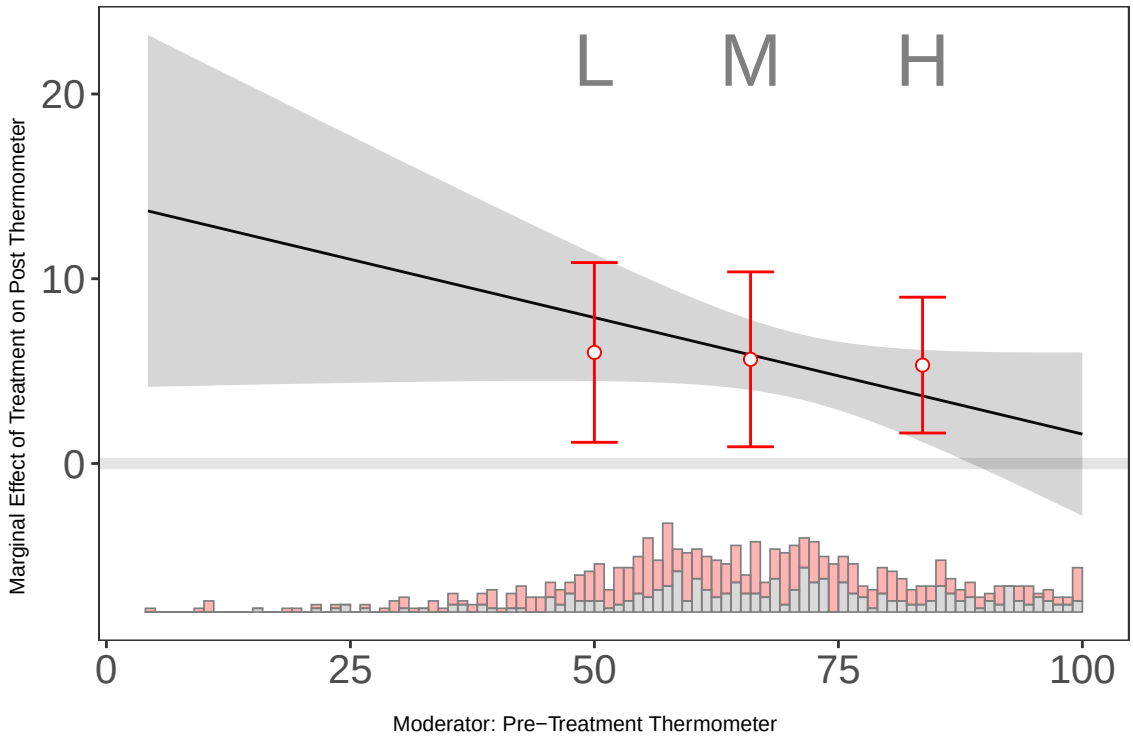


Figure A16: **In Study 2 pre-treatment attitudes do not moderate our interventions' average treatment effects.** This plot reports that the average treatment effect of our intervention conditional on levels of our pre-treatment thermometer index using the binning estimator proposed by Hainmueller et al. (*40*).

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.

We address this concern in Figure A17-A18 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

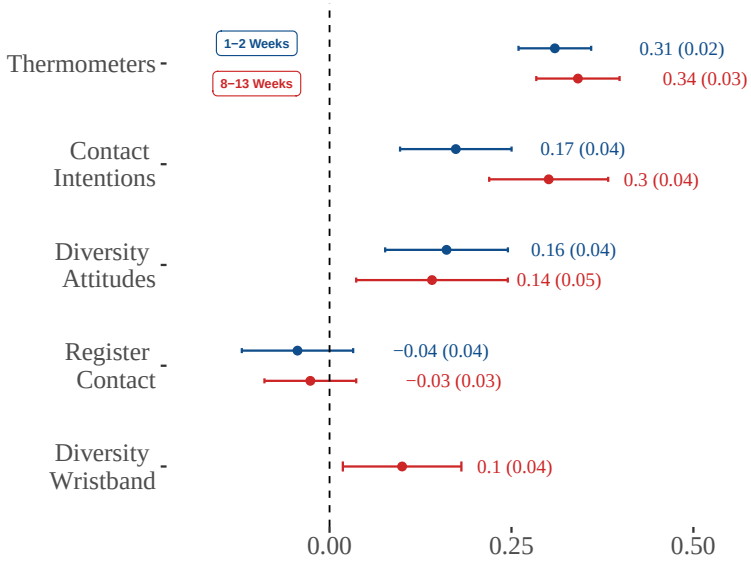


Figure A17: **The main pattern of results is consistent when focusing on outcomes only among respondents that participated in both post-treatment survey waves.** This figure reports point estimates and 95% confidence intervals representing the effect of our intervention on main outcomes amongst students participating in both post-treatment survey waves. Point estimates and standard errors (in parentheses) are reported along each estimate.

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.

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 (*36*). We take the first step in this direction by showing that our main results replicate

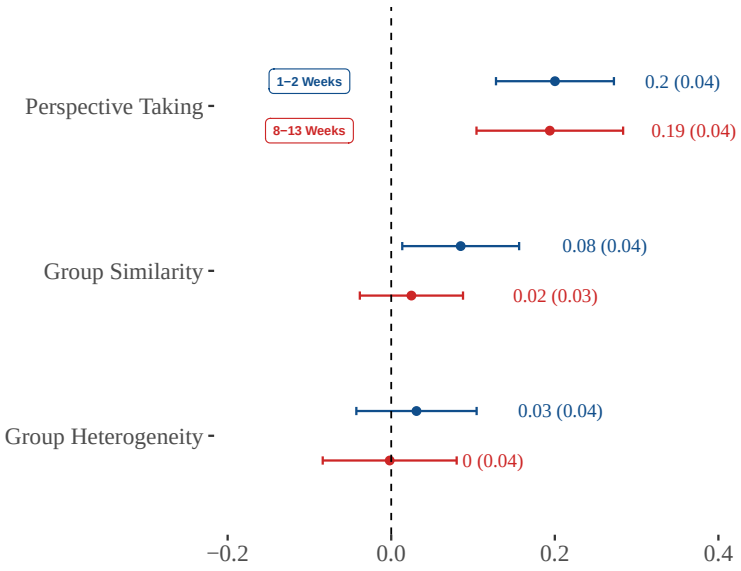


Figure A18: **The main pattern of mechanism results is consistent when focusing on measures only among respondents participating in both post-treatment survey waves.** This figure reports point estimates and 95% confidence intervals representing the effect of our intervention on mechanism outcomes amongst students participating in both post-treatment survey waves. Point estimates and standard errors (in parentheses) are reported along each estimate.

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 (*37*), 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.

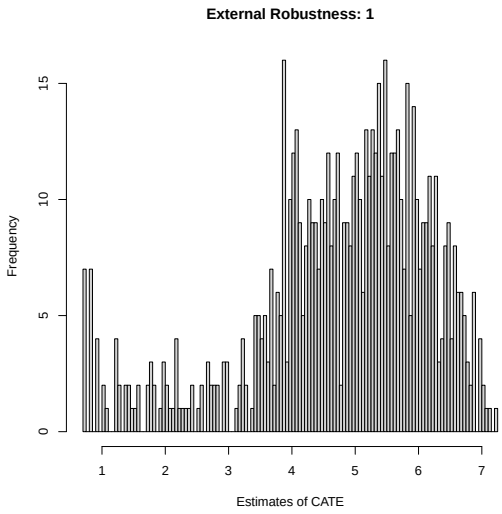
It is important to note that the virtue of this measure depends on the theoretical and practical relevance of covariates used to estimate the CATE. In our case, we employ measures for a student's block, gender, age, and pre-treatment outcome measures. In that sense, as we see it, the main virtue of this 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 A19. 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,

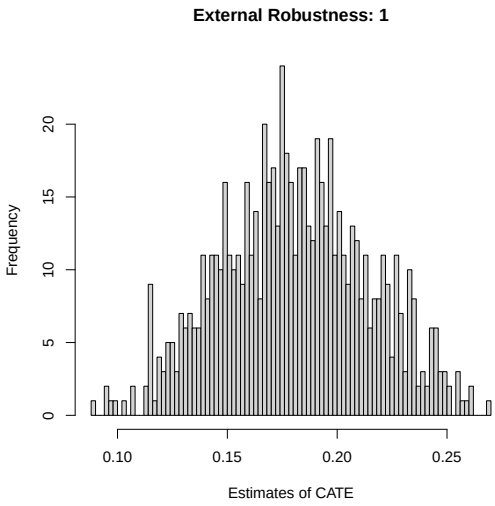
as well as a set of pre-treatment covariates that might generate a degree of heterogeneity in response to treatment. The covariates include respondents' gender, age, experimental block, and all pre-treatment measures of prejudice. The four plots reported in Figure A19 provide reassuring evidence regarding 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 Figure A19 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.

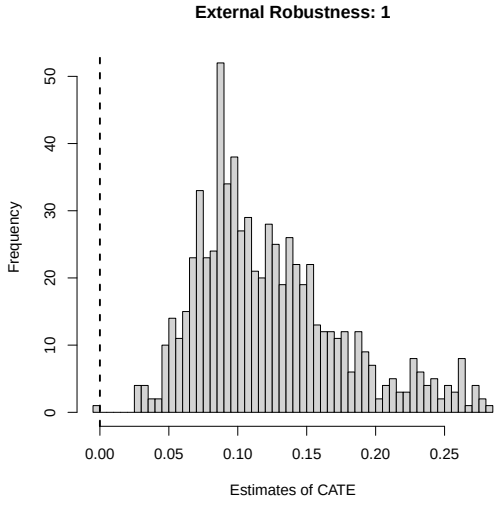
Thus, given our CATEs, re-weighting our sample to resemble some new target population of interest that is more (or less prejudicial), might reduce the size of our estimates, but is unlikely to explain away our average treatment effects. We construe results from this exercise as encouraging with regards 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.



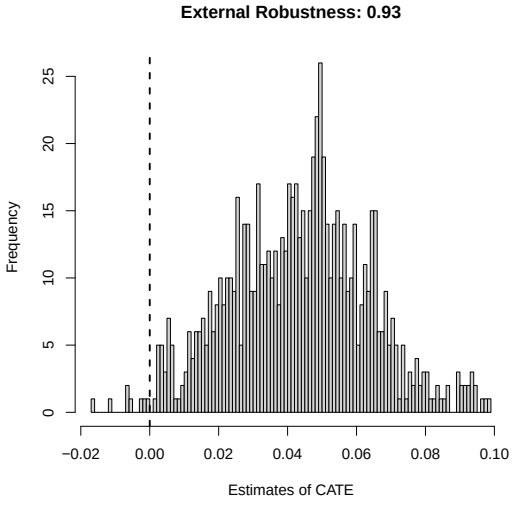
(a) Thermometers Index



(b) Contact Intention Index



(c) Diversity Attitudes Index



(d) Diversity Wrist Band

Figure A19: **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).