Breaking Down Bullying: Empathy, Social Networks, and Adolescents*

Qinyou Hu†

November 27, 2023

Updated frequently. Click here for the latest version.

Abstract

This paper examines the formation of a specific non-cognitive skill – empathy – and its role in determining bullying behavior with a focus on social networks. The analysis centers on a parent-directed empathy-fostering intervention, which successfully increased empathy levels and reduced bullying among students. To disentangle the mechanisms underlying these findings, I develop and estimate a structural model of empathy development, network formation, and bullying decisions. The analysis reveals that 32% of the observed reduction in bullying is attributed to empathy-induced alterations in social networks. Policy counterfactuals show that social network information is valuable. Notably, targeting students based on popularity can lead to up to a 7.5% further reduction in bullying compared to targeting students randomly. Moreover, targeting bullies' friends is more effective than targeting bullies directly. This insight holds promise for refining the efficacy of anti-bullying initiatives, which often focus more on bullies, and highlights the potential of reshaping social networks to mitigate violent behavior among adolescents.

JEL codes: I21, J13, J24

Keywords: Human capital, Social interactions, Violence, Bullying, Field experiment

*I am grateful to Flávio Cunha, Rossella Calvi, Isabelle Perrigne, Xun Tang, and Matthew Thirkettle for guidance and support. I also thank Eric Auerbach, Charles Becker, Marianne Bitler, Vincent Boucher, Isabelle Broca, Christina Brown, Brach Champion, Emily Cook, Victor Delgado, Eric Edmonds, Eliana La Ferrara, Jeremy Fiel, Corrado Giulietti, Alan Griffith, Randi Hjalmarrson, YingHua He, Clemence Idoux, Natasha Jha, Hanchen Jiang, Anne Karing, Kristin Kleinjans, Yunmi Kong, Marcos Lee, Jessica Leight, Lixiong Li, David Liebowitz, Xiaodong Liu, Brendon Mcconnell, Chris Mills, Dimitrios Nikolaou, Fred Oswald, Cynthia Osborne, Mallesh Pai, Alexa Prettyman, Daniel Prudencio, Todd Pugatch, Andrea Salvati, Anya Samek, Shihan Shen, Ketki Sheth, Robin Sickles, Gabriela Smarrelli, Petra Todd, Jackie Wahba, Siyi Wang, Yiming Xia, Xi Yang, Fisher Yu, Rui Zeng, Yves Zenou, and participants at PacDev 2023 conference, Chinese Economists Society North American Conference, North America Summer Meeting of the Econometric Society, NBER Summer Institute, University of North Texas, University of Utah Department of Family & Consumer Studies, 2023 APPAM Fall Research Conference, SEA Annual Meeting, and Rice Economics Department Brown Bag Seminar for valuable comments and suggestions. This paper was supported by the Chiang Ching-kuo Fellowship and the Rice Social Sciences Research Institute Dissertation Grant. All errors remain my own.

†Department of Economics, Rice University. Email: qinyou.hu@rice.edu.

1 Introduction

A growing body of research across various social science disciplines has studied the implications of fostering non-cognitive skills throughout an individual's lifespan (Cunha and Heckman, 2007; Heckman and Kautz, 2012; Heckman et al., 2006; Jones et al., 2015; Kautz et al., 2014). The significance of non-cognitive skills extends beyond academic and workplace success, encompassing crucial implications for prosocial behaviors and interpersonal relationships, ultimately contributing to a better society (Alan et al., 2021a; Deming, 2017; Heckman et al., 2023; Kosse et al., 2020). Adolescence is a sensitive period of social skill development (Deming, 2022; Steinberg and Morris, 2001; Steinberg, 2014). Moreover, adolescents benefit from social skills. With increased exposure to peers during adolescence, social skills assume a more pronounced role (Frith and Frith, 2007; Lam et al., 2014; Orben et al., 2020). Given that adolescents are in a phase of development characterized by an intense desire for peer recognition, social skills, along with the social environment, can significantly influence their behavior, potentially mitigating violent or risky tendencies (Chein et al., 2011; Steinberg, 2017).

However, the specific mechanisms by which social skills can help prevent or mitigate violence among adolescents remain unclear. This study aims to fill this gap by examining the role of empathy (the capacity to sense others' emotions and imagine their thoughts or feelings; Davis, 1996; Hogan, 1969) in determining bullying in schools.[1] Roughly one-third of youth fall prey to bullying worldwide, with far-reaching consequences (UNESCO, 2019).[2] Despite its prevalence and severe effects, bullying has yet to receive the level of public awareness and policy attention it deserves, especially in developing countries.

In this paper, I show that fostering empathy can effectively reduce bullying. I also provide a rigorous illustration of the underlying mechanisms by integrating evidence from the field with insights from a structural model. Besides increasing one's concerns about others' feelings and perspective taking (*individual human capital effect*), I show that empathy can help prevent bullying by reshaping adolescents' social networks and the subsequent influence from peers within those new networks (*social effect*). Understanding the social aspect of bullying and how empathy shapes social networks is critical to inform policy. In low to middle-income countries, where resources are limited or program uptake is often low, a network theory-based targeting design may be particularly appealing.[3] Moreover, the proposed model can be broadly applied

[1] Empathy has been proven to be a powerful solution to address aggressive behavior like bullying and is widely recognized as a vital social-emotional skill that can be cultivated (Decety, 2010; Hodges and Myers, 2007).

[2] The detrimental effects of bullying on mental health, academic performance, labor market outcomes, and even the emergence of suicidal tendencies have been well-documented (Brown and Taylor, 2008; Hinduja and Patchin, 2010; Klomek et al., 2007; Sarzosa, 2021; Sarzosa and Urzúa, 2021).

[3] Empirical studies have established the efficacy and extensive utilization of network theory-based targeting interventions across diverse domains, encompassing technology adoption (Beaman et al., 2021), education (Bennett and Bergman, 2021), labor markets (Beaman and Magruder, 2012), anti-poverty initiatives (Haushofer et al., 2022), immunization (Banerjee et al., 2019), and others.

to understanding how adolescent skill development mitigates socially undesirable behaviors beyond bullying from a social network perspective.

The analysis centers on a cluster-randomized controlled trial (RCT) on empathy development that I conducted in two large middle schools in China with around 2,200 students and their parents (Cunha, Hu, Xia and Zhao, 2023). The intervention takes the form of a parent-directed parental involvement program that includes coaching and education focused on empathy, as described in Section 3. The parent-directed design, motivated by parents' central role in children's social-emotional development, sets this experiment apart from most existing anti-bullying programs.[4] The first-hand program implementation allowed the collection of comprehensive measurements pertaining to standard psychometric measures of empathy skills (Davis, 1983), friendship networks, characteristics of friends, and detailed bullying behavior, which are challenging to obtain in observational data. The intervention successfully enhanced students' empathy levels and concurrently mitigated instances of bullying (Cunha et al., 2023).[5] The present paper takes a step further and investigates the channels through which empathy can successfully reduce bullying among adolescents. It does so while explicitly accounting for the social dimension of empathy and bullying behavior.

I start by showing that the intervention effectively reshapes the friendship network structure of treated students, demonstrating empathy's social effect in Section 4. Specifically, students in these classes on average exhibit a higher level of connectedness measured by in-degree centrality, namely the number of nomination friendship links received, with an increase of 0.2 units (6.5%). Moreover, they develop more friendships with non-bullies, reflected by an increase of 0.22 units (8.6%), relative to those in the control group. I also find that empathy plays a dominant role in predicting changes in network structures. Empathy consistently demonstrates a stronger predictive power and remains statistically significant even after controlling for the treatment indicator and other social-emotional skills.

While the program evaluation has provided valuable insights into empathy's human capital and social effects, it falls short in quantifying their relative contributions to the reduction of bullying among adolescents. It also fails to capture the social aspect of bullying that always involves peer spillovers. To overcome these limitations, I develop a structural model that accounts for the data structure and the experimental evidence discussed so far. As shown in Section 5, in my model, utility-maximizing parents choose whether to participate in the empathy-fostering program, and children make decisions regarding their interactions with peers in the classroom including friendship formation and bullying behavior. Students' empathy, which I model as a non-cognitive skill, evolves according to a human capital production function, taking their baseline empathy skill endowment and parents' participation status as inputs. Students make

[4] See Table E1 in Cunha et al. (2023).

[5] Specifically, the intervention led to a decrease in the prevalence of bullies by 4 percentage points (23%) and victims by 5 percentage points (12%).

friends according to a dyadic link formation model and decide on bullying involvement following a social interaction type model. The key feature of the model is that empathy plays a twofold role in shaping the bullying outcome: first, individual empathy levels directly influence the private utility of their own bullying efforts to capture empathy's human capital effect, and second, empathy affects the probability of friendship links within the class, thus governing the process of social network formation to capture empathy's social effect.

In Section 6, I leverage the random assignment of the program and unique individual-level network data to address two key identification challenges in my structural model. First, I introduce a correlated error term structure between parents' unobserved preference heterogeneity and the empathy formation shock, akin to the classical self-selection model (Heckman and Sedlacek, 1985; Roy, 1951), to account for the potential endogeneity of parents' participation decisions in children's empathy formation. The random assignment nature of the intervention, coupled with the observation that no participants in the control group participated in the program, enables me to estimate the parents' participation decision using the treatment group and verify the validity of the error term assumption using the control group. Second, I tackle the potential endogeneity of students' friendship networks in bullying outcomes by allowing for a correlation between the individual fixed effect when a student nominates others as friends and the unobserved preference shocks in bullying effort utility. I then use a two-stage IV estimation approach following König et al. (2019) with the help of individual-level network data.[6]

The estimates of the empathy production function reveal that program participation leads to a higher marginal return of baseline empathy in forming follow-up empathy, but it does not induce a significant level change. This suggests that students' baseline empathy level plays a central role in driving the program's effect on empathy. It turns out that children's bully status and friendship information do not significantly affect parents' participation decisions, thus further verifying the model's sequential set-up. Moreover, parents of children with better academic outcomes exhibit a higher likelihood of participation, while those who have skipped exams and have poorer mental health outcomes are less likely to participate, emphasizing the selection of participation. I cannot reject the null hypothesis that the estimates using the control groups are statistically the same as those using non-participants in the treated group, which helps validate the empathy model assumptions.

In the network formation model estimates, an individual's empathy level emerges as the primary determinant of changes in friend networks, even after accounting for the baseline network structure. I also find that differences in empathy between individuals and their peers play

[6]In the first stage, I predict the network structure based on dyadic characteristics and empathy level. In the second stage, I construct instruments using the predicted adjacency matrix. Specifically, I utilize the characteristics of an individual's predicted indirect connections as excluded instruments for friends' bullying behavior to handle the "reflection problem" of Manski (1993). The intuition is that the characteristics of peers of peers solely influence an individual's activity through their peers' activity (Bramoullé et al., 2009).

a minimal role in predicting link formation. As for the magnitude, a one standard deviation (SD) increase in empathy corresponds to an additional 0.14 friends per student, on average. When considering a class size of 47 students, which is the average in the study sample, this effect translates to approximately 6.4 more friends per class. Finally, consistent with the experimental evidence in Cunha et al. (2023), I find that empathy significantly reduces bullying. Students exhibit a preference for conformity, and the positive "spillover" of peer bullying behavior influences one's own bullying behavior. Furthermore, male students exhibit a higher likelihood of engaging in bullying than females.

Based on the model estimates, I then proceed to decompose the channels through which empathy can change bullying behavior and quantify their relative importance. I find that the social effect of empathy accounts for 32% of the total effect on bullying behavior. In other words, empathy affects bullying both through individual human capital and social effects, with the latter accounting for 47%, which is almost half of the former's impact. The decomposition analysis underscores the economic significance of empathy's social dimension in explaining bullying reduction.

In Section 7, I use the model estimates to evaluate the impact of alternative targeting schemes on bullying reduction. I simulate bullying change in equilibrium under various targeting groups and compare them with random targeting. This exercise was motivated by that students' friendship networks can serve as proxy measures of their social status, and peer effects of bullying have been proven to be non-negligible. Consequently, a targeting experiment combining network formation may result in a more effective reduction in bullying than random targeting. Results from the policy counterfactuals show that targeting based on students' popularity always yields higher reductions in bullying than treating students randomly. At its maximum, popularity-based targeting leads to a 7.5% further reduction in bullies.[7]

Interestingly, I find that targeting bullies' closest friends is more effective than targeting bullies directly. Delving into the data patterns, I find this is primarily driven by the observation that bullies tend to be less popular than their closest friends. Conversely, those students identified as bullies' closest friends have higher social status and display more desirable personality traits than the bullies themselves. So, targeting bullies' closest friends generates a higher social effect than random targeting, and importantly, the effect will inevitably get transmitted to bullies.

The counterfactual analysis discussed above is highly informative for the development of future anti-bullying programs. On one hand, it highlights the limited effectiveness of Zero Tolerance policies (see Borgwald and Theixos, 2013). On the other hand, it emphasizes the importance of leveraging social network information, particularly within the school environ-

[7]This set of results resonates with the importance of targeting social referents as the optimal treatment assignment to change harmful norms and harassment behavior found in Paluck and Shepherd (2012) and Paluck et al. (2016).

ment. Policymakers or practitioners can weigh the trade-off between the gained efficiency from popularity-targeting and the additional costs associated with collecting social network data. However, the findings pin down students' popularity as a powerful characteristic for effective targeting experiments. Popularity measure is more readily obtainable than other network centrality measures, such as eigenvector centrality. School principals and class teachers may already possess insights into students' popularity, alleviating the necessity for extensive and costly complete network data collection. Moreover, the latest advancements in network data collection methods, such as the "random matching within sample" method (Conley and Udry, 2010)[8] and the aggregated relational data approach (Breza et al., 2020),[9] offer promising avenues for accessing pertinent information regarding students' popularity. Collectively, we anticipate that these advancements would enhance the efficiency and feasibility of conducting network-based targeted experiments in future implementations.

In summary, this study is among the few to probe the causal impact of empathy on bullying by innovatively combining an RCT with a structural model of human capital development, network formation, and individual behavior. It also demonstrates that a parent-directed program designed to foster empathy development in adolescents can reshape the structure of their friendship networks and help them engage in desirable social behaviors. Notably, the networks formed as a result of the program exhibit higher levels of friendliness, as treated students develop more friendships with non-bullies and the prevalence of isolated students diminishes. This finding motivates the necessary consideration of factoring in the social environment when studying adolescent development. Critically for policy design, this study illustrates how to enhance the effectiveness of anti-bullying programs through more precise targeting methods that consider the influence of social networks.

The remainder of the paper is organized as follows. Section 2 discusses the contribution to the relevant literature. Section 3 introduces the experimental setting and key measures for this study. Section 4 presents motivating evidence on experiment-driven changes in network structure and peer effects of bullying. Section 5 develops a structural model with empathy production, network formation, and bullying decisions. Section 6 discusses model identification and estimation, and provides the model estimates. Section 7 reports counterfactual analysis to show the effectiveness of targeting different groups. Section 8 concludes.

[8]Each respondent is asked questions about their relationship ties with a certain number of randomly selected individuals from a larger sample, for example, "Do you know person X?"

[9]They ask the following type of question "How many of your friends have trait X?" to obtain network information and propose a network formation model accordingly.

2 Related Literature

This study contributes to four strands of literature. First, this paper connects to the rich literature on human capital formation. The seminal works of Becker (1962), Becker and Tomes (1979) and Becker and Tomes (1986) set the foundation and first introduce the concept of human capital, which refers to the skills, knowledge, and abilities that individuals acquire through education and training, within the economics literature. Researchers have since developed models to analyze the formation of various skills, including cognitive and non-cognitive skills, using economic frameworks. They have also examined the input factors that contribute to the formation production function (Agostinelli et al., 2020, 2022; Attanasio et al., 2020a; Del Boca et al., 2017, 2013; Bono et al., 2016; Cunha and Heckman, 2008; Cunha et al., 2010).[10] Only recently have researchers started to emphasize the importance of non-academic or higher-order skills (Deming, 2017, 2022). This paper investigates an understudied yet vital social-emotional skill – empathy. Furthermore, it studies the development of skills during the adolescence period, which differs from the existing literature's predominant focus on early childhood. This study examines empathy jointly with friendship networks, suggesting that network effects should be considered in understanding adolescent development.

Second, this paper connects to the literature on parental involvement program evaluation. There is limited evidence on the causal impacts of parental involvement programs on children's peer networks. Existing studies on parenting program evaluation tend to focus more on improving cognitive and non-cognitive abilities for the children (Attanasio et al., 2020b; Barrera-Osorio et al., 2020; Cappelen et al., 2020; Doyle, 2020; Wang et al., 2023).[11] More recently, research has explored the impact of neighborhoods on human capital formation using experimental data (List et al., 2023).[12] In contrast, my paper examines the return to parental involvement on a different margin – the social environment of children captured by their friendship network. Moreover, participating in the program improves students' social-emotional skills, especially empathy, and subsequently increases their popularity within the class.[13] Understanding the "virtuous circle" of non-cognitive skills and social capital can help understand

[10] For instance, Agostinelli et al. (2022) explore the roles of peer interactions and parental involvement in children's education to understand how school closures during the pandemic affected educational inequality. Agostinelli et al. (2020) extend the skill formation model by connecting parenting with peer effects among children.

[11] This strand of literature also tends to focus on the effect of parental involvement programs during early childhood, which ranges from 0 to 6 years old (e.g., Attanasio et al., 2020b; Doyle, 2020; Wang et al., 2023). See also Cunha et al. (2021).

[12] List et al. (2023) use spatial distance to explore spillover effects on skills between children attending a Pre-K program in Chicago and those in the control group. Another related strand of literature uses residential movers to identify neighborhood effects using observational data (e.g., Chetty et al., 2016; Katz et al., 2001). See Graham (2018) for a review.

[13] This finding is consistent with Alan et al. (2021b), who show that a classroom-level intervention on perspective taking, i.e., cognitive empathy, in Turkish primary schools increases friendship and support ties in classroom networks.

the mechanisms of peer effects.

Third, this paper digs into the causes of school bullying, one of the most common violent behaviors among adolescents, through a friendship network angle. Existing literature on bullying summarizes potential reasons into various individual and contextual factors (Álvarez-García et al., 2015), distant parent-child relationships (Li et al., 2019), and family socioeconomic status (Tippett and Wolke, 2014). Most studies from psychology only provide correlational evidence that high impulsiveness and low empathy may be linked to bullying behavior (Cook et al., 2010; Jolliffe and Farrington, 2006). In contrast, this paper investigates the causal impact of empathy on bullying. It also emphasizes empathy's role in shaping the social environment, which has been examined narrowly in the literature thus far. Moreover, I utilize unique data on detailed peer network information to explore the non-linearity of peer effects (e.g., Battaglini et al., 2017; Booij et al., 2016; Boucher et al., 2022b; König et al., 2019) rather than only the leave-one-out average. The empirical test of homophily in adolescents' bullying behavior further adds to the latest discussion on homophily over "malleable" characteristics (Jackson et al., 2022).

Lastly, this study provides insights into network-based interventions to address practical issues. Thus, it connects to one of the benchmark papers by Banerjee et al. (2019), who study the diffusion of information via a micro-finance experiment in Indian rural villages. They construct new centrality measures and document faster diffusion when technologies are seeded with people who are central in the network rather than broadcasting information widely. In the context of education-related intervention, Islam et al. (2021) conduct a two-year field experiment of free provision of an after-school tutoring program in primary schools in rural Bangladesh based on the centrality measure of the students. The treatment effect analysis shows that targeting the most central students leads to more improvement in study outcomes than random targeting and thus proves to be a more cost-effective way. This study adds to this by showing that network structures can also be effective in preventing non-academic outcomes like bullying. It takes into account potential changes in the network structure that may occur in response to interventions, which is relevant to recent papers on policy evaluations in the presence of between-unit interactions (e.g., Comola and Prina, 2021; Griffith, 2022b). This paper also relates to the "key player" literature (e.g., Ballester et al., 2006; Calvó-Armengol et al., 2009; Lee et al., 2021; Peng, 2019; Zenou, 2016). The targeting intervention exercise underscores the importance of indirectly targeting bullies' friends' circles or involving high-status adolescents as leaders to facilitate bullying prevention through positive peer influence, which can be harnessed by the school and community leaders to inform policy design.

3 Empirical Setting and Data Description

3.1 The Parental Involvement Program on Empathy Development

The empirical analysis relies on data collected in an RCT on a parent-directed empathy education program. The program was conducted in two large middle schools in China involving around 2,200 students and their parents. The main goal of the program was to encourage parental involvement and foster adolescents' non-cognitive skills, especially empathy. I briefly discuss the curriculum and settings below. For more details, one can refer to Cunha, Hu, Xia and Zhao (2023).

Curriculum The embedded curriculum includes coaching and education on empathy. The program lasted for four months and the materials were designed based on a monthly theme. In sum, the four monthly themes cover empathy, perspective-taking, respect for uniqueness (i.e., recognizing the value of various personality types), and the role of social-emotional skills in maintaining relationships with others. We used daily activities, such as reading short articles and watching movies, as the key learning methods. All the tasks were empathy-oriented.[14] During the four-month intervention, we provided two biweekly reading materials on the first day of the corresponding week and one movie task on the first day of each month. We also encouraged parents and their children to leave a brief reading reflection to share their thoughts, and this assignment is optional. The curriculum was designed with the aim of cultivating adolescents' empathy through family education. We targeted parents to help them learn empathy and positive parenting skills with their children. As predicted by the simulation theory from the psychology literature (Decety and Jackson, 2004; Preston and De Waal, 2002), children would eventually develop empathy through interactions with their parents.

Delivery We used a stratified randomization design and randomly assigned the 48 classes to the treatment or the control group. All the families in the treatment classes received the intervention, while the control classes received no information during the intervention except invitations for the follow-up surveys. The delivery of the intervention used a social media app – *WeChat* [15] – to trace the enrollment and participation for the study sample. We uploaded biweekly and monthly materials through a platform *Daka* on the app, and all the program materials were exclusively accessed through the app. The platform automatically comes with a

[14] Although we collected rich measures on bullying behavior, the program was mainly advertised as an empathy development program and all the curriculum content was about empathy. Empathy development, which is irrelevant to the exam-based evaluation education system, is not prioritized by the school system in China. In fact, even in many developed countries, the main focus has been on raising test scores. This is a common "unintended" consequence of test-based accountability policies (Loveless et al., 2005). Therefore, we anticipate that the marginal emphasis on empathy development may contribute to the issue of non-compliance. However, the marginality helps alleviate reporting issues regarding empathy skills to some extent. Additionally, the absence of information about bullying behavior or prevention measures also helps minimize the Hawthorne effects.

[15] *WeChat*, similar to *Facebook* or *WhatsApp*, is the most commonly used social media platform in China.

check-in feature so that we can check the participation records of each participant.

3.2 Uniqueness of the Program and the Data

The design of the parental involvement program allows me to study the role of friendship networks in determining how empathy as a social-emotional skill works on bullying reduction.

First, the delivery of the intervention requires students to participate and learn empathy with parents at home rather than at school. This setting minimizes the direct impact of the program on school life, such as the teacher and student network, which avoids confounders and establishes the empathy effect on the student network. During the intervention, teachers were only informed that it is a parental involvement program on empathy, and the only task they were required to do was to help us forward a biweekly reminder message with the access link to the platform. We thus expect the teacher effect to be minimal. There are also two additional pieces of evidence that can support the argument. First, empathy skill is a marginal skill in schools due to the emphasis on test scores and the prevail of test-based policies.[16] Teachers do not have further incentives to encourage participation as participation is not linked to their salary. Second, through the enrollment data, I observe no teacher ever registered for the courses, and hence, they had no access to any material.

Second, we also expect a minimum direct impact on the student network due to the nature of family activities. Parents know little about their children's entire friendship networks,[17] and the likelihood that parents intervene in relationships is low — from the baseline data of the student survey, only around 20% of students reported that their parents ever intervened in their friend-making decisions.

Lastly, the random assignment provides the exogenous variation needed for identifying the causal effect of empathy on bullying reduction. In addition, the richness of the individual-level friendship network data allows for more flexibility to decompose the empathy effects at a more granular level.

3.3 Key Measures

The data for this study include measures from two rounds of surveys: the baseline survey collected in January 2021 and the follow-up survey collected at the end of June 2021.[18] The key measures used in this study are introduced below:

[16]Unfortunately, this is an issue not only prevalent in the developing world, such as China, but also in the developed world, such as the United States (Chen et al., 2021; Ryan et al., 2017).

[17]The later analysis of participation decision also resonates with this finding — Table 5 shows that only child ability measures are determinants of participation while network features are not relevant.

[18]Data collection relied on the administration of standardized student questionnaires and was realized before and right after the end of the program under the supervision of class teachers. Students had the right to discontinue at any point in time or skip sections of the questionnaire part. In case students chose not to participate, they were asked to work on a worksheet provided by the teacher who was present during the entire data collection phase.

School Bullying In both surveys, we asked students' detailed bullying behavior covering five domains: (1) threatening/verbal abuse for verbal bullying, (2) hitting/kicking for physical bullying, (3) lying and spreading rumors for social bullying, (4) social isolation, and (5) abusive or hurtful texts online as cyberbullying. We used multiple questions with specific examples to alleviate the concern of misreporting bullying behavior caused by misunderstanding. Comparing results between different domains also gives us more confidence in handling the Hawthorne reporting effects. We collected detailed frequencies of the five types of bullying behavior, as well as whether they were victims of any event during the intervention semester.

Regarding the misreporting issue, Cunha et al. (2023) carefully test and argue that students' self-reports of bullying behavior may be superior to other resources researchers use to avoid misreporting issues, such as teachers' reports or administrative records. The subtlety of school bullying makes teachers know little (Hazler et al., 2001; Huang et al., 2013). Even if teachers are aware, they may have higher incentives to misreport it than students due to potential negative effects on their performance evaluation by the school.[19] Moreover, even if teachers do report aggressive behavior, their reports are more likely to be influenced by the Hawthorne effect compared to students' self-reports, because adults are more sophisticated than adolescents as well as their expertise in education. In contrast, we expect that students in both groups have no differential incentive to misreport, as bullying is merely mentioned in the program. Last, we also verify the robustness of our results to the misreporting issue by applying the misclassification method in Cunha et al. (2023).

For the empirical analysis, I construct an indicator and a continuous measure for students' bullying behavior: (1) the "repetitive bully" indicator equals 1 if students ever engaged in at least one of the five bullying events more than once to capture the repetitive nature of school bullying incidents [20] and (2) a bullying score, calculated using factor score, to capture bully intensity.[21]

Friendship Network We followed the data collection methods of National Longitudinal Study of Adolescent to Adult Health (Add Health) to collect the friendship information for each student in the study sample at both baseline and follow-up.[22] We asked students to name

[19] We requested a list of records regarding aggressive behavior from teachers, but no incidents have been reported. We attribute this outcome to several factors: (1) students generally avoid bullying others in the presence of teachers, (2) teachers may be too burdened to report non-academic behavior within the current test-oriented education system, and (3) reporting aggressive behavior goes against teachers' incentives, as it could potentially lower their bonuses.

[20] For robustness checks, I also show results in the Appendix as a supplementary analysis using the "bully once" indicator, which equals 1 if students engaged in at least one of the five bullying events for at least once. The pattern of the results remains the same.

[21] To alleviate the concern of measurement error issue of the reported bullying behavior, a factor model that maps the reported bullying frequencies to their latent bullying efforts is applied using a Two-parameter Logistic model from Item Response Theory developed in psychology (Lord, 2012). More details are discussed in Appendix Section E.

[22] The Add Health data have been widely used in network and peer effects studies, including Boucher et al.

at most five best friends within the classroom and rank them in order of closeness. In this study, I construct the following network statistics: at the individual level, I use *in-degree, eigenvector centrality, reciprocal link*, and *unilateral link*; at the classroom level, I quantify the degree of homophily using the *Coleman Index* proposed in Coleman (1958) and measure segregation using the number of isolated students within each class. Details about these measures are explained in Appendix Section A.1.

Empathy Skill We used self-reported instruments for students' empathy skill and follow Alan et al. (2021a) to measure two dimensions at the baseline: *perspective taking* and *empathetic concern*.[23] However, our sample students are, on average, about 5 years older than students in Alan et al. (2021a). Therefore, we added another dimension, *prosociality*, in the follow-up, to construct a more valid empathy measure for the adolescents. The new measure is closer to the modified *Interpersonal Reactivity Index* in the psychology literature (Davis, 1983). I follow Anderson (2008) to construct an inverse covariance weighting empathy index to summarize the three sub-components.

Children's Other Outcome Variables To better understand the mechanisms beyond the treatment, we collected other outcome variables for the children, including stress, mental health (measured by CES-D), positive personality, time with parents, and time spent on study and leisure. In the empirical analysis, the inverse covariance weighting indices are constructed for stress and positive personality following Anderson (2008). Appendix Section A.2 discusses these measures in detail.

3.4 Summary Statistics and Balance Test

The original study sample contains 2,246 student responses. After dropping those with invalid network information,[24] the total estimation sample of this study composes of 1,025 students from the control classes and 1,206 students from the treatment classes. Appendix Table I2 reports the attrition rates by comparing with the original student sample with all valid responses. The attrition rate is small and negligible: 0.4% from control groups and 0.9% from treatment groups.

Table 1 displays the summary statistics of the study sample at baseline. Panel A exhibits the results for individual characteristics. On average, students are aged 14, with 53% of them being male, 58% having urban *hukou*, and more than 70% having at least one sibling. Additionally, 38% of the students reported engaging in aggressive bullying behavior at least once, while

(2022b), Boucher et al. (2022a), Bifulco et al. (2011), Lleras-Muney et al. (2020), Calvó-Armengol et al. (2009), Boucher et al. (2023) and many others.

[23] Kamas and Preston (2021) discuss different types of measures of empathy and conclude that a self-reported survey is considered to be a valid measure of empathy.

[24] There exist some cases ($N = 15$) of students' responses where they either gave invalid friends' names or I cannot successfully match the student ID.

approximately 70% of the study sample reported experiencing any of the events at least once. When considering the repetitiveness of bullying behavior, 21% of students are classified as bullies, while 57% are classified as victims.

Panel B of Table 1 reports results for baseline network statistics. On average, each student has 3.1 links to other students. Among these links, 1.3 are reciprocal links, and 1.8 are unilateral links. The average eigenvector centrality is 0.2, suggesting that the friend network is quite loose in the sense that there are few "clubs" that exist within a class. The two groups of classes display similar classroom network patterns. No significant differences are detected. To complement those statistics, Appendix Figure I1 illustrates the distribution of the total number of nomination links received by each student. Students in the study sample receive an average of 3 nominations. I also plot the distribution of the number of isolated students within each class in Appendix Figure I2. On average, 4 to 5 students in the study sample receive no friend nominations in each class.

The individual characteristics as well as network statistics are balanced across two groups. The joint test also verifies the successful randomization.

4 Motivating Evidence

Bullying is a complex societal issue that encompasses various individuals and includes a desire for social status (De Bruyn et al., 2010; Espelage, 2014; Espelage et al., 2003). Meanwhile, empathy plays a vital role as a social-emotional skill, potentially influencing one's social circle. Cunha et al. (2023) shows the effectiveness of the empathy education program in cultivating students' empathy and further reducing bullying, as summarized in Appendix Section B. However, it leaves out empathy's social effect on the friendship network. This study enriches our initial analysis by examining the social aspect of empathy in order to further understand its impact on reducing bullying, thereby providing valuable policy recommendations for preventing violent behavior.

In this section, I detail findings from data exploration. I first present descriptive patterns from the network data. I then discuss the results of evaluating the program's impacts on the network structure. I also lay out a simple model by transforming insights from psychology literature on the empathy effect to help interpret the results in Appendix Section C.

4.1 Descriptive Patterns from the Network Data

In Figure 1, I visualize the friendship network structures of a randomly selected class in our study sample at baseline (Panel A) and follow-up (Panel B). At baseline, students tend to associate with those who share a similar status, and both bullies and non-bullies can be popular. However, at follow-up, the social circles of bullies break down into pieces. These patterns sug-

gest that after the intervention, bullies become less popular within the class. I observe a similar pattern in Appendix Table I3 when examining the association between students' network structure and bully status at baseline and follow-up for the whole sample.

In Table 2, I report the regression estimates, which capture the statistical relationship between students' own bullying perpetration and peers' bullying perpetration using the baseline data. The estimates are robust and stable, with a correlation of around 0.04. It suggests that conditional on the number of friends one has, one additional bully friend is associated with a 4-percentage-point increase in the likelihood of myself being a bully. These estimates imply the presence of peer effects on bullying behavior, and it can be salient among middle school students.

4.2 Program Effects on Friendship Network Dynamics

I estimate the effects of offering the intervention, or the intention to treat (ITT) effects, on the network statistics using the following specification:

$$Y_{ic1} = \alpha + \beta_1 T_c + Y_{ic0} + \tau_s + \varepsilon_{ic}, \tag{1}$$

where Y_{ic1} is a vector of outcome variables for individual i in class c at the follow-up, T_c is the treatment group indicator for class c, which was assigned at the baseline, and τ_s is a set of strata fixed effects, and ε_{ic} is an i.i.d error. I also control for baseline outcome variables Y_{ic0} for a more robust analysis. For all regressions, I cluster the standard error at the class level. Given 48 clusters, which is marginally greater than the rule of thumb, I complement it with Cameron et al. (2008)'s wild cluster bootstrap (WCB) p-values using 9,999 resampling.

Effects on Network Statistics Table 3 reports the ITT effects on various network measures. At the individual level, I find that the offering of the intervention leads to an average 0.2 unit increase in students' in-degree centrality, suggesting that students in treatment classes are more likely to have friends than students in control classes after the intervention. On the other hand, it leads to an insignificant decrease (-0.02) in students' eigenvector centrality. The results suggest that students in treatment classes seem to make more friends with less "important" peers, such as those who were originally isolated or those with very few friends.

How links are formed also matters in network architecture. I find that the program leads to increases in both reciprocal (0.05 units) and unilateral links (0.16 units). In Appendix Figure I3, I show the heterogeneous treatment effects by baseline empathy level and baseline bully status. Results suggest that most of the significant changes in in-degree happen among those with lower baseline empathy skills.

I detect a significant decrease in the total number of isolated students, i.e., those who have no friend nominations, within the treated class. This suggests that the classroom atmosphere

in treated groups becomes more harmonious due to the intervention. I do not find a significant program impact on the degree of homophily at the classroom level, measured by *Coleman index*.

I also find there is a strong correlation between empathy, network status, and bullying behavior. Using the baseline data, I conducted a correlation analysis between empathy and these variables while controlling for demographics. The results, presented in Appendix Table I4, indicate that individuals who possess a greater degree of empathy are likely to receive more peer nominations, experience less social isolation, and have more reciprocal links. These findings are in line with previous research in psychology, which suggests that empathy helps form intimate relationships (Kardos et al., 2017).

Effects on Who Makes Friends with Whom The above analysis explores the treatment effects on the network structure as a whole. The next question pertains to the impact of the intervention on the formation of bully-bully and bully-nonbully ties. Therefore, I conduct subgroup analyses to explore whether the program affects students' choices of friends at the follow-up separately by their baseline bully status. Table 4 reports that, among baseline bullies, there is a 0.19 unit increase in non-bully friends, while there is no change in the number of bully friends. Among baseline non-bullies, receiving treatment leads to a decrease in bully friends, as opposed to a significant increase of 0.22 units in non-bully friends. Therefore, bullies become less "popular," and students are more likely to befriend non-bullies in treatment classes, especially those who were non-bullies at the baseline. These results remain robust to the alternative definition of being a bully, which categorizes bullies as students who have engaged at least once among the five types of events, as shown in Appendix Table I5.

This finding can be attributed to two potential explanations: (i) students' initial friends, i.e., those nominated at the baseline, transitioned to a non-bully status; or (ii) students severed friendship ties with individuals who engaged in bullying behavior and made new friends who are non-bullies. To examine the mechanisms, I construct the aggregate count of bully and non-bully friends falling into two distinct categories: newly formed connections and pre-existing nominations. Subsequently, I employ Eq (1) to scrutinize the treatment effects associated with each category. The findings, presented in Appendix Table I6, indicate that both channels seem to be equally plausible. This finding suggests the need for a comprehensive exploration of the mechanisms, as elaborated in Section 5.[25]

[25] I also explore the effects on the victim network structure in Appendix Tables I7 and I8. Overall, the tables report that both victims and non-victims tend to befriend non-victims after the intervention, and the change is more likely to be driven by the reduction of victims. Moreover, I examine how the program changes friends' characteristics across other dimensions in addition to bully status. More details can be found in Appendix Section D.

5 Model

Two patterns have been documented so far — (1) the changes in students' network structure, especially the number of links students received, due to the empathy program, and (2) the existence of peer effects on bullying behavior. The treatment effect analysis in Section 4 implies the existence of a link between empathy skills and network structures. However, the treatment effect analysis is silent in quantifying the exact contribution of each channel, which is critical to inform policy design. Moreover, bullying often involves multiple agents and needs to be studied as a social event, and the above treatment effect analysis is limited in capturing the equilibrium effect, namely the potential spillover effect of bullying behavior.

Therefore, I proceed to develop a model aiming to (i) unpack the impact of empathy on adolescents' involvement in bullying and (ii) quantify the peer effect in bullying behavior after accounting for empathy and network structure.

Set-up The model is composed of two sets of agents: parents decide on whether to participate in the program, which helps cultivate their child's empathy, and children decide on the interactions with peers in the classroom including friendship formation and bullying involvement.

As an overview, the model includes four interconnected components focusing on the follow-up outcomes in a sequential manner. Notably, my modeling strategy specifically targets the follow-up outcomes to facilitate the decomposition of the causal effect of empathy change on bullying outcomes. The first component encompasses parents' participation decisions by solving a utility maximization problem that governs the evolution of students' empathy. The students' follow-up empathy levels are determined by the interplay between the empathy evolvement equation and the participation decisions made by parents. Subsequently, leveraging the updated empathy levels, students adjust their friendship networks. This network adaptation process considers the new empathy levels and the existing friendship links at the baseline. In the final step, given the updated empathy levels and friendship network, students weigh the private utility of engaging in bullying against the cost associated with deviating from the social norm.

A key distinguishing feature of the model lies in incorporating students' empathy skills into both the private utility of their bullying decisions and the stage of friendship network formation. This integration enables me to capture both the individual human capital and social effects of empathy on bullying.

Model Details In the first step, parents make the decision of whether to participate in the program while considering the constraints imposed by their child's empathy formation process. Modeling the take-up decision is necessary for counterfactual simulations of reassigning the

intervention. Parents would opt to participate if there is an additional gain in their child's empathy, which also outweighs the opportunity cost of participation. Otherwise, they would choose not to participate. Formally, the parents' utility function is as follows:

$$U_i = \gamma_0 + \gamma_1 \left(H^1_{i,1} - H^0_{i,1}\right) + \gamma_2 X_i - \gamma_3 C_i + \vartheta_i, \tag{2}$$

where $H^0_{i,1}$ denotes their child i's empathy at the follow-up when not participating and $H^1_{i,1}$ denotes the new empathy when participating, the vector X_i encompasses the characteristics of both the parents and the children, and C_i represents the opportunity cost of participation depending on parents' income Y_i. Additionally, ϑ_i stands for the unobserved parents' preference shock.

I specify the child's empathy formation function known to her parents as follows:

$$H^0_{i,1} = \beta_0 + \beta_{1,0} H_{i,0} + \varepsilon_i, \tag{3}$$

$$H^1_{i,1} = \beta_0 + \delta + \beta_{1,1} H_{i,0} + \varepsilon_i. \tag{4}$$

In the empathy formation model, I allow participation to have two potential effects on the subsequent development of empathy: The first is the change in level, denoted by δ, while the second is the differential marginal return of baseline empathy $H_{i,0}$ induced by participation, i.e., $\beta_{1,0}$ when not participating and $\beta_{1,1}$ when participating. The parameter β_0 represents the average change in empathy level from baseline to follow-up invariant to participation status. The term ε_i denotes the unobserved shock in empathy formation experienced by student i. I assume that parents' information set includes $\Omega_i = \{X_i, C_i, \beta_0, \delta, \beta_{1,0}, \beta_{1,1}, \varepsilon_i\}$.

Parents choose $P_i \in \{0,1\}$ following the decision rule:[26]

$$P_i = \begin{cases} 0, & \text{if } U_i < 0, \\ 1, & \text{if } U_i \geq 0. \end{cases}$$

In the next step, students form friendship networks in the classroom. Let $\mathbf{W}$ represent the $(n \times n)$ adjacency matrix, where elements are $w_{ij} = 1$ if and only if student i nominates j as a friend and zero otherwise.[27] Note that to fully utilize the information from the students' survey, I adopt a directed network framework. This means that the friendship network matrix $\mathbf{W}$ need not be symmetric.[28]

[26] One may be concerned about the spillover of parents' participation decisions. However, it should not be a big concern in the existing context due to (1) parents having little information about their children's complete friendship network and (2) the take-up rate is not incredibly high and only 40% of treated parents participated.

[27] For this study, each sub-network consists of all students in the same classroom. There is a total of 48 sub-networks, corresponding to the 48 classes included in the analysis.

[28] To clarify, if student i nominates student j as a friend but not the other way around, the corresponding entry w_{ij} in the weighting matrix is set to one, while entry w_{ji} is assigned zero.

To model the formation of the follow-up friendship network $\mathbf{W_1}$, which consists of links w_{1ij}, I use the concept of homophily. Homophily implies that individuals tend to form friendships with others who share similar characteristics (König et al., 2019; Lleras-Muney et al., 2020; Santavirta and Sarzosa, 2019). Moreover, I incorporate the information of previous connections from the baseline friendship networks $\mathbf{W_0}$, which encompass links w_{0ij}, to better capture the changes in the network structure following the intervention. Including the baseline connections aims to enhance our understanding of how the network evolves over time and how the intervention impacts its formation by cultivating students' empathy. Therefore, student i sends a friendship nomination link to j if there is a positive value in doing so, which is represented by the following equation:

$$w_{1ij} = \mathbf{1}\left(w_{0ij}\psi_0 + V'_{ij}\psi_v + X'_j\psi_\gamma + \Xi_{ij} > 0\right), \tag{5}$$

where $V_{ij} = \left(\sum_{\rho=1}^{R}\left(V_{i\rho} - V_{j\rho}\right)^2\right)^{\frac{1}{2}}$ denotes the distance between i and j in R dimensions of characteristics, including the follow-up empathy levels H_{1i} and H_{1j}, and the vector X_j denotes a series of student j's characteristics which also includes j's follow-up empathy H_{1j}. I let Ξ_{ij} denote the unobserved link heterogeneity specific to the student pair $(i,\ j)$.

In the last stage, students make decisions on bullying engagement solving the following utility maximization problem:

$$\max_{b_i} \quad U^s(b_i) = \underbrace{d_i(\mathbf{x})b_i - \frac{1}{2}b_i^2}_{\text{private net benefit}} - \underbrace{\frac{\phi}{2}\left(b_i - \sum_{j\neq i}^{N} w_{1ij}b_j\right)^2}_{\text{conformity cost}}, \tag{6}$$

where b_i denotes the continuous bullying score,[29] $d_i(\mathbf{x})$ represents the private marginal benefit of engaging in bullying affected by characteristics $\mathbf{x}$ including empathy, ϕ denotes the peer effect or the experienced degree of peer pressure, and w_{1ij} is the follow-up friendship link. A scholarly debate exists regarding the appropriate form that peer effects ought to follow, either spillover or conformist, as evidenced by the literature (Boucher et al., 2022b; Liu et al., 2014). The present context assumes that bullying behavior among adolescents is predominantly influenced by social norms,[30] a finding that has been established in prior studies (e.g., Perkins et al., 2011).

We obtain the following best response function for every student's bullying effort, b_i, in

[29]Note that I model for continuous measures of bullying to ensure the existence and uniqueness of equilibrium for tractability. Empirically, in the estimation, I map the reported bullying frequencies to their latent bullying scores using a Two-parameter Logistic model from Item Response Theory in psychology (Lord, 2012). More details are discussed in Appendix Section E.

[30]In order to align with the social norm narrative, I perform row normalization on the matrix $\mathbf{W_1}$ during the estimation process. Although the notation remains unchanged for brevity, it is important to note that all calculations and analyses are based on the row-normalized version of $\mathbf{W_1}$.

equilibrium:

$$b_i^* = \underbrace{d_i(\mathbf{x})}_{\text{human capital effect}} + \underbrace{\frac{\phi}{1+\phi}[\sum_{j\neq i}^{N} w_{1ij} b_j^* - d_i(\mathbf{x})]}_{\text{social effect}}, \quad \forall i \in N. \tag{7}$$

Equivalently, it can be further expressed in a matrix form as:

$$\mathbf{b}^* = [\mathbf{I} + \lambda(\mathbf{I} - \mathbf{W_1})]^{-1}\mathbf{d}(\mathbf{x}), \tag{8}$$

where $\lambda \equiv \frac{\phi}{(1+\phi)}$ is a scalar that captures the degree of peer effect, $\mathbf{I}$ is the identity matrix with dimension $n \times n$, and $\mathbf{W_1}$ denotes the follow-up friendship network matrix. There exists a unique Nash equilibrium of bullying score as long as regularity conditions hold, i.e., $|\lambda| < 1$ (Ballester et al., 2006; Debreu and Herstein, 1953).

To capture empathy's individual human capital effect on bullying, I formulate the private marginal benefit of engaging in bullying as $d_i(\mathbf{x}) = \sum_{m=1}^{M} \beta_m x_{mi} + \nu_i$,[31] where $\mathbf{x}$ is a M-dimensional vector of students' characteristics, including the follow-up empathy H_{1i}, and ν_i denotes individual i's unobserved heterogeneity. Therefore, the equilibrium effort of individual i, b_i, in (7) is given by

$$b_i = (1-\lambda)\sum_{m=1}^{M} \beta_m x_{mi} + \lambda \sum_{j\neq i}^{N} w_{1ij} b_j + \varepsilon_i, \tag{9}$$

where $\varepsilon_i \equiv (1-\lambda)\nu_i$.

Key Assumptions of the Model Setup The above model exhibits several simplifying assumptions. First, I simplify parents' decision-making of participation. In the current model, I abstract away from factoring expected bullying outcomes and future friendship network structures into parents' decision-making process. This assumption is reasonable for several reasons: (i) it echoes our program design in that the program was advertised as an empathy education program instead of a bullying prevention program, (ii) parents have limited knowledge about their children's friendship networks, as discussed in Section 3.2, and (iii) I empirically test that child bully status and popularity are not correlated with the participation decision, as shown in Table 5. Moreover, it is plausible that parents can only make decisions based on limited information, in line with the bounded rationality literature (Kahneman, 2003).[32]

Second, I also assume that students' decision-making regarding their social connections does not incorporate the consideration of final bullying outcomes b_i. In other words, I assume

[31] This specification is flexible for the additional inclusion of contextual effects. In that case, I can let $d_i(\mathbf{x}) = \sum_{m=1}^{M} \beta_m x_{mi} + \frac{1}{\sum_{j\neq i} w_{ij}} \sum_{m=1}^{M} \sum_{j\neq i}^{n} \gamma_m w_{1ij} x_{mj} + \nu_i$ and the parameter γ_m captures the contextual effects, i.e., the effects of peers' characteristics on one's own behavior.

[32] I will leave modeling the feedback effect of bullying on network formation for future studies when longer-term data are available.

that students make two decisions separately – one on friends (which depends on friends' bullying decisions) and another decision on their own bullying outcomes. This setting is motivated by the finding that adolescents are less sophisticated than adults in decision-making and their goals are more likely to maximize immediate pleasure (Reyna and Farley, 2006). Additionally, this assumption greatly enhances the model's tractability and aligns with the approach employed by scholars such as Griffith (2022b) and Boucher et al. (2023).[33]

Furthermore, the network formation model assumes the absence of utility derived from "indirect" connections. This assumption is shared by researchers such as Graham (2017), Griffith (2022b), König et al. (2019), and Lee et al. (2021). Nevertheless, accounting for externalities or additional utility gains stemming from popularity may unveil a more pronounced social effect of empathy on reducing instances of bullying. I leave this issue for future research.

Lastly, I assume "sequential exclusion" such that the treatment only affects students' network structures by improving students' empathy skills. I justify it empirically and find that conditional on the follow-up empathy and the baseline network links, receiving the treatment and the other follow-up non-cognitive abilities are not significantly correlated with the follow-up friendship links, as shown in Appendix Table I10. This result supports the "sequential exclusion" assumption.

6 Empirical Results

In this section, I first discuss the identification of the model. I then explain the estimation strategy, followed by providing the estimation results.

6.1 Identification Strategy

As illustrated in the previous section, the whole model system is composed of estimating four components: Parents' participation decision (Eq (2)), empathy formation (Eqs (3) and (4)), network formation (Eq (5)), and bullying involvement (Eq (9)). The model contains three endogeneity issues: (i) the endogeneity of the take-up decision P_i in empathy formation, this issue may arise due to the correlated error term structure involving parents' preference shock ϑ_i and empathy formation shock ε_i.[34] Then, (ii) the endogeneity of the friendship network matrix $\mathbf{W_1}$ in the final bullying outcome equation (9), if there exists an unobservable factor that affects both the bullying efforts, d_i and d_j, and the friendship matrix, w_{1ij}, it is possible that the non-random formation of network links is a concern. For example, two students exposed to similar violent content on social media are more likely to be friends, and at the same time, both have a higher private utility of engaging in bullying. Lastly, (iii) the reflection problem, i.e.,

[33] This setup is distinct from the models used by Badev (2021) and Hsieh et al. (2022), where they explicitly model the co-evolution of network links and individual behavior.

[34] Essentially, it presents a selection on unobservables problem.

the simultaneity of my own and my peers' bullying behavior, d_i and d_j, because an individual may both influence her group's behavior and be affected by it at the same time. I discuss how to account for these endogeneity issues accordingly.

To address the endogeneity of the take-up decision in empathy formation, I adopt the classic self-selection framework (Heckman and Sedlacek, 1985; Roy, 1951). I model endogeneity by assuming a bivariate normal distribution of parent preference heterogeneity in participation decisions and the unobserved shock in students' empathy formation, allowing for a correlation between these two shocks. I assume that the same empathy shock is invariant to the take-up status to ensure point identification of parameters. The normality structure leads to the inclusion of inverse Mills ratio terms to correct for the selection bias in the expected follow-up empathy regression conditioning on participation status. The variances of the two shocks are identified from the observed variances of empathy in both the compliance (take-up) and non-compliance (non-take-up) groups. I estimate the model using the treatment group. Thanks to the random assignment of the empathy program and the observation that no participants in the control group participate in the program, I can validate my estimation results and the error term assumption using the control group sample. Further implementation details can be found in Section 6.2 and the Appendix Section F.

To address the endogeneity of the friendship matrix $\mathbf{W_1}$ in the bullying outcome, I use a two-stage IV estimation approach based on the method proposed by König et al. (2019) and Lee et al. (2021). In the first stage, I predict $\mathbf{W_1}$ based on predetermined dyadic characteristics,[35] and in the second stage, I construct IVs using the predicted adjacency matrix $\widehat{\mathbf{W_1}}$.

To address the simultaneity of my own and my peers' bullying behavior, I explore the structure of interactions in a directed friendship network to identify peer effects using an instrumental variable strategy, similar as in König et al. (2019). Given that each individual's outcomes can be influenced by their nominated friends, i.e., those direct connections, then under the presence of intransitivity, the characteristics of an individual's indirect connections can be used as instruments to identify peer effects. Intuitively, the identification condition says that the characteristics of the friends' friends of a student who are not her direct friends could serve as valid instruments for the bullying behavior of her own friends. Therefore, I construct $\left(\widehat{\mathbf{W_1}}^2\mathbf{X}, \widehat{\mathbf{W_1}}^3\mathbf{X}, \ldots\right)$ as excluded instruments for peers' bullying behavior $\mathbf{W_1 b}$, given the existence of intransitivity, that is $\mathbf{I}, \widehat{\mathbf{W_1}}, \widehat{\mathbf{W_1}}^2$, and $\widehat{\mathbf{W_1}}^3$ are linear independent.

[35] The underlying rationale is based on the concept of homophily, which posits that individual-level characteristics only impact the formation of links when they align with similar attributes of another agent at the dyad level. Consequently, the functional form of the outcome equation that models action choices inherently excludes dyad-level variables. This exclusion stems from the disparity between the dimensions of the dyad-level link formation equation (5) and the individual-level outcome equation (9), thus establishing a natural exclusion restriction.

6.2 Empirical Specification and Estimation

Parents' Problem Assuming the error terms ε and ϑ are correlated with covariance $\sigma_{\varepsilon v}$ and follow a normal distribution and with mean zero, variances σ_{ε}^2 and σ_{ϑ}^2, the joint distribution of the two error terms follows $\begin{pmatrix} \varepsilon_i \\ \vartheta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{\varepsilon}^2 & \sigma_{\varepsilon v} \\ \sigma_{\varepsilon v} & \sigma_{\vartheta}^2 \end{pmatrix}\right)$. Equivalently, one can also write $\varepsilon_i = \frac{\sigma_{\varepsilon v}}{\sigma_{\vartheta}^2}\vartheta_i + \eta_i$, where $\vartheta_i \perp \eta_i$.

Parents choose not to participate, i.e., $P_i = 0$, if $U_i < 0$. This gives:

$$\begin{aligned} U_i < 0 &\Longleftrightarrow \gamma_0 + \gamma_1\delta + \gamma_1(\beta_{1,1} - \beta_{1,0})H_{i,0} + \gamma_2\mathbf{X_i} - \gamma_3 C_i + \vartheta_i < 0 \\ \vartheta_i &< \underbrace{\gamma_3 C_i - \gamma_2\mathbf{X_i} - \gamma_1(\beta_{1,1} - \beta_{1,0})H_{i,0} - \gamma_1\delta - \gamma_0}_{:=\bar{h}}. \end{aligned}$$

Likewise, $P_i = 1$ if $U_i \geq 0$, which corresponds to $\vartheta_i \geq \bar{h}$. Given the normality assumption of parents' preference shock ϑ and let $\bar{h}^* \equiv \frac{\bar{h}}{\sigma_{\vartheta}}$, we can calculate the parent i's participation probability, where Φ denotes the standard normal CDF:

$$P(P_i = 1) = \Phi(\bar{h}^*). \tag{10}$$

In order to obtain the parameters governing the empathy formation process, we need to rely on the first-order moments. Given the error term structure, we can derive the conditional expectation of child i's follow-up empathy when not participating as follows:

$$E(H_{i,1}|H_{i,0}, C_i, X_i, P_i = 0) = \beta_0 + \beta_{1,0}H_{i,0} - \sigma_{\varepsilon v}\frac{\phi(\bar{h}^*)}{\Phi(\bar{h}^*)}. \tag{11}$$

The last term on the RHS, referred to as the Inverse Mills Ratio, serves to control for selection. Similarly, the conditional expectation of child i's follow-up empathy when participating is derived as:

$$E(H_{i,1}|H_{i,0}, C_i, X_i, P_i = 1) = \beta_0 + \delta + \beta_{1,1}H_{i,0} + \sigma_{\varepsilon v}\frac{\phi(\bar{h}^*)}{1 - \Phi(\bar{h}^*)}. \tag{12}$$

To obtain estimates of the variances, we also need to utilize the second-order moments. Specifically, for those who choose to participate:

$$Var(H_{i,1}|H_{i,0}, C_i, X_i, P_i = 1) = \frac{\sigma_{\varepsilon v}^2}{\sigma_{\vartheta}^2}\left(1 - \frac{\bar{h}^*\phi(\bar{h}^*)}{\Phi(\bar{h}^*)} - \left(\frac{\phi(\bar{h}^*)}{\Phi(\bar{h}^*)}\right)^2\right) + \sigma_{\varepsilon}^2 - \frac{\sigma_{\varepsilon v}^2}{\sigma_{\vartheta}^2}. \tag{13}$$

Similarly for those who choose not to participate,

$$Var(H_{i,1}|H_{i,0},C_i,X_i,P_i=0)=\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}\left(1+\frac{\bar{h}^*\phi(\bar{h}^*)}{1-\Phi(\bar{h}^*)}-\left(\frac{\phi(\bar{h}^*)}{\Phi(\bar{h}^*)}\right)^2\right)+\sigma_\varepsilon^2-\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}. \tag{14}$$

The estimation of the parameters in the parents' problem involves three steps: The first step requires a Probit regression of the participation dummy P_i over C_i, X_i, and the baseline empathy $H_{i,0}$ on the treated group as specified in Eq (10).[36] When performing estimation, I let C_i depend on the family income level, and $\mathbf{X_i}$ denotes the vector of additional controls for parents' characteristics, such as age and whether the mother responded to the survey. Furthermore, I account for various characteristics of the children, including child bully status, inverse CES-D score, test score, study pressure, popularity, and peers' bullying level.[37] I can then construct the selection correction terms in Eqs (11) and (12). This step provides estimates of $\frac{\gamma_0+\gamma_1\delta}{\sigma_\vartheta}, \frac{\gamma_1(\beta_{1,1}-\beta_{1,0})}{\sigma_\vartheta}, \frac{\gamma_2}{\sigma_\vartheta}, \frac{\gamma_3}{\sigma_\vartheta}$.

The second step involves conducting OLS regressions using Eqs (11) and (12). Here, I impose the coefficient constraint to estimate $\sigma_{\varepsilon v}$. This step provides estimates of $\beta_0, \delta, \beta_{1,0}, \beta_{1,1}, \sigma_{\varepsilon v}$.

In the third step, I first transform the two conditional variance equations (13) and (14) to be conditional solely on the participation status, i.e., $Var(H_{i,1}|P_i=1)$ and $Var(H_{i,1}|P_i=0)$. Using nonlinear least square estimation, I match the observed variances of follow-up empathy with the model-implied variances.[38] This step provides estimates for the variances of both shocks, σ_ε^2 and σ_ϑ^2.

Network Formation In estimating the dyadic network formation described by equation (5), I specify the surplus structure of the network formation equation (5) following Graham (2017). Specifically, I let $\Xi_{ij}=\tau_i+\xi_{ij}$ and $\varepsilon_i=\tau_i+a_i$ as compounded error terms, where $V_{ij}\perp\xi_{ij}$ and $\xi_{ij}\perp a_i$.[39] Note that the incorporation of unobserved heterogeneity of individual i is flexible enough to address the potential endogeneity of the follow-up empathy H_{1i} in network formation. Assuming that ξ follows a logistic distribution and that they are independently and identically distributed across dyads, one can express the likelihood of observing network $\mathbf{W_1}$

[36] I only use data from the treatment group for estimation. This decision is based on our records of program take-up, which indicate that only individuals in the treatment group participate in the program. Consequently, there is no variation in take-up decisions within the control group.

[37] The inclusion of characteristics X_i and C_i can also serve as exclusion restrictions to solve the selection issue. Additionally, testing whether the children's bully status and popularity play a significant role in the take-up estimation equation also helps address concerns regarding the potential imbalance in counterfactual targeting experiments based on students' popularity and bully status.

[38] I lay out the estimation details in Appendix Section F.

[39] The endogeneity of the peer network in the final bullying outcomes arises from the unobserved heterogeneity τ_i, which affects both ε_i and Ξ_{ij}. Specifically, one can derive that $E[(\sum_j w_{1ij})\varepsilon_i]=\sum_j E[(w_{1ij})\varepsilon_i]=\sum_j E[\mathbf{1}\left(w_{0ij}\psi_0+V_{ij}'\psi_v+X_j'\psi_\gamma+\tau_i+\xi_{ij}\right)\varepsilon_i]=\sum_j E[\mathbf{1}\left(w_{0ij}\psi_0+V_{ij}'\psi_v+X_j'\psi_\gamma+\tau_i+\xi_{ij}\right)(\tau_i+a_i)]\neq 0$.

as:

$$\Pr(\mathbf{W_1} = w_{1ij} \mid \mathbf{V}, \tau) = \prod_{i \neq j} \left[\frac{1}{1+\exp\left(w_{0ij}\psi_0 + V'_{ij}\psi_v + X'_j\psi_\gamma + \tau_i\right)} \right]^{1-w_{1ij}} \left[\frac{\exp\left(w_{0ij}\psi_0 + V'_{ij}\psi_v + X'_j\psi_\gamma + \tau_i\right)}{1+\exp\left(w_{0ij}\psi_0 + V'_{ij}\psi_v + X'_j\psi_\gamma + \tau_i\right)} \right]^{w_{1ij}}.$$

Thus, the link formation probability can then be estimated using the following conditional logistic regression function allowing for degree heterogeneity:

$$\Pr\left(w_{1ij} = 1 \mid \mathbf{V}, \tau\right) = \frac{\exp\left(w_{0ij}\psi_0 + V'_{ij}\psi_v + X'_j\psi_\gamma + \tau_i\right)}{1 + \exp\left(w_{0ij}\psi_0 + V'_{ij}\psi_v + X'_j\psi_\gamma + \tau_i\right)}. \tag{15}$$

I incorporate controls for the previous link connection w_{0ij} at the baseline. This inclusion allows for a more accurate capture of network changes following the intervention. In addition, the model also controls for dyadic covariate V_{ij} including differences in ages, height, time spent on studying and on leisure, pocket money (to capture differences in socio-economic backgrounds), and test scores, all measured at the baseline. I also include dummies for whether students were both bullies or victims at the baseline to capture potential taste changes in making friends with bullies. While additional determinants could have been included based on theoretical grounds, the chosen covariates are predetermined from the perspective of the students, making them valid instruments for link formation.

Following the approach of Santavirta and Sarzosa (2019), I use a fixed effect Logit model to estimate Eq (15). The model incorporates individual i's fixed effects to account for sender i's average tendency to nominate others, while controlling for observed variations in the characteristics of the receiver of the friendship nomination, denoted as friend j. Specifically, the vector of j's characteristics X_j in the model includes gender, an indicator for being an only child, follow-up empathy, and baseline standardized test scores as covariates. In contrast to the estimation method proposed by Graham (2017) that requires undirected links, this estimation approach considers friendships to be asymmetric due to the nature of how students nominated their friends within classrooms in the survey. This leads to estimating the model under the setting of directed links, so that it can account for the information on who nominates whom and reflects the asymmetry inherent in the formation of friendships.

Bullying Involvement To estimate the peer effects equation of bullying outcomes (9), I first construct the predicted adjacency matrix $\widehat{\mathbf{W}_1}$. I apply the same structure as Eq (15). However, one needs to remove τ_i as these terms are correlated with the error term in (9). Therefore, based on (15), I define

$$\hat{w}_{1ij} = \frac{\exp\left(w_{0ij}\widehat{\psi_0} + V'_{ij}\widehat{\psi_v} + X'_j\widehat{\psi_\gamma}\right)}{1 + \exp\left(w_{0ij}\widehat{\psi_0} + V'_{ij}\widehat{\psi_v} + X'_j\widehat{\psi_\gamma}\right)},$$

where $\widehat{\psi_0}$, $\widehat{\psi_v}$, and $\widehat{\psi_\gamma}$ are obtained from the logistic regression of Eq (15).[40] Therefore, the predicted adjacency matrix $\widehat{\mathbf{W_1}}$ is composed of $\hat{w}_{1ij}$ for $i \neq j$ and 0 for all other entries.

Once we obtain the predicted adjacency matrix, let $\mathbf{Q} = [\mathbf{X}, \widehat{\mathbf{W_1}}\mathbf{X}, \widehat{\mathbf{W_1}}^2\mathbf{X}, \widehat{\mathbf{W_1}}^3\mathbf{X}]$ denote the IV matrix. The final estimator of the peer effects equation becomes $\hat{\theta}^{2SLS} = \left(\widetilde{\mathbf{X}}'\mathbf{M}\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}'\mathbf{M}\mathbf{b}$, where $\widetilde{\mathbf{X}}$ denotes the matrix of regressors $[\mathbf{X}, \mathbf{W_1}\mathbf{X}, \mathbf{W_1}\mathbf{b}]$ and $\mathbf{M}$ is the projection matrix $\mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}$. In the estimation, the vector $\mathbf{X}$ includes the follow-up empathy, allowing for the examination of its impact on bullying, as well as gender, an indicator for being an only child, *hukou* status, baseline bully status, and baseline victim status. I also control for strata fixed effects in estimating the peer effects of bullying due to the randomization nature. The strata fixed effects can help alleviate the concern of correlated effects, as discussed by Manski (1993).[41]

To account for the multiple steps involved in the estimation procedure for the final bullying outcome equation (9), I use bootstrap methods with 1,000 repetitions, redrawing a sample of clusters each time, to obtain standard errors. This adjustment helps address the issue that the second-stage errors are incorrect as they inherit noise from the first stage.

6.3 Estimation Results

6.3.1 Parents' Decisions

Table 5 shows the results of parents' decisions including the participation decision (Panel A) and the child empathy formation (Panel B). From Panel A, the gain in empathy due to participation has a significant effect on parents' participation decisions. I also find that the participation decision is highly correlated with children's academic outcomes. Parents of children with better academic outcomes are more likely to participate, while those who skip exams and have worse mental health outcomes are less likely to participate. Older parents are more likely to participate than younger parents. Neither children's baseline bully status nor network-related measures have any significant effects on parents' participation decisions. This finding alleviates the self-selection bias issue in subsequent counterfactual experiments targeting bullies and students based on popularity, as discussed in Section 7.

As shown in Panel B, participation in the empathy program increases both the average level of follow-up empathy and the marginal return of baseline empathy in empathy development. Despite the small statistical significance, the magnitude of the mean-level empathy improvement due to participation amounts to about 0.25 SD of the empathy score, with an additional

[40] As noted by Lee et al. (2021), the exclusion of the individual fixed effect τ_i from the predicted link probability $\hat{w}_{1ij}$ does not undermine the validity of my estimation strategy. The consistency of parameter estimates in Eq (6) is ensured as long as the estimated vector $\widehat{\Psi} = (\widehat{\Psi_0}, \widehat{\Psi_v})$ converges. Therefore, I use the estimates obtained from the link formation equation (15) in my subsequent analysis.

[41] According to Manski (1993), the correlated effect is defined as the issue that people in the same reference group often react similarly, but not because they are influenced by one another but rather because they have similar unobserved characteristics in common, such as institutional contexts or common shocks.

0.05 SD increase in marginal return. The covariance between parents' preference heterogeneity and empathy shocks turns out to be negative, though not significant. One possible explanation is that I observe parents of children with high test scores and high baseline empathy are more likely to participate. However, those participating children are also more likely to experience diminishing returns compared to those with low baseline empathy levels.

6.3.2 Network Formation

Panel A of Table 6 shows the results of link formation prediction. To interpret these results, I compare the set of results with the one obtained from not conditioning on the baseline network (as shown in Appendix Table I11). When not controlling for the baseline links, there is strong evidence of homophily, especially in gender, *hukou* status, and age. Those with the same gender and the same *hukou* are more likely to be friends. I also find that those who were victims at the baseline are more likely to be friends after the intervention. This finding adds to the homophily feature of bullying perpetration by showing that the homophily pattern is also non-negligible among victims. Additionally, students with a significant gap in test scores are less likely to form friendships. This phenomenon may be attributed to China's exam-oriented educational evaluation system, which can influence how children choose their friends.

After adjusting for the previous link structure, the overall homophily pattern becomes less pronounced. Moreover, the homophily associated with being a bully or a victim at the baseline loses statistical significance, suggesting that the intervention may have disrupted the bully-bully network. Notably, the difference between my own empathy and my peer's empathy does not appear to influence link formation at the follow-up. However, the impact of the receiver's empathy on forming friendship links remains highly significant, even when controlling for baseline links. In terms of effect size, the estimated coefficient suggests that a 1 SD increase in empathy would bring 0.14 more friend nomination links per student. Considering the average class size of 47 in the study sample, the effect is equivalent to 6.4 more links compared to the baseline within an average-sized class. This finding corroborates the work of Kardos et al. (2017), who demonstrate that empathic abilities can predict network sizes, indicating that being more empathetic tends to attract more friends.

6.3.3 Bullying Involvement

Panel B of Table 6 presents the results for estimating bullying involvement. I observe that empathy directly reduces bullying through its human capital effect. In terms of magnitude, a one SD increase in own empathy skill corresponds to a 0.14 SD decrease in bullying score. When examining the peer effects estimate, the positive sign indicates a preference for conformity among students. This implies a positive "spillover" of peer bullying behavior, and the effect is statistically significant at the 1% level. According to the results, a one SD increase in the

average peer bullying score would result in a 0.48 SD increase in my own bullying score. These figures translate to an 18% higher likelihood of being a bully at follow-up. Additionally, male students exhibit a higher likelihood of engaging in bullying, consistent with previous research findings (e.g., Lavy and Schlosser, 2011). Furthermore, those with an urban *hukou* are also more inclined to bully others.

Remarkably, bullying behavior displays a degree of persistence, with individuals who report ever bullying others at the baseline being significantly more likely to continue bullying at the follow-up. This persistent effect is sizable, approximately 0.56 SD, which corresponds to a 21% higher likelihood of being a bully. Meanwhile, results suggest there is also a "revenge" effect from those victims at the baseline, as being a victim during this period is linked to a 0.16 SD increase in bullying scores at the follow-up. This result suggests the intricate dynamics at play between bullies and victims.

Censored Network Concern I use the Add Health survey methodology to collect friendship network data. As highlighted in Griffith (2022a), there is a concern that our data may suffer from incompleteness or censorship due to the five-name cap. To address this concern, I fully leveraged our data structure by recording names in order of closeness. Consequently, I conducted robustness checks by deliberately truncating the list of friends at specific rank-order positions for individuals who reported five names. Appendix Table I12 presents the simulation results. In line with the findings in Griffith (2022a), where partial networks were allowed but endogenous networks were not considered, the estimate of peer effect coefficient in the current context also gradually attenuates to zero as I truncate more important friends. This pattern suggests that a censored network with an endogenous network could also lead to an underestimation of the social effect, although formal proof will be deferred to future studies.[42] With a complete network, we would expect the estimated social effect of empathy to be more substantial. Furthermore, the estimates capturing empathy's direct effect on bullying and its impact on network formation remain relatively stable.

Comparison with Alternative Approaches For comparative purposes, I present the results from an OLS estimation, the 2SLS estimation proposed by Bramoullé et al. (2009), and the 2SLS method used in this study in Appendix Table I13. The OLS estimation from Column (1) is smaller than the estimations accounting for potential endogeneity, which may be due to measurement error or omitted variable bias. Comparing the 2SLS estimation in Column (2), which does not consider an endogenous network, with the method utilized in the main estimation in Column (3), the human capital effect of empathy diminishes while the peer effects estimate increases.[43] This finding provides further evidence that empathy impacts bullying not

[42] As far as I know, the formal solution to address both issues—endogenous network formation and censored network—is still under investigation.

[43] This finding aligns with Qu and Lee (2015), who also observe that the commonly used estimates under

only directly, but also indirectly. Ignoring empathy or altruistic preferences in social networks can lead to bias in peer effect estimation.

6.3.4 Model Validation

Within-Sample Fit I first present the within-sample fit in Appendix Table I14. The model predicts the overall mean of the participation rate and empathy fairly well. It slightly overestimates the mean of in-degree but underestimates the mean of the change in bully scores. When examining the overall distribution of bully scores, I find that the model exhibits a good fit. In Appendix Figure 3, I draw the density plots for observed and predicted bullying scores in Panel A. I also compare the observed and predicted proportions of bullies and non-bullies in the study sample, as shown in Panel B.[44] The model fits the distribution of the bully scores well, and it also precisely predicts the proportion of bullies and non-bullies given the optimal cutoff.

Out-of-Sample Validation The random assignment, coupled with the observation that no individuals from the control group ever participated in the program, allows me to test the validity of the error term structure assumption of children's empathy formation using the control group. From the lower segment of Panel B of Table 5, I present the results of estimating the empathy production function (3) using the observations solely from the control group. I also test the equality of the key parameters. The large p-values indicate that we cannot reject the null hypothesis stating that the key parameters estimated for the non-compliers (i.e., those who did not participate) in the treatment group and the ones in the control group are equal. This finding suggests that the framework is capable of explaining data patterns related to parents' participation decisions and children's empathy formation. It also alleviates the concern of potential spillover of empathy from compliers to non-compliers within treatment classes.

I also conduct a validation exercise similar to that of Todd and Wolpin (2006), where I compare the impacts predicted by the model to the experimental impacts. The model performs less satisfactorily for in-degree, as shown in Appendix Figure I4, possibly due to the abstraction of strategic interactions among students in the network formation model. However, the model performs quite well in estimating the treatment effects on empathy and bullying.

6.3.5 The Decomposition Exercise

One of the advantages of the structural model is that it can quantify the exact contribution of empathy effect on bullying. As discussed in the previous section, improving my own empathy

exogenous weight matrix suffer from a downward bias when the true weight matrix is endogenous.

[44] To transform the continuous bully scores into dummy variables, I assume there is a cutoff point for the bullying score. An individual is predicted as a "bully" if his/her score is above the cutoff, and a "non-bully" if the score is below. I use the `cutpointr` package in `R` to find the optimal cutoff that maximizes the sum of sensitivity (true positive rate) and specificity (true negative rate).

can reduce my own bullying because of my own empathetic concerns and perspective-taking (*individual human capital effect*). Furthermore, improving my own empathy can also lead to changes in the social networks and the associated peer spillovers of bullying (*social effect*).

Based on the equilibrium bullying behavior derived from the model (Eq (8)), one can derive the change in the bullying score as:[45]

$$\Delta \mathbf{b} = \mathbf{b}^*_{\text{post}} - \mathbf{b}^*_{\text{pre}} = [\mathbf{I} + \phi(\mathbf{I} - \mathbf{W}_{\text{post}})]^{-1}(\beta \Delta H_i + \phi(\mathbf{W}_{\text{post}} - \mathbf{W}_{\text{pre}})\mathbf{b}^*_{\text{pre}}). \tag{16}$$

Therefore, the *individual human capital effect* of empathy change on bullying score is obtained by disabling the peer effects parameter ϕ in the model (i.e., setting $\phi = 0$) and simulating the new bullying outcomes in the sample. The *social effect* of empathy change on bullying score—due to the reshaped network structures and the associated peer spillovers of bullying—can then be calculated by subtracting the total bullying score change from the simulated bullying changes by setting $\phi = 0$. Figure 2 illustrates the decomposition exercise results. The social effect of empathy on bullying accounts for 32% of the total change in the bullying score.

Alternatively, one can fully exploit the structure of Eq (16). In Appendix Figure I5, I show additional results when (i) only shutting down empathy's human capital effect, i.e., setting $\phi = \hat{\phi}$ while $\beta = 0$, and (ii) assuming the network links remain the same across the two periods, i.e., $\mathbf{W}_{\text{post}} = \mathbf{W}_{\text{pre}}$. As demonstrated in the figure, the alternative approach of calculating the *social effect* by setting $\phi = \hat{\phi}$ and $\beta = 0$ yields a similar result regarding the relative importance of empathy's human capital and social effects compared to setting $\phi = 0$. I find that when assuming the links remain the same at both baseline and follow-up, but allowing for friends' types to change, such as from bullies to non-bullies, it results in the new simulated change in bullying score accounting for around 70% of the total change. This exercise underscores the advantage of combining treatment effect analysis with the decomposition framework, providing a precise breakdown of each channel for a better understanding of empathy's effect on the network structure.

7 Targeting Experiments With Network Information

Now that we have an understanding of empathy's human capital and social effects, the next question is whether we can achieve greater reductions in bullying through alternative targeting experiments using social network information, compared to random assignments.

First, I investigate the effects of targeting experiments based on students' popularity, motivated by the "social referents" theory in the social psychology literature (Paluck and Shepherd, 2012; Paluck et al., 2016). This theory suggests that the behavior of highly connected individ-

[45] I show the detailed derivation steps in Appendix Section G.

uals in a group can shape perceived collective norms and, consequently, mitigate undesirable behaviors. I then compare the effectiveness of this approach with nondiscriminatory allocations, i.e., random assignment.[46]

Figure 4 illustrates the findings, wherein the effectiveness is quantified by the percentage-point decrease in the number of bullies. The results in Figure 4 demonstrate that targeting based on popularity consistently leads to greater reductions in bullies compared to the random assignment. At its maximum, popularity-based targeting leads to a further 1.5 percentage-point decrease in bullying. Given that the proportion of bullies in the study sample is around 20%, this translates to a 7.5% further reduction in bullies. The disparity in effectiveness between the two sets of experiments follows a non-linear pattern: it initially increases, peaks when treating approximately half of the sample, and then gradually diminishes, eventually converging as all units are treated. This shrinkage in the effectiveness gap could be attributed to the congestion effect in social networks, arising from the overlap of students' networks when a larger number of individuals are involved.

Second, I apply the proposed framework to simulate the program's impact on reducing bullying for each of the following scenarios: (i) individuals selected based on specific characteristics (e.g., being a bully), and (ii) individuals selected based on both characteristics and their network ties (e.g., bullies' best friends). The former scenario aligns with the Zero Tolerance policies that are relatively prevalent in American schools (Borgwald and Theixos, 2013). Relatedly, Şahin (2012) evaluates an empathy program exclusively targeting bullies in Turkish primary schools. As for the latter scenario, interventions relying on peer support have been applied in various contexts. Examples include efforts to enhance women's reproductive agency in India (Anukriti et al., 2023), and interventions employing group therapies and spousal-level counseling to decrease alcohol abuse in western Kenya (Murphy, 2023). As these simulations involve different numbers of treated units and the effects on bullying are highly non-linear due to the incorporated social effect, I also compare these targeting experiments against corresponding random assignments with an equivalent number of treated units. The comparison ensures a rigorous assessment of the relative effectiveness across the different targeting experiments.

Figure 5 shows the results. I find that targeting bullies, whether they are popular or non-popular,[47] consistently leads to fewer reductions in bullying in comparison to random assign-

[46]To implement the popularity-based experiment, I rank all students based on their in-degree measure. Subsequently, I select a specific number of students from this ranked sample to compose the treated group. I repeat this process nine times to create a smooth line for extrapolation. For each nondiscriminatory experiment, I randomly assign a certain number of treated units from the original study sample (which is not popularity-ranked) 100 times and simulate the predicted changes in empathy, network structures, and, finally, bullying scores. I then average the results over the 100 simulations and report the mean of the changes in bullying scores. This entire process was repeated nine times to capture the pattern of the simulated bully outcomes.

[47]I classify popular and non-popular bullies based on a comparison of each student's in-degree measure with the median in-degree of the 456 bully sample. A student with an in-degree higher than the median is labeled as a popular bully, and vice versa.

ments involving the same number of treated units, as depicted in Panel A. In contrast, targeting bullies' social circles, in particular their best friends, is more effective than random assignments, though the difference is not statistically significant, probably due to the small number of treated units, as shown in Panel B. Targeting the top 10% of most popular students,[48] is also more effective than random assignment. This targeting approach results in a similar number of students being assigned to the treatment group ($N = 480$) as targeting bullies ($N = 456$). However, the former yields nearly twice the reduction in bullying as the latter. This discrepancy underscores the limited effectiveness of directly targeting bullies.

This finding aligns with the results obtained from our decomposition exercise. Applying the decomposition framework, I observe that when targeting bullies' best friends ($N = 394$) or the most popular students ($N = 480$), the social effect of improving empathy, driven by the induced changes in network structures and associated peer effects, outweighs the individual human capital effect, as illustrated in Appendix Figure I6. In contrast, targeting bullies themselves ($N = 456$) generates a smaller social effect than the human capital effect. The greater magnitude of the social effect when targeting bullies' best friends can be attributed to the fact that among the bullies' best friends, only about 23% of them are bullies. Additionally, when targeting bullies' best friends, the effect inevitably gets transmitted to bullies. However, in other targeting experiments, only a proportion of the effect is transmitted to bullies, as not all of their friends, or even best friends, are bullies.

To further illustrate this point, I compare the distribution of in-degrees for bullies and bullies' best friends, as displayed in Appendix Figure I7. Here, I observe that the distribution for bullies' best friends is more right-skewed. Consequently, one should anticipate a smaller human capital effect and a larger social effect when targeting bullies' best friends. Similarly, as indicated in the same figure, the higher social effect when targeting the most popular students can also be attributed to the fact that these students tend to have a higher level of popularity within the treated population.

These findings so far suggest that for socially undesirable behaviors such as bullying, targeting perpetrators' social circles may perform better than targeting other groups, and sometimes even more effective than directly targeting perpetrators. This is important as much effort tends to intervene on bullies or victims directly (e.g., Dake et al., 2003), while social network literature usually identifies "key players" using various centrality measures or sometimes simply based on popularity (Ballester et al., 2006; Zenou, 2016). Policymakers or practitioners can weigh the trade-off between the gained efficiency from popularity-targeting and the additional costs associated with pre-collecting social network data. If possible, the whole sample treatment is preferable as it always yields the highest effect.[49]

[48] To select the top 10% of most popular students, I rank the study sample by in-degree and choose the first 10 students with the highest in-degree measures in each class.

[49] However, in the presence of constraints, we should prioritize targeting students with higher in-degree, lower

8 Conclusion

This study sheds light on the social side of cultivating adolescents' non-cognitive skills, particularly empathy, in relation to bullying. I collect unique data from conducting a parental involvement program focused on youth empathy development in two middle schools in China. The innovation of this study lies in its ability to connect individual program impacts with the broader social environment. Leveraging detailed friendship network data, I empirically test homophily in adolescents' bullying behavior. The analysis of treatment effects reveals that enhancing social-emotional skills can help mitigate peer status differences within the classroom. In particular, improving adolescents' empathy can help lower bullies' social status by diminishing their popularity, devaluing the gains they derive from bullying, and distancing them from the newly established social norm that opposes bullying.

Furthermore, the richness of the data also allows the flexibility to investigate and quantify how empathy reduces bullying by indirectly shaping peer relationships. By constructing and estimating a unified model with empathy and network formation, and accounting for peer spillovers in bullying, I find that empathy's social effect constitutes 32% of its total effect on bullying reduction. This proportion is equivalent to about half of its human capital effect. Policy counterfactuals suggest that utilizing the friendship network structure can result in greater reductions in bullying compared to random assignments. In this setting, bullies' best friends can be considered potential "key players." On the other hand, targeting bullies is among the least effective. Researchers can weigh the trade-off between the gained efficiency from popularity-targeting and the additional costs associated with pre-collecting social network data. Depending on specific contexts, the traditional and easily implementable approach of random assignment may still hold promise.

Several limitations of the current study are worth mentioning. First, the network data only allows the students to report up to five best friends. If resources are sufficient, future studies should try to collect the complete friendship network as much as possible, therefore conveying a complete picture of the social effects of empathy. With complete network information, one would detect an even larger social effect of empathy on bullying reduction. Second, the current model assumes parents' bounded rationality mainly due to a lack of corresponding measures from the parents. Future studies could relax this assumption by collecting more data, for example, parents' beliefs on their children's bullying behavior or friends' circle.

Despite these limitations, the findings of this study underscore the interplay between social-emotional skills and the social environment, as well as the value of one's social capital in reducing aggressive behavior, such as bullying. This calls for the incorporation of the social environment when analyzing adolescent development in future research. While the current

empathy, and lower test scores first. See detailed discussions in Appendix Section H.

analysis specifically focuses on how empathy reduces bullying, the proposed framework holds the potential to inform prevention strategies addressing a wider range of youth violence and socially undesirable behaviors.

Moreover, understanding the mechanisms of a field experiment using a structural model contributes to the ongoing debate regarding the external validity of RCTs. It serves as an initial step in discussing external validity (Findley et al., 2021; List, 2020). This study delves into the additional mechanisms behind the efficacy of an empathy-based intervention in reducing aggressive behavior such as bullying. Specifically, I examine how the intervention impacts the social environment and associated peer effects, thus enhancing our understanding of intervention effectiveness presented in the previous analyses. By considering social effects along the causal pathway from intervention implementation to the observation of reduced bullying, this study establishes a broader connection between empathy and the reduction of bullying. Equipped with these comprehensive mechanisms, one can employ mechanism mapping to assess the external validity of the experiment across diverse settings.

Table 1: Balance Test

	Mean		Difference	SE
	Control	Treat	C-T	
Panel A. Individual characteristics				
𝟙(male)	0.526	0.532	-0.006	(0.015)
age	14.028	14.012	0.016	(0.138)
𝟙(urban hukou)	0.476	0.447	0.029	(0.033)
𝟙(only child)	0.295	0.299	-0.005	(0.020)
height in cm	161.785	161.936	-0.151	(0.716)
weight in half kilo	101.599	100.270	1.329	(1.400)
empathy score	47.659	48.504	-0.846	(0.614)
empathy index	-0.000	0.083	-0.083	(0.059)
bully score	0.002	-0.010	0.012	(0.036)
𝟙(bully)(at least once)	0.380	0.381	-0.002	(0.024)
𝟙(bully)(more than once)	0.214	0.197	0.017	(0.019)
𝟙(victim)(at least once)	0.704	0.710	-0.005	(0.023)
𝟙(victim)(more than once)	0.574	0.574	-0.000	(0.024)
positive personality index	0.000	0.030	-0.030	(0.040)
stress index	-0.000	0.027	-0.027	(0.042)
time with parents	10.680	10.711	-0.031	(0.760)
reported number of friends	3.968	4.025	-0.057	(0.128)
𝟙(in a small group)	0.619	0.635	-0.017	(0.026)
class size	46.856	46.648	0.207	(1.118)
time spent on studying (per day)	5.904	6.086	-0.182	(0.229)
time spent on playing (per day)	4.876	4.859	0.017	(0.276)
Panel B. Network statistics				
in-degree	3.054	3.158	-0.104	(0.154)
eigenvector centrality	0.213	0.215	-0.002	(0.017)
# reciprocal links	1.340	1.369	-0.029	(0.167)
# unilateral links	1.713	1.788	-0.075	(0.098)
𝟙(isolation)	0.100	0.101	-0.002	(0.013)
N	1,025	1,206		
P-Value (Joint Significance)			0.403	

Note. This table shows the results of the summary statistics and balance test for individual characteristics (Panel A) and network statistics (Panel B). *In-degree*: Represents the number of nominations a student received. *Eigenvector centrality*: Measures the connectedness of a student to other popular students in the network. *Reciprocal link*: Occurs when both students nominate each other as friends. *Unilateral link*: Occurs when one student nominates the other as a friend, but the other doesn't reciprocate. *Isolation*: A dummy variable taking the value one when the student receives no nominations. More details about the network statistics can be found in Appendix Section A.1. Columns (1) and (2) report the mean of each variable for the control and treatment groups, respectively. Column (3) is the difference between the first two columns. Column (4) reports the standard errors for item-wise regressions using the variables labeled in the first column as the dependent variables and the treatment indicator as the only independent variable. Classroom-level clustered standard errors are presented in parentheses.

Table 2: Relationship between Own Perpetration and Peers' Bullying Perpetration

	(1) Being a bully	(2) Being a bully	(3) Being a bully	(4) Being a bully
# of bully friends	0.041***	0.041***	0.039***	0.040***
	(0.013)	(0.013)	(0.013)	(0.013)
# of friends	-0.014**	-0.010	-0.009	-0.008
	(0.007)	(0.007)	(0.007)	(0.010)
Baseline empathy index		-0.084***	-0.056***	-0.056***
		(0.014)	(0.017)	(0.017)
Baseline positive personality			-0.057***	-0.057***
			(0.013)	(0.013)
# of victim friends				-0.002
				(0.011)
Strata FE	Y	Y	Y	Y
Demographics	Y	Y	Y	Y
R^2	0.021	0.037	0.047	0.047
N	2,231	2,231	2,231	2,231

Note. This table shows the correlation estimates between the likelihood of an individual being a bully and her friends' bully status using only the baseline data. Details about the variable construction methods can be found in Appendix Section A.2. In all regressions, I control for demographic indicator variables, including male, urban *hukou*, and only-child status. For each regression, I also control for strata fixed effects. Standard errors are clustered at the classroom level and presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table 3: Program Effects on Network Statistics

	(1) Control	(2) Intention to treat	(3) WCB p-value	(4) N
Panel A. Individual level				
In-degree	3.030	0.195*	0.075	2,231
	(2.137)	(0.101)		
Eigenvector centrality	0.233	-0.020	0.269	2,231
	(0.279)	(0.017)		
Reciprocal link	1.467	0.051	0.608	2,231
	(1.401)	(0.102)		
Unilateral link	1.563	0.159*	0.049	2,231
	(1.396)	(0.080)		
Panel B. Classroom level				
Coleman homophily index				
High empathy	0.043	0.005	0.888	48
	(0.145)	(0.036)		
Low empathy	-0.096	-0.007	0.915	48
	(0.175)	(0.049)		
# of isolated students	5.140	-0.976*	0.075	48
	(3.686)	(0.559)		

Note. This table shows the intention-to-treat (ITT) estimates on various network measures. The individual-level measures consist of in-degree, eigenvector centrality, reciprocal links, and unilateral links; the classroom-level measures consist of the homophily index and the number of isolated students. *In-degree*: Represents the number of nominations a student received. *Eigenvector centrality*: Measures the connectedness of a student to other popular students in the network. *Reciprocal link*: Occurs when both students nominate each other as friends. *Unilateral link*: Occurs when one student nominates the other as a friend, but the other doesn't reciprocate. *Coleman homophily index*: Measures the degree of homophily along a particular characteristic within a network. *Isolation*: A dummy variable taking the value one when the student receives no nominations. More details about the network statistics can be found in Appendix Section A.1. Column (1) reports the means and the standard deviations for the corresponding outcome variables for those in control classes. Column (2) reports the ITT estimates and standard errors, while Columns (3) and (4) report the Cameron et al. (2008) wild cluster bootstrapped (wcb) p-values after 9,999 resampling and the number of valid observations for each analysis. Standard errors are clustered at the classroom level and presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table 4: Who Makes Friends with Who? A Subgroup Analysis

	(1) Control	(2) Intention to treat	(3) WCB p-value	(4) N
Panel A. Bully (baseline)				
# peers are bullies	0.447	0.004	0.966	456
	(0.717)	(0.084)		
# peers are non-bullies	2.174	0.185	0.160	456
	(1.647)	(0.131)		
Panel B. Non-bully (baseline)				
# peers are bullies	0.459	-0.063	0.404	1,775
	(0.746)	(0.078)		
# peers are non-bullies	2.566	0.220*	0.072	1,775
	(1.641)	(0.112)		
Panel C. Homophily				
Bully homophily	-0.166	-0.055	0.601	48
	(0.392)	(0.101)		
Non-bully homophily	0.105	0.012	0.698	48
	(0.172)	(0.031)		

Note. This table shows the intention-to-treat (ITT) estimates for the subgroup analyses of the program's impact on network structure. I define bullies after accounting for the repetition of each of the five events. I report the ITT estimates on whether friends at the follow-up are bullies and non-bullies for baseline bullies (Panel A) and non-bullies (Panel B) separately. At the classroom level, I report the ITT estimates on the bully and non-bully homophily index (Panel C). Column (1) reports the means and the standard deviations for the corresponding outcome variables for those in control classes. Column (2) reports the ITT estimates and standard errors, while Columns (3) and (4) report the Cameron et al. (2008) wild cluster bootstrapped (wcb) p-values after 9,999 resampling and the number of valid observations for each analysis. Standard errors are clustered at the classroom level and presented in parentheses (* $p<0.10$, ** $p<0.05$, *** $p<0.01$). Another set of results with an alternative definition of being a bully (at least once) is reported in Appendix Table I5.

Table 5: Parents' Participation and Child Empathy Production Functions Parameter Estimates

Parameter	Label	(1) Value	(2) Std. Err
Panel A. Parents' participation decision			
γ_0	intercept	-0.343***	(0.036)
γ_1	parental preference on the gain in empathy	0.820***	(0.278)
X variable			
γ_{21}	child bully	0.009	(0.033)
γ_{22}	child popularity	-0.004	(0.011)
γ_{23}	child friend bully proportion	-0.077	(0.059)
γ_{24}	child inverse CES-D score	-0.006**	(0.002)
γ_{25}	standardized test score	0.037***	(0.012)
γ_{26}	child test score missing indicator	-0.125**	(0.063)
γ_{27}	child study pressure	-0.017**	(0.008)
γ_{28}	parent age	0.003**	(0.001)
γ_{29}	mother	-0.015	(0.031)
C variable			
γ_3	income	-0.002	(0.012)
Panel B. Child empathy formation			
Using treatment group (T=1)			
β_0^1	average empathy change at level	-0.083	(0.159)
δ	average empathy change at level when participating	0.246	(0.210)
$\beta_{1,0}^1$	marginal return of baseline empathy skill when not participating	0.235***	(0.037)
$\beta_{1,1}^1$	marginal return of baseline empathy skill when participating	0.287***	(0.057)
$\sigma_{\varepsilon v}$	covariance of empathy shock ε and parents' utility shock ϑ	-0.113	(0.140)
σ_ϑ^2	variance of parents' preference shock ϑ	0.061	(0.069)
σ_ε^2	variance of empathy shock ε	1.060	(0.077)
Using control group (T=0)			
β_0^0	average empathy change at level	-0.066	(0.039)
$\beta_{1,0}^0$	marginal return of baseline empathy skill when not participating	0.239***	(0.035)
Validation: Tests of equality of coefficients			
$\beta_0^1 = \beta_0^0$		p = 0.916	
$\beta_{1,0}^1 = \beta_{1,0}^0$		p = 0.926	
Joint tests		p = 0.989	

Note. This table shows the estimation results of parents' decisions including the participation decision (Panel A) and the technology of empathy formation functions (Panel B). I construct the empathy skill index for both the baseline and the follow-up using the Bartlett factor score. In Panel A, the estimation is performed using a probit regression model, where the dependent variable is the predicted probability of participation (1) versus non-participation (0). All variables are measured at the baseline. In Panel B, I show the estimation results using the treatment group and the validation using the control group. Bootstrapped standard errors are presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

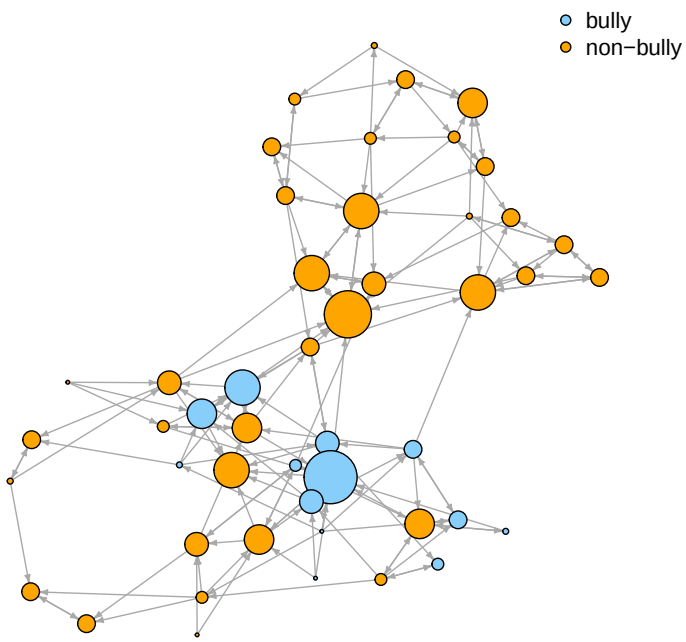
Table 6: Child Network Formation and Bullying Involvement Functions Parameter Estimates

Parameter	Label	(1) Value	(2) Std. Err
Panel A. Network formation			
ψ_0	friends at baseline	0.463***	(0.016)
$\mathbb{1}\left[x_i = x_j\right]$			
ψ_{v1}	bully at baseline	0.002	(0.001)
ψ_{v2}	victim at baseline	0.003	(0.002)
$\left\|x_i - x_j\right\|$			
ψ_{v3}	age	-0.007**	(0.003)
ψ_{v4}	empathy index	-0.001	(0.001)
ψ_{v5}	height	-0.002***	(0.002)
ψ_{v6}	time spent on studying	-0.001	(0.001)
ψ_{v7}	time spent on leisure	-0.002*	(0.001)
ψ_{v8}	pocket money	-0.002	(0.001)
ψ_{v9}	test score	-0.001	(0.002)
x_j			
$\psi_{\gamma 1}$	male	0.002	(0.002)
$\psi_{\gamma 2}$	only child	-0.001	(0.002)
$\psi_{\gamma 3}$	empathy index	0.003***	(0.001)
$\psi_{\gamma 4}$	test score rank	-0.001	(0.001)
Panel B. Bullying involvement			
λ	peer bullying score	0.484***	(0.129)
β_1	empathy index	-0.139***	(0.024)
β_2	male	0.161***	(0.040)
β_3	urban hukou	0.082**	(0.038)
β_4	bully at baseline	0.562***	(0.074)
β_5	victim at baseline	0.162***	(0.042)

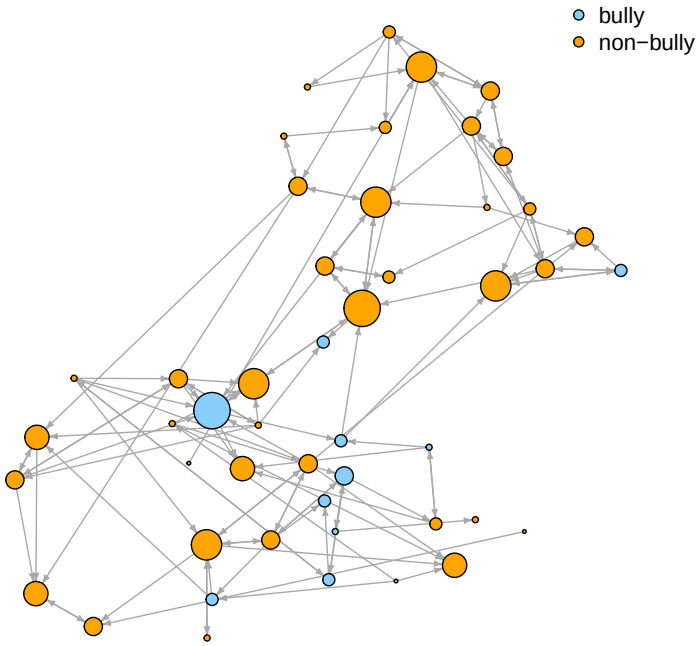
Note. This table shows the estimation results of the network formation model (Eq (15)) in Panel A and the bullying involvement equation (Eq (9)) in Panel B. For network formation, I model the directed links. The dependent variable w_{1ij} takes value 1 if i nominates j (and zero otherwise, even in the event that j nominates i, i.e., $w_{1ji} = 1$). I estimate fixed-effect logit regressions. I include agent i (i.e., sender) fixed effects. The number of observations (i.e., potential links) in all the regressions is 102,054, which stems from a sample of 2,231 unique observations. Standard errors are two-way clustered by nominating and nominated students and presented in parentheses. For bullying involvement, the dependent variable is the standardized bullying score at the follow-up constructed by the IRT model (See details in Appendix Section E). It is worth noting that only the empathy index is obtained from the follow-up measure, while all other variables are measured at the baseline. I use the second- and third-order predicted friends' characteristics as the instrument. The Cragg-Donald F-stat testing validity of the instrument is 20.534. For each regression, I also control for strata fixed effects. Bootstrapped standard errors are presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Figure 1: Visual Illustration of Classroom Network Structure

Panel A. Baseline

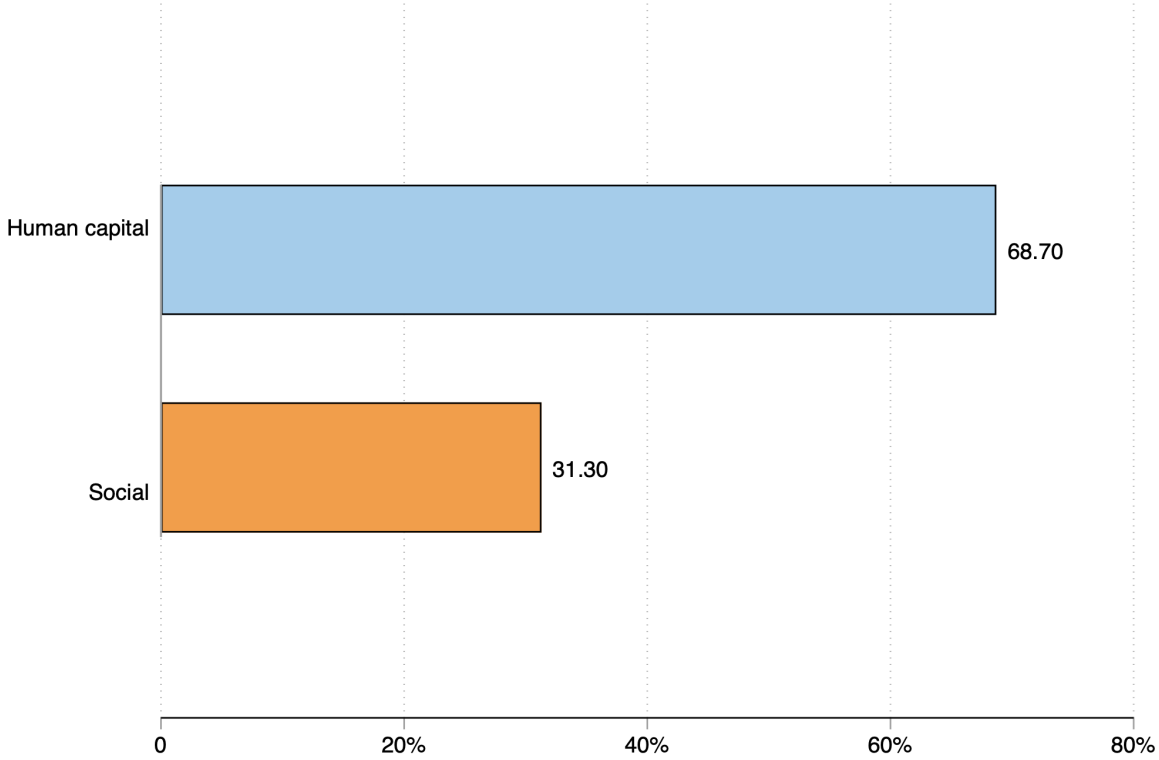


Panel B. Follow-up



Note. The two figures visualize the friendship network structure of one randomly chosen class in the study sample at the baseline (Panel A) and the follow-up (Panel B). The blue dots represent bullies, while the orange dots represent non-bullies. The size of each dot (network node) is scaled by in-degree, meaning the nomination links the student received. Larger dots indicate greater popularity.

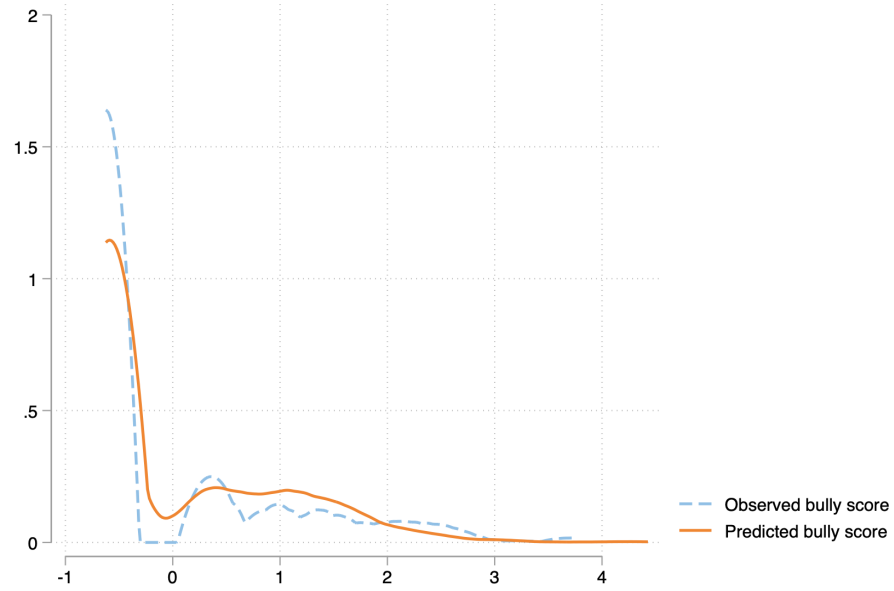
Figure 2: Decomposition Results



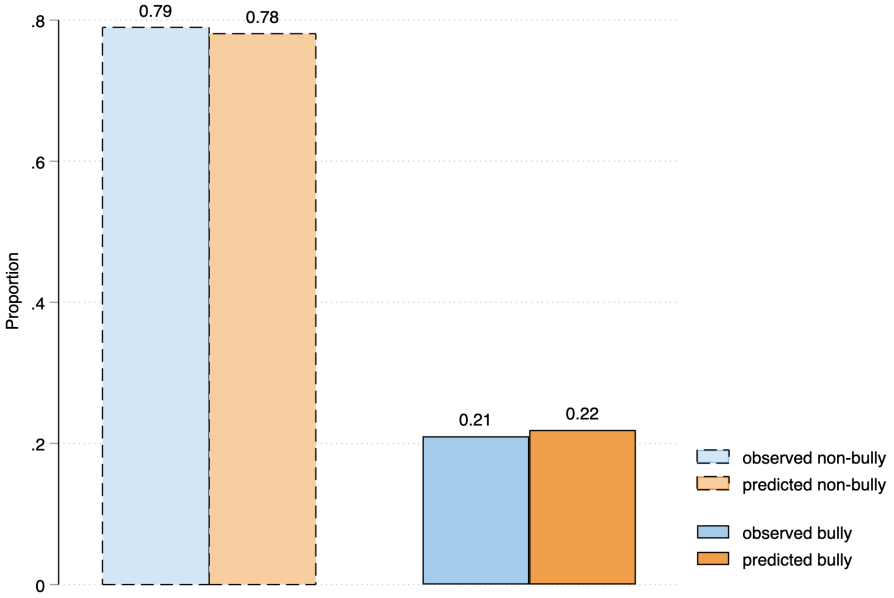
Note. The figure shows the results of the decomposition exercise aiming to obtain empathy's individual human capital and social effects on bullying perpetration. The human capital effect is defined as the change in own bullying score due to changes in own empathy. The social effect is defined as the change in own bullying score due to empathy-induced changes in peer networks and the associated spillovers from peers' bullying on own bullying score. I obtain the human capital effect by setting the peer effect parameter in Eq (16) equal to zero and simulating the new bullying score. The social effect is obtained by subtracting the simulated change in bullying scores when calculating the human capital effect from the total change in bullying scores, which is observed from our experimental data. The orange-filled box denotes the proportion of the total change in bullying scores explained by the social effect, while the blue-filled box denotes the remaining part explained by the human capital effect. The social effect accounts for 32% of the empathy effect, and it is almost identical to results calculated using the model estimates in Appendix Figure I5.

Figure 3: Model Goodness of Fit

Panel A. Density plot of bullying score

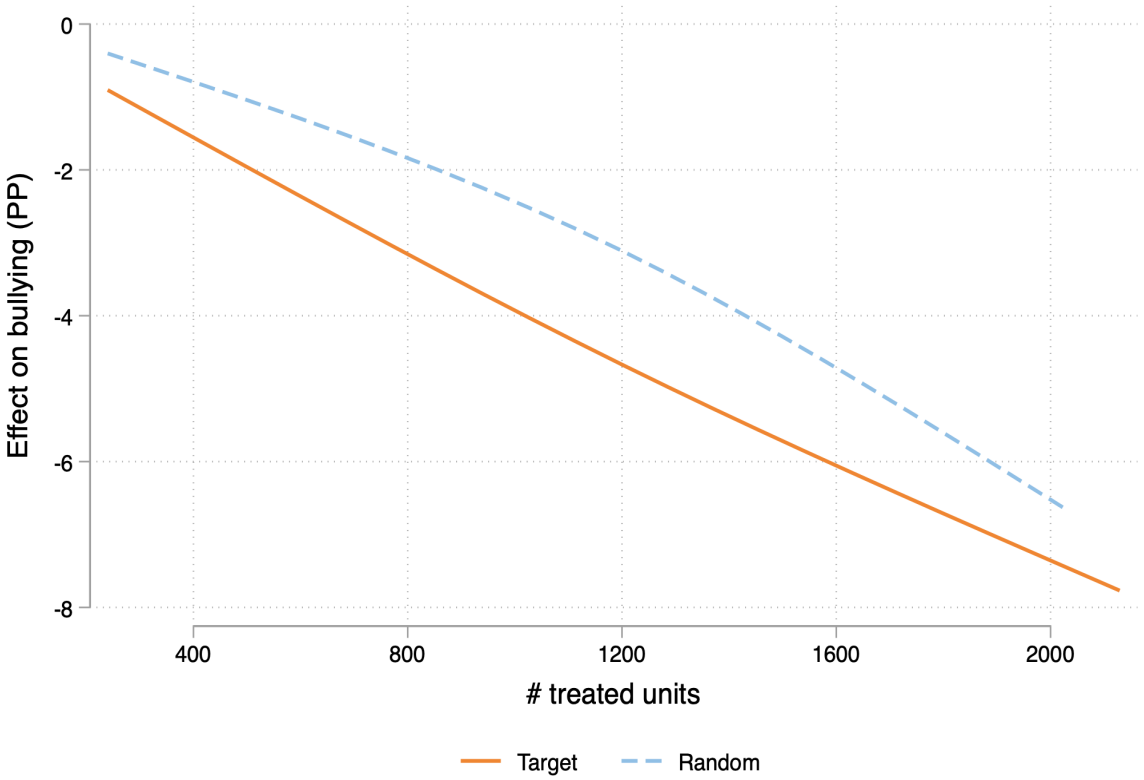


Panel B. Proportion of bullies and non-bullies



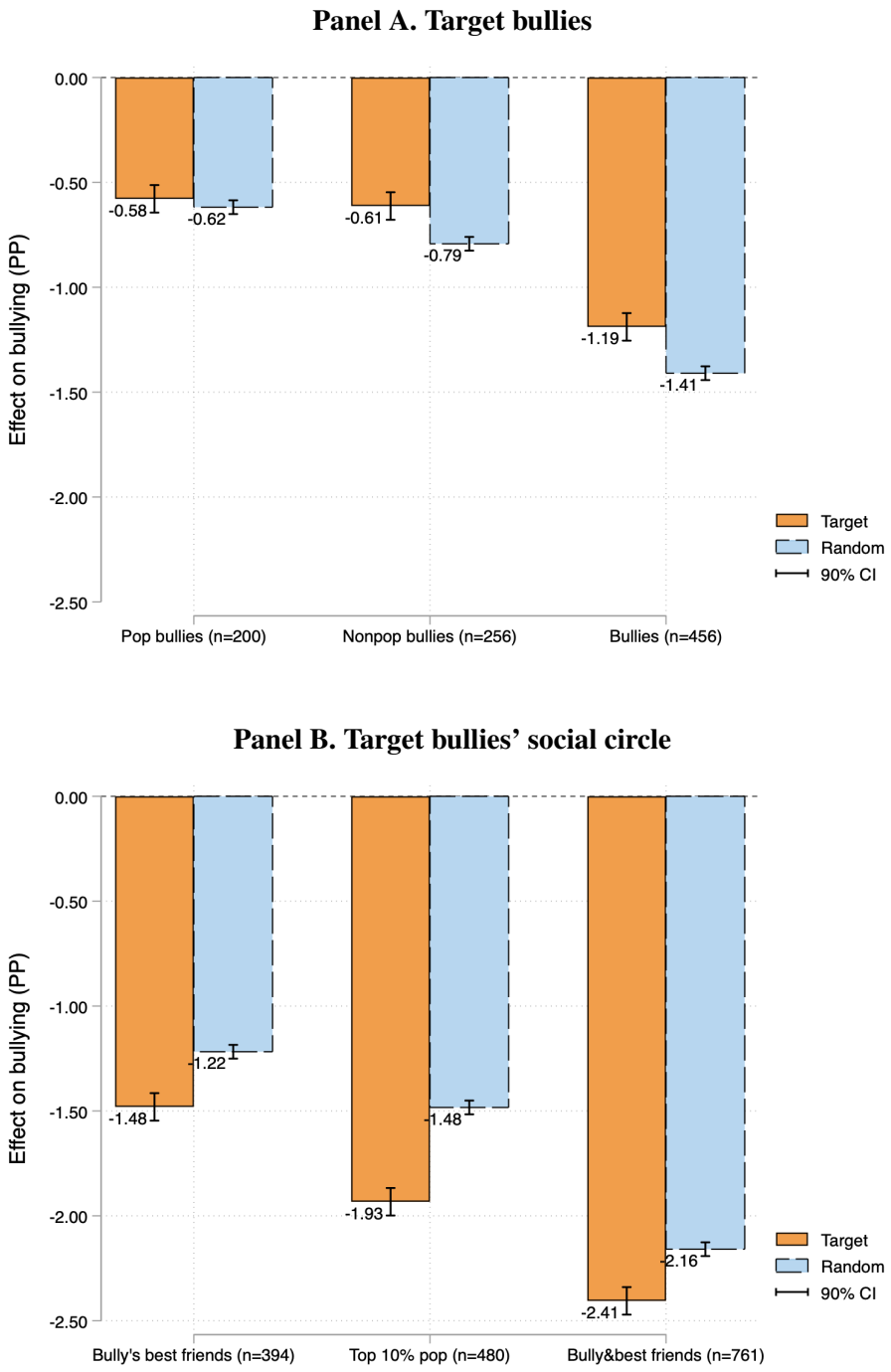
Note. The figures show the model fit. Panel A shows the density plot of the observed and predicted bullying score. Panel B shows the proportions of bullies and non-bullies from the data and the model. I use the `cutpointr` package in `R` to transform the continuous bullying score measures into dummy classifications. In Panel A, the orange solid line denotes the predicted bully score and the blue dashed line denotes the observed bully score. In Panel B, the blue dashed bar denotes the proportion of non-bully students observed in the data, while the orange dashed bar denotes the proportion of non-bully students predicted by the model. The blue solid bar denotes the proportion of bully students observed in the data, while the orange solid bar denotes the proportion of bully students predicted by the model.

Figure 4: Popularity-based Targeting VS. Random Assignment



Note. The figure plots the percentage-point (pp) decrease in the number of bullies for targeting experiments based on students' popularity and random treatment assignments. The x-axis denotes the number of treated units. The y-axis denotes the pp decrease in bullies. I use a blue dashed line to represent the results for the random treatment assignment, while the orange solid line denotes the results for the targeting experiment.

Figure 5: Alternative Targeting Experiments VS. Random Assignment



Note. The figures plot the percentage-point (pp) decrease in the number of bullies for alternative targeting and random assignment experiments. Panel A shows the results of targeting bullies, including targeting popular bullies, non-popular bullies, and bullies. Panel B shows the results of targeting bullies' social circle, including the top 10% most popular students, bullies' best friends, and bullies & best friends. The orange bar shows the results of targeting experiments considering imperfect compliance. The exact numbers are reported in Panel A of Table I15. The light blue bar shows the results of the corresponding random assignment, which assigns the same number of treated units for each scenario. Results of the random experiments are obtained from calculating the average bullying score reduction for 100 simulations of empathy $\times$ 10,000 simulations of network structure. I also plot the 90% confidence interval on top of each bar.

References

Agostinelli, Francesco, Matthias Doepke, Giuseppe Sorrenti, and Fabrizio Zilibotti, "It Takes a Village: The Economics of Parenting with Neighborhood and Peer Effects," Working Paper 27050, National Bureau of Economic Research April 2020. [6]

_ , _ , _ , and _ , "When the great equalizer shuts down: Schools, peers, and parents in pandemic times," *Journal of Public Economics*, 2022, *206*, 104574. [6]

Alan, Sule, Ceren Baysan, Mert Gumren, and Elif Kubilay, "Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking," *Quarterly Journal of Economics*, 03 2021, *136* (4), 2147–2194. [1, 11]

_ , Enes Duysak, Elif Kubilay, and Ipek Mumcu, "Social Exclusion and Ethnic Segregation in Schools: The Role of Teacher's Ethnic Prejudice," *Review of Economics and Statistics*, 2021, pp. 1–45. [6]

Álvarez-García, David, Trinidad García, and José Carlos Núñez, "Predictors of school bullying perpetration in adolescence: A systematic review," *Aggression and Violent Behavior*, 2015, *23*, 126–136. [7]

Anderson, Michael L, "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103* (484), 1481–1495. [11]

Anukriti, S, Catalina Herrera-Almanza, and Mahesh Karra, "Bring A Friend: Leveraging Financial and Peer Support to Improve Women's Reproductive Agency in India," 2023. Unpublished Manuscript. [29]

Attanasio, Orazio, Costas Meghir, and Emily Nix, "Human capital development and parental investment in India," *Review of Economic Studies*, 2020, *87* (6), 2511–2541. [6]

_ , Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina, "Estimating the production function for human capital: results from a randomized controlled trial in Colombia," *American Economic Review*, 2020, *110* (1), 48–85. [6]

Badev, Anton, "Nash equilibria on (Un)stable networks," *Econometrica*, 2021, *89* (3), 1179–1206. [19]

Ballester, Coralio, Antoni Calvó-Armengol, and Yves Zenou, "Who's who in networks. Wanted: The key player," *Econometrica*, 2006, *74* (5), 1403–1417. [7, 18, and 30]

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson, "Using gossips to spread information: Theory and evidence from two randomized controlled trials," *Review of Economic Studies*, 2019, *86* (6), 2453–2490. [1, 7]

Barrera-Osorio, Felipe, Paul Gertler, Nozomi Nakajima, and Harry Patrinos, "Promoting Parental Involvement in Schools: Evidence From Two Randomized Experiments," Working Paper 28040, National Bureau of Economic Research October 2020. [6]

Battaglini, Marco, Carlos Díaz, and Eleonora Patacchini, "Self-control and peer groups: An empirical analysis," *Journal of Economic Behavior & Organization*, 2017, *134*, 240–254. [7]

Beaman, Lori and Jeremy Magruder, "Who Gets the Job Referral? Evidence from a Social Networks Experiment," *American Economic Review*, December 2012, *102* (7), 3574–93. [1]

_ , **Ariel BenYishay, Jeremy Magruder, and Ahmed Mushfiq Mobarak**, "Can Network Theory-Based Targeting Increase Technology Adoption?," *American Economic Review*, June 2021, *111* (6), 1918–43. [1]

Becker, Gary S., "Investment in Human Capital: A Theoretical Analysis," *Journal of Political Economy*, 1962, *70* (5, Part 2), 9–49. [6]

_ **and Nigel Tomes**, "An equilibrium theory of the distribution of income and intergenerational mobility," *Journal of Political Economy*, 1979, *87* (6), 1153–1189. [6]

_ **and** _ , "Human capital and the rise and fall of families," *Journal of Labor Economics*, 1986, *4* (3, Part 2), S1–S39. [6]

Bennett, Magdalena and Peter Bergman, "Better Together? Social Networks in Truancy and the Targeting of Treatment," *Journal of Labor Economics*, 2021, *39* (1), 1–36. [1]

Bifulco, Robert, Jason M. Fletcher, and Stephen L. Ross, "The Effect of Classmate Characteristics on Post-secondary Outcomes: Evidence from the Add Health," *American Economic Journal: Economic Policy*, February 2011, *3* (1), 25–53. [11]

Boca, Daniela Del, Chiara Monfardini, and Cheti Nicoletti, "Parental and child time investments and the cognitive development of adolescents," *Journal of Labor Economics*, 2017, *35* (2), 565–608. [6]

_ , **Christopher Flinn, and Matthew Wiswall**, "Household Choices and Child Development," *Review of Economic Studies*, 2013, *81* (1), 137–185. [6]

Bono, Emilia Del, Marco Francesconi, Yvonne Kelly, and Amanda Sacker, "Early maternal time investment and early child outcomes," *The Economic Journal*, 2016, *126* (596), F96–F135. [6]

Booij, Adam S, Edwin Leuven, and Hessel Oosterbeek, "Ability Peer Effects in University: Evidence from a Randomized Experiment," *Review of Economic Studies*, September 2016, *84* (2), 547–578. [7]

Borgwald, K. and H. Theixos, "Bullying the bully: Why zero-tolerance policies get a failing grade," *Social Influence*, 2013, *8* (2-3), 149–160. [4, 29]

Boucher, Vincent, Carlo L. Del Bello, Fabrizio Panebianco, Thierry Verdier, and Yves Zenou, "Education Transmission and Network Formation," *Journal of Labor Economics*, 2023, *41* (1), 129–173. [11, 19]

_ , **F. Antoine Dedewanou, and Arnaud Dufays**, "Peer-induced beliefs regarding college participation," *Economics of Education Review*, 2022, *90*, 102307. [11]

_ , **Michelle Rendall, Philip Ushchev, and Yves Zenou**, "Toward a general theory of peer effects," May 2022. Unpublished Manuscript. [7, 10, and 17]

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin, "Identification of peer effects through social networks," *Journal of Econometrics*, 2009, *150* (1), 41–55. [3, 26, and 20]

Breza, Emily, Arun G. Chandrasekhar, Tyler H. McCormick, and Mengjie Pan, "Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data," *American Economic Review*, 2020, *110* (8), 2454–84. [5]

Brown, Sarah and Karl Taylor, "Bullying, education and earnings: evidence from the National Child Development Study," *Economics of Education Review*, 2008, *27* (4), 387–401. [1]

Bruyn, Eddy H De, Antonius HN Cillessen, and Inge B Wissink, "Associations of peer acceptance and perceived popularity with bullying and victimization in early adolescence," *Journal of Early Adolescence*, 2010, *30* (4), 543–566. [12]

Calvó-Armengol, Antoni, Eleonora Patacchini, and Yves Zenou, "Peer Effects and Social Networks in Education," *Review of Economic Studies*, 2009, *76* (4), 1239–1267. [7, 11]

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller, "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, 2008, *90* (3), 414–427. [13, 35, and 36]

Cappelen, Alexander, John List, Anya Samek, and Bertil Tungodden, "The effect of early-childhood education on social preferences," *Journal of Political Economy*, 2020, *128* (7), 2739–2758. [6]

Chein, Jason, Dustin Albert, Lia O'Brien, Kaitlyn Uckert, and Laurence Steinberg, "Peers Increase Adolescent Risk Taking by Enhancing Activity in the Brain's Reward Circuitry," *Developmental Science*, 2011, *14* (2), F1–F10. [1]

Chen, Yi, Rong Huang, Yuanping Lu, and Kangyi Zhang, "Education fever in China: Children's academic performance and parents' life satisfaction," *Journal of Happiness Studies*, 2021, *22*, 927–954. [9]

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz, "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment," *American Economic Review*, 2016, *106* (4), 855–902. [6]

Coleman, James, "Relational analysis: The study of social organizations with survey methods," *Human Organization*, 1958, *17* (4), 28–36. [11]

Comola, Margherita and Silvia Prina, "Treatment effect accounting for network changes," *Review of Economics and Statistics*, 2021, *103* (3), 597–604. [7]

Conley, Timothy G. and Christopher R. Udry, "Learning about a New Technology: Pineapple in Ghana," *American Economic Review*, 2010, *100* (1), 35–69. [5]

Cook, Clayton R, Kirk R Williams, Nancy G Guerra, Tia E Kim, and Shelly Sadek, "Predictors of bullying and victimization in childhood and adolescence: a meta-analytic investigation.," *School Psychology Quarterly*, 2010, *25* (2), 65. [7]

Cunha, Flavio and James Heckman, "The technology of skill formation," *American Economic Review*, 2007, *97* (2), 31–47. [1]

_ **and** _ , "Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation," *Journal of Human Resources*, 2008, *43* (4), 738–782. [6]

_ , **Eric Nielsen, and Benjamin Williams**, "The Econometrics of Early Childhood Human Capital and Investments," *Annual Review of Economics*, 2021, *13*, 487–513. [6]

_ , **James Heckman, and Susanne Schennach**, "Estimating the technology of cognitive and noncognitive skill formation," *Econometrica*, 2010, *78* (3), 883–931. [6]

_ , **Qinyou Hu, Yiming Xia, and Naibao Zhao**, "Reducing Bullying: Evidence from a Parental Involvement Program on Empathy Education," Working Paper 30827, National Bureau of Economic Research 2023. [2, 4, 8, 10, and 12]

Currarini, Sergio, Matthew O Jackson, and Paolo Pin, "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica*, 2009, *77* (4), 1003–1045. [7]

Dake, Joseph A, James H Price, and Susan K Telljohann, "The nature and extent of bullying at school," *Journal of School Health*, 2003, *73* (5), 173–180. [30]

Davis, Mark H., "Measuring individual differences in empathy: Evidence for a multidimensional approach.," *Journal of Personality and Social Psychology*, 1983, *44* (1), 113. [2, 11]

_ , *Empathy: A Social Psychological Approach*, 1 ed., Routledge, 1996. [1]

Debreu, Gerard and Israel Nathan Herstein, "Nonnegative square matrices," *Econometrica*, 1953, pp. 597–607. [18]

Decety, Jean, "The neurodevelopment of empathy in humans," *Developmental Neuroscience*, 2010, *32* (4), 257–267. [1]

_ **and Philip L Jackson**, "The functional architecture of human empathy," *Behavioral and Cognitive Neuroscience Reviews*, 2004, *3* (2), 71–100. [8]

Deming, David J., "The growing importance of social skills in the labor market," *Quarterly Journal of Economics*, 2017, *132* (4), 1593–1640. [1, 6]

_ , "Four Facts about Human Capital," *Journal of Economic Perspectives*, 2022, *36* (3), 75–102. [1, 6]

Doyle, Orla, "The First 2,000 Days and Child Skills," *Journal of Political Economy*, 2020, *128* (6), 2067–2122. [6]

Espelage, Dorothy L, "Ecological theory: Preventing youth bullying, aggression, and victimization," *Theory into Practice*, 2014, *53* (4), 257–264. [12]

_ , **Melissa K Holt, and Rachael R Henkel**, "Examination of peer-group contextual effects on aggression during early adolescence," *Child Development*, 2003, *74* (1), 205–220. [12]

Findley, Michael G, Kyosuke Kikuta, and Michael Denly, "External validity," *Annual Review of Political Science*, 2021, *24*, 365–393. [32]

Frith, Chris D and Uta Frith, "Social Cognition in Humans," *Current Biology*, 2007, *17* (16), R724–R732. [1]

Graham, Bryan S., "An econometric model of network formation with degree heterogeneity," *Econometrica*, 2017, *85* (4), 1033–1063. [19, 22, and 23]

_ , "Identifying and Estimating Neighborhood Effects," *Journal of Economic Literature*, 2018, *56* (2), 450–500. [6]

Griffith, Alan, "Name Your Friends, but Only Five? The Importance of Censoring in Peer Effects Estimates Using Social Network Data," *Journal of Labor Economics*, 2022, *40* (4), 779–805. [26]

_ , "Random Assignment with Non-Random Peers: A Structural Approach to Counterfactual Treatment Assessment," *Review of Economics and Statistics*, 05 2022, pp. 1–40. [7, 19]

Haushofer, Johannes, Paul Niehaus, Carlos Paramo, Edward Miguel, and Michael W Walker, "Targeting Impact versus Deprivation," Working Paper 30138, National Bureau of Economic Research 2022. [1]

Hazler, Richard J, Dina L Miller, Jolynn V Carney, and Suzy Green, "Adult recognition of school bullying situations," *Educational Research*, 2001, *43* (2), 133–146. [10]

Heckman, James and Guilherme Sedlacek, "Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market," *Journal of Political Economy*, 1985, *93* (6), 1077–1125. [3, 20]

_ **and Tim Kautz**, "Hard evidence on soft skills," *Labour Economics*, 2012, *19* (4), 451–464. [1]

_ **, Bridget Galaty, and Haihan Tian**, "The Economic Approach to Personality, Character and Virtue," Working Paper 31258, National Bureau of Economic Research 2023. [1]

_ **, Jora Stixrud, and Sergio Urzua**, "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor Economics*, 2006, *24* (3), 411–482. [1]

Hinduja, Sameer and Justin W Patchin, "Bullying, cyberbullying, and suicide," *Archives of Suicide Research*, 2010, *14* (3), 206–221. [1]

Hodges, Sara and M.W. Myers, "Encyclopedia of social psychology," *Empathy*, 01 2007, pp. 296–298. [1]

Hogan, Robert, "Development of an empathy scale.," *Journal of Consulting and Clinical Psychology*, 1969, *33* (3), 307. [1]

Hsieh, Chih-Sheng, Michael D. König, and Xiaodong Liu, "A Structural Model for the Coevolution of Networks and Behavior," *Review of Economics and Statistics*, 03 2022, *104* (2), 355–367. [19]

Huang, Hui, Jun Sung Hong, and Dorothy L Espelage, "Understanding factors associated with bullying and peer victimization in Chinese schools within ecological contexts," *Journal of Child and Family Studies*, 2013, *22*, 881–892. [10]

Islam, Asad, Michael Vlassopoulos, Yves Zenou, and Xin Zhang, "Centrality-based spillover effects," *CEPR Discussion Paper No. DP16321*, 2021. [7]

Jackson, Matthew O, Stephen M Nei, Erik Snowberg, and Leeat Yariv, "The Dynamics of Networks and Homophily," Working Paper 30815, National Bureau of Economic Research 2022. [7]

Jolliffe, Darrick and David P. Farrington, "Examining the relationship between low empathy and bullying," *Aggressive Behavior*, 2006, *32* (6), 540–550. [7]

Jones, Damon E, Mark Greenberg, and Max Crowley, "Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness," *American Journal of Public Health*, 2015, *105* (11), 2283–2290. [1]

Kahneman, Daniel, "Maps of Bounded Rationality: Psychology for Behavioral Economics," *American Economic Review*, 2003, *93* (5), 1449–1475. [18]

Kamas, Linda and Anne Preston, "Empathy, gender, and prosocial behavior," *Journal of Behavioral and Experimental Economics*, 2021, *92*, 101654. [11]

Kardos, Peter, Bernhard Leidner, Csaba Pléh, Péter Soltész, and Zsolt Unoka, "Empathic people have more friends: Empathic abilities predict social network size and position in social network predicts empathic efforts," *Social Networks*, 2017, *50*, 1–5. [14, 25]

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman, "Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment," *Quarterly Journal of Economics*, 05 2001, *116* (2), 607–654. [6]

Kautz, Tim, James Heckman, Ron Diris, Bas ter Weel, and Lex Borghans, "Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success," Working Paper 20749, National Bureau of Economic Research 2014. [1]

Klomek, Anat Brunstein, Frank Marrocco, Marjorie Kleinman, Irvin S Schonfeld, and Madelyn S Gould, "Bullying, depression, and suicidality in adolescents," *Journal of the American Academy of Child & Adolescent Psychiatry*, 2007, *46* (1), 40–49. [1]

König, Michael D, Xiaodong Liu, and Yves Zenou, "R&D networks: Theory, empirics, and policy implications," *Review of Economics and Statistics*, 2019, *101* (3), 476–491. [3, 7, 17, 19, and 20]

Kosse, Fabian, Thomas Deckers, Pia Pinger, Hannah Schildberg-Hörisch, and Armin Falk, "The formation of prosociality: causal evidence on the role of social environment," *Journal of Political Economy*, 2020, *128* (2), 434–467. [1]

Lam, Chun Bun, Susan M McHale, and Ann C Crouter, "Time with peers from middle childhood to late adolescence: Developmental course and adjustment correlates," *Child Development*, 2014, *85* (4), 1677–1693. [1]

Lavy, Victor and Analia Schlosser, "Mechanisms and Impacts of Gender Peer Effects at School," *American Economic Journal: Applied Economics*, 2011, *3* (2), 1–33. [26]

Lee, Lung-Fei, Xiaodong Liu, Eleonora Patacchini, and Yves Zenou, "Who is the Key Player? A Network Analysis of Juvenile Delinquency," *Journal of Business & Economic Statistics*, 2021, *39* (3), 849–857. [7, 19, 20, and 24]

Li, Jiameng, Aissata Mahamadou Sidibe, Xiaoyun Shen, and Therese Hesketh, "Incidence, risk factors and psychosomatic symptoms for traditional bullying and cyberbullying in Chinese adolescents," *Children and Youth Services Review*, 2019, *107*, 104511. [7]

List, John A, "Non est disputandum de generalizability? a glimpse into the external validity trial," Technical Report, National Bureau of Economic Research 2020. [32]

_ , Fatemeh Momeni, Michael Vlassopoulos, and Yves Zenou, "Neighborhood Spillover Effects of Early Childhood Interventions," 2023. Unpublished Manuscript. [6]

Liu, Xiaodong, Eleonora Patacchini, and Yves Zenou, "Endogenous peer effects: local aggregate or local average?," *Journal of Economic Behavior & Organization*, 2014, *103*, 39–59. [17]

Lleras-Muney, Adriana, Matthew Miller, Shuyang Sheng, and Veronica T Sovero, "Party On: The Labor Market Returns to Social Networks and Socializing," Working Paper 27337, National Bureau of Economic Research June 2020. [11, 17]

Lord, Frederic M, *Applications of item response theory to practical testing problems*, Routledge, 2012. [10, 17]

Loveless, Tom, Robert M Costrell, and Larry Cuban, "Test-based accountability: The promise and the perils," *Brookings Papers on Education Policy*, 2005, (8), 7–45. [8]

Manski, Charles F., "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 1993, *60* (3), 531–542. [3, 24]

Murphy, David M.A., "Sobriety, social capital, and village network structures," *World Development*, 2023, *166*, 106210. [29]

Orben, Amy, Livia Tomova, and Sarah-Jayne Blakemore, "The effects of social deprivation on adolescent development and mental health," *The Lancet Child & Adolescent Health*, 2020, *4* (8), 634–640. [1]

Paluck, Elizabeth Levy and Hana Shepherd, "The salience of social referents: a field experiment on collective norms and harassment behavior in a school social network.," *Journal of Personality and Social Psychology*, 2012, *103* (6), 899. [4, 28]

_ , _ , and Peter M. Aronow, "Changing climates of conflict: A social network experiment in 56 schools," *Proceedings of the National Academy of Sciences*, 2016, *113* (3), 566–571. [4, 28]

Peng, Sida, "Heterogeneous endogenous effects in networks," *arXiv Preprint arXiv:1908.00663*, 2019. [7]

Perkins, H Wesley, David W Craig, and Jessica M Perkins, "Using social norms to reduce bullying: A research intervention among adolescents in five middle schools," *Group Processes & Intergroup Relations*, 2011, *14* (5), 703–722. [17]

Preston, Stephanie D and Frans BM De Waal, "Empathy: Its ultimate and proximate bases," *Behavioral and Brain Sciences*, 2002, *25* (1), 1–20. [8]

Qu, Xi and Lung-Fei Lee, "Estimating a spatial autoregressive model with an endogenous spatial weight matrix," *Journal of Econometrics*, 2015, *184* (2), 209–232. [26]

Reyna, Valerie F. and Frank Farley, "Risk and Rationality in Adolescent Decision Making: Implications for Theory, Practice, and Public Policy," *Psychological Science in the Public Interest*, 2006, *7* (1), 1–44. PMID: 26158695. [19]

Roy, Andrew Donald, "Some thoughts on the distribution of earnings," *Oxford Economic Papers*, 1951, *3* (2), 135–146. [3, 20]

Ryan, Shannon V, Nathaniel P von der Embse, Laura L Pendergast, Elina Saeki, Natasha Segool, and Shelby Schwing, "Leaving the teaching profession: The role of teacher stress and educational accountability policies on turnover intent," *Teaching and Teacher Education*, 2017, *66*, 1–11. [9]

Şahin, Mustafa, "An investigation into the efficiency of empathy training program on preventing bullying in primary schools," *Children and Youth Services Review*, 2012, *34* (7), 1325–1330. [29]

Santavirta, Torsten and Miguel Sarzosa, "Effects of disruptive peers in endogeneous social networks," 2019. Unpublished Manuscript. [17, 23]

Sarzosa, Miguel, "Victimization and Skill Accumulation: The Case of School Bullying," *Journal of Human Resources*, 2021, pp. 0819–10371R2. [1]

_ **and Sergio Urzúa**, "Bullying among adolescents: The role of skills," *Quantitative Economics*, 2021, *12* (3), 945–980. [1]

Steinberg, Laurence, "A social neuroscience perspective on adolescent risk-taking," in "Biosocial Theories of Crime," Routledge, 2017, pp. 435–463. [1]

_ **and Amanda Sheffield Morris**, "Adolescent Development," *Annual Review of Psychology*, 2001, *52* (1), 83–110. PMID: 11148300. [1]

Steinberg, Laurence D, *Age of opportunity: Lessons from the new science of adolescence*, Houghton Mifflin Harcourt, 2014. [1]

Tippett, Neil and Dieter Wolke, "Socioeconomic status and bullying: A meta-analysis," *American Journal of Public Health*, 2014, *104* (6), e48–e59. [7]

Todd, Petra E. and Kenneth I. Wolpin, "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility," *American Economic Review*, 2006, *96* (5), 1384–1417. [27]

UNESCO, *Behind the numbers: Ending school violence and bullying*, UNICEF, 2019. [1]

Wang, Lei, Yiwei Qian, Nele Warrinnier, Orazio Attanasio, Scott Rozelle, and Sean Sylvia, "Parental investment, school choice, and the persistent benefits of an early childhood intervention," *Journal of Development Economics*, 2023, *165*, 103166. [6]

Wölfer, Ralf, Kai S. Cortina, and Jürgen Baumert, "Embeddedness and empathy: How the social network shapes adolescents' social understanding," *Journal of Adolescence*, 2012, *35* (5), 1295–1305. [7]

Zenou, Yves, "Key players," *Oxford Handbook on the Economics of Networks*, 2016, pp. 244–274. [7, 30]

APPENDIX

A More Details on Measures

A.1 Friendship Network Statistics

To rigorously examine how social status in friendship networks relates to bully status, I construct five measures of social status based on the network literature.[50]

In-degree Centrality In-degree centrality measures the number of nominations a student received in the friendship network survey. This measure is also equivalent to a measure of popularity.

Eigenvector Centrality In addition to only counting the number of links one student received, eigenvector centrality also accounts for the status of individuals who are connected to the student. Eigenvector centrality is calculated by taking the sum of the centrality of the connected individuals and multiplying it by a scalar. Thus, a student who is nominated as a friend by three students with relatively more ties has a higher eigenvector centrality than those who are nominated by three less connected students. Equivalently, eigenvector centrality is a measure of connectedness to other well-connected individuals (Alan et al., 2021b). This centrality measure is a number between 0 and 1, with a higher value indicating more connectedness. The mean of the measure in the study sample is 0.21.

Reciprocal Link In this directed friendship network, a reciprocal link is denoted as 1 when both individuals, denoted as i and j, list each other as friends; otherwise, it is 0. The average number of reciprocal links in the study sample is 1.36, ranging from 0 to a maximum of 5 ties due to the 5-friendship nomination limit.

Unilateral Link As there may be an asymmetry between students i and j's nominations, I then construct a measure of unilateral links following Lin and Weinberg (2014). Unilateral link is counted as 1 if person i lists person j as a friend but person j does not nominate i as a friend, or if person i does not list person j as a friend but person j nominates i as a friend; otherwise, unilateral link equals 0. The average number of reciprocal links in the study sample is 1.75, with a minimum of zero and a maximum of 7 ties.

Coleman Index I look at the distribution of the quality of the social circles. Specifically, I construct the *Coleman index* to measure the degree of homophily to depict link quality (Coleman, 1958).[51] Since the theme of the intervention is on cultivating empathy. I thus calculate the *Coleman index* for high and low empathy, respectively. I classify students as low- or high-empathetic by comparing their empathy levels with the classroom-level median empathy level. A student is classified as having low empathy if he/she has lower empathy skills than the classroom median. The formal definition of the *Coleman index* is as follows: Let L and H

[50] Note that all these measures are constructed using only the baseline data.

[51] The same methodology is used by Alan et al. (2021b) who study social exclusion and ethnic segregation among host children and refugee children in Turkish schools.

denote low- and high-empathy students, respectively. Denote the number of within-group ties of group i in classroom j as s_{ij}, and the total number of ties of group i in classroom j as t_{ij}, where $i \in \{L,H\}$. Then, s_{ij}/t_{ij} gives the share of within-group (homophilic) ties for group i. Let w_{ij} denote the population share of group i in classroom j. Then, *Coleman's Homophily Index* for group $i \in \{L,H\}$ in classroom j is:

$$C_{ij} = \frac{\frac{s_{ij}}{t_{ij}} - w_{ij}}{1 - w_{ij}}.$$

Intuitively, the *Coleman index* compares the observed friendship links with links formed at random. Moreover, it takes into account groups with very large size w_{ij} and a larger *Coleman index* indicates more segregation.[52] Note that since we have two groups (low- and high-empathy students), the *Coleman index* gives us two normalized excess homophily scores, one for each group. Appendix Figure I8 shows two examples of classroom friendship networks by students' empathy level at the baseline. I scale the size of the network node by in-degree, that is, the nomination links the student received. The figure on the left panel depicts a friendship network with higher segregation for low-empathy students than for high-empathy students. The figure on the right panel depicts a friendship network with higher segregation for high-empathy students than for low-empathy students.

A.2 Students' Outcome Measures

The detailed measures and questions for children's outcomes are listed below:

- School bullying: Students' self-report, using a 5-point Likert scale, on whether they threatened others, physically bullied (hitting/kicking) others, spread rumors about others, socially isolated others, and cyberbullied (abusive or hurtful texts online) others during the semester of the intervention. The frequency is coded as 1) never, 2) once, 3) two or three times, 4) once or twice a month, and 5) at least once a week. Similarly, we also record whether they were bullying victims of any of these behaviors. The detailed questions are specified as follows: "1) During this past semester, has anyone spread rumors or false information about you at school, trying to make other students dislike you? 2) During this past semester, have you ever been threatened by classmates? 3) During this past semester, have any classmates pushed you, hit you, or intentionally tripped you? 4) During this past semester, have any classmates ever verbally abused you online, such as on gaming platforms, school forums, or public discussion boards? and 5) During this past semester, have any classmates deliberately excluded you from certain activities and prevented you from participating?" In addition, they were asked whether they had ever witnessed school bullying incidents in the follow-up survey. We

[52] In cases where there is excess heterophily, i.e. $s_{ij}/t_{ij} < w_{ij}$, the measure is normalized by w_{ij} instead of $1 - w_{ij}$. This adjustment ensures that the *Coleman Homophily Index* is between -1 and 1.

also asked them whether they would help those bullying victims when they witness a bullying incident.

- Empathy: To avoid a longsome questionnaire, we use a 9-item empathy measurement to explore two dimensions of empathetic concerns and perspective taking, which is also used in Alan et al. (2021a). For most items, we use a 7-point Likert scale for both baseline and follow-up surveys: from completely disagree (1) to completely agree (7). We add another dimension, prosociality, to the follow-up survey. The questions include the hypothetical scenarios about helping other children in difficulty, doing others a favor, helping my mother do housework, becoming a charitable person, and rescuing a drowning child adapted from the official guide from Centers for Disease Control (CDC) (Dahlberg et al., 2005). For each scenario, we ask students whether they have ever imagined the scenarios, and they are asked to choose (1) Never; (2) Sometimes; or (3) Very frequent.

- Mental health: Mental health is measured using the 10-item Center for the Epidemiological Studies of Depression Short Form, or CES-D-10, is a 10-item Likert scale questionnaire (Yang et al., 2018).[53] The depression indicator is generated with a threshold value of 12. The *inverse CESD index* is constructed by 30 minus the CESD score so that a higher score indicates better mental health status. We construct a happiness score using a scale of 1-7, with 7 being the happiest.

- Stress score: We elicit students' stress by three categories of sources: (i) studies at school; (ii) peer relationships; and (iii) rank/test scores in the class. For each item, we use a 7-point Likert scale for both baseline and follow-up surveys: from the least stressed (1) to the most stressed (7). We then construct the *inverse stress index* consisting of these three components so that a higher score indicates less stress.

- Positive personality: The positive personality measure is composed of two dimensions: positive self-image and perseverance. Four aspects of self-image were measured by four single-item questions: (i) I am satisfied with myself (self-satisfied); (ii) I have many valuable traits (self-worth); (iii) I can do well in most cases (self-confident); (iv) I am not worse than others and proud of myself (self-esteem). For each item, we use a 7-point Likert scale for both baseline and follow-up surveys: from completely disagree (1) to completely agree (7). In the empirical analysis, we use inverse covariance matrix weighting methods to construct the self-esteem index that includes these four components. For perseverance, we ask students whether they agree or disagree with the statement "Frustration and difficulty will not stop me from reaching my goals." We use a 7-point Likert scale for both baseline and follow-up surveys: from completely disagree (1) to completely agree (7). We then construct the *Positive personality index* consisting of the self-esteem index and perseverance score, so that a higher score indicates a more positive personality.

[53] The items are also employed in China Family Panel Studies (CFPS) 2012 survey.

- Time with parents: As a cross-check of parental time investment, we ask students to count the total number (ranges from 0 - 7) of each activity that parents have been involved in a normal week in the past semester. The activities include having dinner, talking/discussing school life, watching TV, checking homework, and playing outdoor activities.

- Time spent on study and leisure per day: We ask for students' time use on a range of activities per day in a normal week in the past semester. To measure time spent on study, we sum up the time they reported on doing homework and attending tutoring classes. To measure time spent on leisure, we sum up the time they reported on playing video games, hanging out with friends, and doing workouts. The time use is measured in minutes.

B Findings From Previous Analyses

The findings from Cunha et al. (2023), who evaluate the program's effects on students' outcomes, help establish the causal link between adolescents' empathy skills and bullying reduction. Specifically, we find that the empathy development program is effective in improving students' empathy skills. We also find that the intervention is effective in reducing bullying across various dimensions, including bullies (4 percentage-point), victims (5 percentage-point), and witnesses (6 percentage-point), as summarized in Appendix Table I1. Moreover, data exported from the app shows that 71% of eligible parents enrolled ($N = 872$) and the overall take-up or participation rate of eligible parents is 41% ($N = 495$).[54]

While our previous study briefly examines the social aspect through an investigation of spillover effects, we find no evidence of spillovers on bullying reduction between classes, thereby confirming the Stable Unit Treatment Value Assumption (SUTVA). However, we observe a significant spillover effect of bullying reduction within classes, specifically from students who complied with the program to those who did not. This finding underscores the significance of exploring the social dynamics involved in the empathy effect on school bullying.

C A Simple Theoretical Framework

De Vignemont and Singer (2006) suggests that empathy has two roles. One is the epistemological role in the sense that empathy helps people understand others' feelings. Another is the social role which makes people internalize others' feelings. Therefore, empathy skills can shape an individual's utility function by changing their utility gains from peers. For instance, a student with high empathy can experience a loss in utility from knowing peers' suffering. Additionally, an empathetic student endogenizes friends' utility and may change his/her behavior to friends' preferences/behaviors. In this section, the Psychology insights are translated into Economics terms using the utility maximization framework.

[54] We define take-up or participation as completing at least half of the tasks.

C.1 Model Human Capital Effect of Empathy

One possible explanation is that empathy skills, which consist of perspective-taking, empathetic concerns, and prosociality, may directly affect the utility of bullying others. Let peer victimization enter one's utility function, and bullying others can make the student worse off if the student's empathy is high enough. One can construct the following utility specification for student i:

$$U_i = U(l_i, s_i, s_{-i}),$$

where l_i denotes leisure, s_i denotes my own status, and s_{-i} denotes peers' status.

Assuming the social status s_i has a formation function following

$$s_i = S(t_i, b_i, v_i),$$
$$s_{-i} = \frac{1}{N-1}\sum_{j\neq i} S(t_j, b_j, v_j),$$

where t_i is student i's test score, b_i is i's bullying effort, measured by time spent on bullying, v_i is the severity level of victimization, and N is the total number of students in the class.

To make it concrete, suppose l_i, s_i, and s_{-i} additively enter the utility function and let empathy affect the parameter ϕ, which is the marginal utility gained from the peer's status. The other two parameters, α and β, capture one's utility from leisure and own status, respectively. It gives:

$$U_i = \alpha l_i + \beta s_i + \phi s_{-i},$$

subject to the time constraint $T_i = l_i + b_i$.

The marginal utility with respect to the amount of effort on bullying, b, is derived as:

$$\frac{dU_i}{db_i} = -\alpha + \beta\frac{ds_i}{db_i} + \underbrace{\phi\frac{ds_{-i}}{dv_{-i}}\frac{dv_{-i}}{db_i}}_{\text{human capital effect}}, \tag{17}$$

where the $\phi\frac{ds_{-i}}{dv_{-i}}\frac{dv_{-i}}{db_i}$ explains the reduction in school bullying by improving empathy skills ϕ, which lower the marginal utility of bullying others, as $\frac{ds_{-i}}{dv_{-i}}$ is negative and $\frac{dv_{-i}}{db_i}$ is positive. Hence, we reach the first implication:

Proposition 1. A higher level of empathy implies a larger ϕ, and subsequently leads to a lower bullying rate as students internalize peers' victimization.

This implication summarizes the general explanation from the psychology literature of empathy's effect on bullying reduction in economic terms.

C.2 Model Social Effect of Empathy

Empathy may affect one's behavior through the friend circle. To consider the social sides of empathy, we assume that the student not only cares about peers' statuses but also cares about their good friends' utility. The difference between peers and friends is that individual i only knows peers' statuses, but they know more information about their good friends, so we assume the student internalizes their good friends' utility in their own utility function. In a simple situation with one good friend j, the student i's utility function takes the following form:

$$U_i = U(l_i, s_i, s_{-i}, U_j),$$

where U_j is good friend $j's$ utilities. Using the same social status formation function and assuming linear additive of the inputs of the utility function, we have:

$$U_i = \alpha l_i + \beta s_i + \phi s_{-i} + \lambda U_j.$$

We simplify the chains and assume that friend j's utility U_j is not reciprocal. The marginal utility with respect to the bullying effort b takes the following form:

$$\frac{dU_i}{db_i} = -\alpha + \beta \frac{ds_i}{db_i} + \underbrace{\phi \frac{ds_{-i}}{dv_{-i}} \frac{dv_{-i}}{db_i}}_{\text{human capital effect}} + \underbrace{\lambda \frac{dU_j}{ds_{-i}} \frac{ds_{-i}}{dv_{-i}} \frac{dv_{-i}}{db_i}}_{\text{social effect}}.$$

The third item $\lambda \frac{dU_j}{ds_{-i}} \frac{ds_{-i}}{dv_{-i}} \frac{dv_{-i}}{db_i}$ suggests that friend j also cares about peers $-i$'s victimization which further lowers i's marginal utility from increasing bullying effort.

In this case, empathy may affect the bullying decision through the human capital effect $\phi \frac{ds_{-i}}{dv_{-i}} \frac{dv_{-i}}{db_i}$ and the social effect $\lambda \frac{dU_j}{ds_{-i}} \frac{ds_{-i}}{dv_{-i}} \frac{dv_{-i}}{db_i}$. The parameter ϕ depends on student i's empathy level and captures the effect on reduction in bullying as a result of empathizing peers' potential victimization. Or one could consider that empathy can directly lower i's utility through their own empathetic concerns about victims. The parameter λ captures the social effect as it implies that empathy makes student i internalize his/her friend j's empathetic concerns about victims. The social effects accumulate as the number of good friends increases. In sum, empathy skill controls λ and controls for the marginal disutility from the reduction in friend's utility as a result of bullying perpetration, and we have the second implication:

Proposition 2. A higher level of empathy implies a larger λ, and subsequently leads to a lower bullying rate as students internalize friends' disutility from seeing victimization.

Unlike the first implication, Proposition 2 suggests the existence of the social side of empathy's effect on bullying reduction.

D Additional Evidence

D.1 Understanding the Network Changes

Appendix Table I6 helps to understand the ITT on network changes by examining the effects on the number of new or old friends who are bullies or non-bullies.

Panel A suggests that the treatment leads to a small and insignificant decrease in the number of friends who are bullies. Either making fewer new bully friends or maintaining friendships with fewer old bully friends can help explain this change.

In Panel B, I evaluate the effect through the angle of non-bullies. On average, I observe a 0.28 unit increase in making non-bully friends at the follow-up. Though not significant, I find that a larger proportion of the effects is driven by having non-bully old friends.

In sum, the ITT estimates shown in Appendix Table I6 suggest that the program slightly reduces the number of bully friends but significantly increases the number of non-bully friends.

D.2 Program Effects on Homophily

Empathy makes people more inclusive, especially to individuals from diverse backgrounds (Wölfer et al., 2012). Therefore, I continue to examine whether the intervention affects homophily, namely, whether the intervention encourages students to make friends who are less similar. I explore the effects across several other dimensions in addition to bully status: gender, high/low achiever (based on test scores), *hukou* status, college aspiration, self-esteem, perseverance, self-confidence, and empathy.[55] Appendix Table I9 shows the results — there is no significant change across all the examined dimensions after the intervention. The insignificant effects on homophily indicate that changing the quality of links, especially the bond of "best friend," takes time, and the construction of the social capital measured by the quantity of links might drive more short-run effects.

E Item Response Theory and Related Models

In the survey, students' involvement in bullying was measured using a detailed frequency scale ranging from 0 (*never*) to 4 (*at least once a week*). However, directly using the raw frequency score is problematic as it is difficult to interpret the meaning of the gap between two frequency scores.[56] To address this issue, I employed Item Response Theory (IRT), a method commonly used in psychology introduced by Lord (2012). First, I generate binary indicators of bullying involvement, assigning a value of one if an individual engages in at least one bullying event more than once, and zero otherwise. In the second step, I apply the two-parameter Logistic model

[55] I construct a *Homophily index* following Currarini et al. (2009), which has a similar flavor as the *Coleman index* but is built at the individual level.

[56] For instance, comparing 2 with 3 may deliver different meaning about the bullying involvement rates from comparing 3 with 4.

from IRT to map these binary indicators to latent bullying scores. This method considers the probability of correctly answering a question as a function of various question characteristics, including difficulty levels and a respondent's latent bullying effort.

Let Y_{ij} represent the response to question i (i.e., a particular type of bullying event i) from student j. Note that I use binary measures of bullying involvement for a more balanced distribution of the latent bullying effort. $Y_{ij} = 1$ denotes student i ever involved in a particular type of bullying event, and vice versa. Consider a Two-parameter Logistic (2PL) model with parameters $\Gamma \equiv (\kappa, \lambda)$, the probability of student j with latent bullying effort ζ_j providing a correct response to the item i is given by:

$$\Pr\left(Y_{ij} = 1 \mid \Gamma, \zeta_j\right) = \frac{\exp\left\{\kappa_i\left(\zeta_j - \lambda_i\right)\right\}}{1 + \exp\left\{\kappa_i\left(\zeta_j - \lambda_i\right)\right\}},$$

where κ_i represents discrimination, and λ_i represents the difficulty of item i. This setup implies that in the neighborhood of a given difficulty level, questions with higher discrimination indicate that two students with distinct bullying efforts would have different predicted probabilities of responding to their involvement in bullying.

Students' latent bullying efforts ζ are assumed to have a standard Normal distribution. Then the student j 's contribution to the likelihood is:

$$\mathbb{L}_j(\Gamma) = \int_{\zeta_j} \prod_{i=1}^{I} \Pr\left(Y_{ij} \mid \Gamma, \zeta_j\right)^{\mathbb{1}\{Y_{ij}=1\}} \left[1 - \Pr\left(Y_{ij} \mid \Gamma, \zeta_j\right)\right]^{\mathbb{1}\{Y_{ij}=0\}} \phi\left(\zeta_j\right) d\zeta_j,$$

where the integral is with respect to latent bullying effort ζ_j and is approximated numerically. For estimation, I recover the expected latent bullying efforts for each student in the data using the empirical likelihood and the score density via empirical Bayesian updating after estimating the Two-parameter Logistic model.

F Estimation Details of Parents' Problem

To obtain the estimates of variances of the shocks, we also need to rely on the second-order moments from the data. Specifically, for those who choose to participate:

$$\begin{aligned} Var(H_{i,1}|H_{i,0}, C_i, X_i, P_i = 1) &= Var(\varepsilon_i | \vartheta_i < \bar{h}) \\ &= Var\left(\frac{\sigma_{\varepsilon v}}{\sigma_\vartheta^2}\vartheta_i + \eta_i | \vartheta_i < \bar{h}\right) = \frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^4}\underbrace{\left(Var(\vartheta_i | \vartheta_i < \bar{h})\right)}_{①} + \underbrace{Var(\eta_i)}_{②}. \end{aligned}$$

To calculate ① $Var(\vartheta_i | \vartheta_i < \bar{h}) = \sigma_\vartheta^2 (Var(\frac{\vartheta_i}{\sigma_\vartheta} | \frac{\vartheta_i}{\sigma_\vartheta} < \bar{h}^*)) = \sigma_\vartheta^2 (1 - \frac{\bar{h}^* \phi(\bar{h}^*)}{\Phi(\bar{h}^*)} - (\frac{\phi(\bar{h}^*)}{\Phi(\bar{h}^*)})^2)$, due to the property of truncated normal distribution.

To calculate ② $Var(\eta_i) = Var(\varepsilon_i - \frac{\sigma_{\varepsilon v}}{\sigma_\vartheta^2}\vartheta_i) = \sigma_\varepsilon^2 - \frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}$.

Therefore, plugging ① and ②, one obtains:

$$Var(H_{i,1}|H_{i,0},C_i,X_i,P_i=1)=\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}\left(1-\frac{\bar{h}^*\phi(\bar{h}^*)}{\Phi(\bar{h}^*)}-\left(\frac{\phi(\bar{h}^*)}{\Phi(\bar{h}^*)}\right)^2\right)+\sigma_\varepsilon^2-\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}.$$

Similarly, for those who choose not to participate:

$$Var(H_{i,1}|H_{i,0},C_i,X_i,P_i=0)=\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}\left(1+\frac{\bar{h}^*\phi(\bar{h}^*)}{1-\Phi(\bar{h}^*)}-\left(\frac{\phi(\bar{h}^*)}{\Phi(\bar{h}^*)}\right)^2\right)+\sigma_\varepsilon^2-\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}.$$

I then need to obtain the unconditional variances to map to the data moments from the above conditional variances, since each individual has different $H_{i,0}$. Following $Var(Y)=E(Var(Y|X))+Var(E(Y|X))$, one derives:

$$Var(H_{i,1}|P_i=1)=\underbrace{E(Var(H_{i,1}|H_{i,0},C_i,X_i,P_i=1))}_{③}+\underbrace{Var(E(H_{i,1}|H_{i,0},C_i,X_i,P_i=1))}_{④}.$$

$$Var(H_{i,1}|P_i=0)=\underbrace{E(Var(H_{i,1}|H_{i,0},C_i,X_i,P_i=0))}_{⑤}+\underbrace{Var(E(H_{i,1}|H_{i,0},C_i,X_i,P_i=0))}_{⑥}.$$

The sample counterpart of $Var(H_{i,1}|P_i=1)$ is $\frac{1}{(N_1-1)}\sum_{i=1}^{N_1}(H_{i,1}^1-\overline{H_{i,1}^1})^2$, where N_1 is the total number of participants. The sample counterpart of ③ is $\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}\frac{1}{N_1}\sum_{i=1}^{N_1}f_1(\bar{h_i}^*)+\sigma_\varepsilon^2-\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}$, where $f_1(\bar{h_i}^*):=1-\frac{\bar{h_i}^*\phi(\bar{h_i}^*)}{\Phi(\bar{h_i}^*)}-(\frac{\phi(\bar{h_i}^*)}{\Phi(\bar{h_i}^*)})^2$. The sample counterpart of ④ is $\frac{1}{(N_1-1)}\sum_{i=1}^{N_1}(\widehat{H_{i,1}^1}-\overline{H_{i,1}^1})^2$, where $\widehat{H_{i,1}^1}$ is the predicted empathy obtained from estimating Eq (12).

Similarly, the sample counterpart of $Var(H_{i,1}|P_i=0)$ is $\frac{1}{(N_0-1)}\sum_{i=1}^{N_0}(H_{i,1}^0-\overline{H_{i,1}^0})^2$, where N_0 is the total number of individuals in the treatment group who did not participate. The sample counterpart of ⑤ is $\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}\frac{1}{N_0}\sum_{i=1}^{N_0}f_0(\bar{h_i}^*)+\sigma_\varepsilon^2-\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}$, where $f_0(\bar{h_i}^*):=1+\frac{\bar{h_i}^*\phi(\bar{h_i}^*)}{1-\Phi(\bar{h_i}^*)}-(\frac{\phi(\bar{h_i}^*)}{\Phi(\bar{h_i}^*)})^2$. The sample counterpart of ⑥ is $\frac{1}{(N_0-1)}\sum_{i=1}^{N_0}(\widehat{H_{i,1}^0}-\overline{H_{i,1}^0})^2$, where $\widehat{H_{i,1}^0}$ is the predicted empathy obtained from estimating Eq (11).

Therefore, I can estimate the variance of both shocks $\{\sigma_\varepsilon^2,\sigma_\vartheta^2\}$ using a nonlinear least square estimator, where the objective function is:

$$\arg\min_{\{\sigma_\varepsilon^2,\sigma_\vartheta^2\}}\frac{1}{N}\sum_{i=1}^{N}\Bigg[(H_{i,1}^1-\overline{H_{i,1}^1})^2P_i+(H_{i,1}^0-\overline{H_{i,1}^0})^2(1-P_i)-\left(\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}(f_1(\bar{h_i}^*)P_i+f_0(\bar{h_i}^*)(1-P_i))+\sigma_\varepsilon^2-\frac{\sigma_{\varepsilon v}^2}{\sigma_\vartheta^2}+(\widehat{H_{i,1}^1}-\overline{H_{i,1}^1})^2P_i+(\widehat{H_{i,1}^0}-\overline{H_{i,1}^0})^2(1-P_i)\right)\Bigg]^2$$

G Derivation of Equation (16)

I show the steps to obtain Eq (16) as follows:

$$\begin{aligned}\mathbf{b}^*_{\text{post}} &= [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}\mathbf{d}_{\text{post}} \\ &= [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}\mathbf{d}_{\text{pre}} + [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}\Delta\mathbf{d}\end{aligned}$$

As $\mathbf{b}^*_{\text{pre}} = [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{pre}})]^{-1}\mathbf{d}_{\text{pre}}$, we can get $\mathbf{d}_{\text{pre}} = [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{pre}})]\mathbf{b}^*_{\text{pre}}$. After plugging in, we get $\mathbf{b}^*_{\text{post}} = [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}[\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{pre}})]\mathbf{b}^*_{\text{pre}} + [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}\Delta\mathbf{d}$.

Therefore,

$$\begin{aligned}\Delta\mathbf{b} &= \mathbf{b}^*_{\text{post}} - \mathbf{b}^*_{\text{pre}} \\ &= [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}[\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{pre}})]\mathbf{b}^*_{\text{pre}} + [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}\Delta\mathbf{d} - \mathbf{b}^*_{\text{pre}} \\ &= [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}[\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{pre}})]\mathbf{b}^*_{\text{pre}} + [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}\beta\Delta H_i - \mathbf{b}^*_{\text{pre}} \\ &= [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}\left(\left[\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{pre}}) - (\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}}))\right]\mathbf{b}^*_{\text{pre}} + \beta\Delta H_i\right) \\ &= [\mathbf{I}+\phi(\mathbf{I}-\mathbf{W}_{\text{post}})]^{-1}(\beta\Delta H_i + \phi(\mathbf{W}_{\text{post}} - \mathbf{W}_{\text{pre}})\mathbf{b}^*_{\text{pre}}).\end{aligned}$$

H Key Player Exercise

In addition to comparing targeting by popularity with random assignment, I empirically investigate the "key player" theorem, following the methodologies outlined in Zenou (2016) and Lee et al. (2021). Instead of identifying key players by removing individuals, the applied approach involves the selection of optimal candidates for intervention using a greedy algorithm, resulting in a rank order for each student. This analysis simulates a potential policy query: "When confronted with limited resources and the imperative to select only ten students from each class for program participation, which ten students should be chosen?" Here, I use ten as an illustrative example.

Similar to Lee et al. (2021), I also report the Spearman's rank correlation coefficients in Appendix Table I16. Notably, the identification of key players strongly correlates with their network positions, especially in-degree. When considered alongside individual characteristics, it becomes clear that students with higher in-degree, lower empathy, and lower test scores should be prioritized targeting when faced with constraints. However, it is important to note that the level of bullying involvement is not closely related to the targeting order, indicating that these prioritized students may not necessarily exhibit the highest levels of bullying involvement.

I Appendix Tables and Figures

Table I1: Findings in Previous Analysis

	(1) Control mean	(2) ITT	(3) WCB p-value
	Panel A. Parents' outcomes		
Time investment (weekday)	3.725	0.513**	0.020
	(3.288)	(0.204)	
Time investment (weekend)	5.413	0.417*	0.103
	(3.649)	(0.230)	
Empathy index	0	0.100**	0.040
	(1)	(0.046)	
Authoritative parenting	0.789	0.039**	0.033
	(0.408)	(0.017)	
N	848	1,852	
	Panel B. Non-cognitive skills		
Empathy index	0	0.154**	0.050
	(1)	(0.073)	
Positive trait index	0	0.157***	0.007
	(1)	(0.056)	
Mental health index	0	0.164**	0.038
	(1)	(0.076)	
N	1,029	2,246	
	Panel C. Bullying		
Bullying score	0	-0.104*	0.108
	(1)	(0.059)	
Bully	0.176	-0.040*	0.107
	(0.381)	(0.024)	
Victim	0.421	-0.052*	0.118
	(0.494)	(0.031)	
Witnesses	0.420	-0.061*	0.092
	(0.494)	(0.034)	
Willing to help victims	0.767	0.053**	0.014
	(0.423)	(0.021)	
N	1,029	2,246	

Note. This table shows results in Cunha et al. (2023). Regression follows ANCOVA specification, similar to Eq (1). Panel A reports the intention-to-treat (ITT) estimates on parents' time investments and parental skills (N=1,852), while Panel B reports the ITT estimates on students' non-cognitive skills (N=2,246), and Panel C reports the ITT estimates on bullying-related outcomes (N=2,246). I also report the wild cluster bootstrapped (wcb) p-values after 9,999 resampling. Standard errors are clustered at the classroom level and presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table I2: Attrition in Network Information

	(1) Control	(2) Treatment	(3) Difference	(4) Total number
Number of classes	22	26	4	48
Number of students who complete survey	1,029	1,217	188	2,246
Number of students with valid network information	1,025	1,206	181	2,231
Number of attrition	4	11	7	15
Attrition rate	0.004	0.009	0.005	

Note. This table shows the attrition pattern once accounted for network information. Column (1) reports the statistics for control classes. Column (2) reports the statistics for treatment classes. Column (3) is the difference between the first two columns. Column (4) is the sum of the first two columns.

Table I3: Pattern of Friends' Structure by Bully Status

	Bully (1)	Non-bully (2)	Total (3)
	Panel A. Baseline		
Bully	0.706	2.175	2.882
Non-bully	0.567	2.468	3.035
	p = 0.001	p = 0.001	p=0.091
	Panel B. Follow-up		
Bully	0.637	1.944	2.581
Non-bully	0.389	2.681	3.070
	p = 0.000	p = 0.000	p=0.000

Note. This table shows the distribution of the number of friends by bully status at baseline (top panel) and follow-up (bottom panel). I define a student being a bully as engaging in more than once for each of the five events. I also report the p-value for the t-test comparing the differences between the two numbers in the same column.

Table I4: Empathy and Social Status

	(1) In-degree	(2) Isolation	(3) Eigenvector centrality	(4) Reciprocal link	(5) Unilateral link
empathy index	0.361***	-0.037***	0.021**	0.175***	0.123*
	(0.075)	(0.011)	(0.009)	(0.050)	(0.066)
$\mathbb{1}$(male)	-0.050	0.021	-0.030	-0.221**	0.058
	(0.101)	(0.016)	(0.033)	(0.093)	(0.070)
age	0.153*	0.003	0.019**	0.071	-0.013
	(0.082)	(0.014)	(0.008)	(0.047)	(0.055)
$\mathbb{1}$(urban hukou)	0.114	0.010	0.003	-0.001	-0.003
	(0.090)	(0.015)	(0.012)	(0.060)	(0.056)
$\mathbb{1}$(only child)	0.122	0.013	0.029**	0.024	-0.032
	(0.112)	(0.013)	(0.014)	(0.062)	(0.078)
class size	0.064***	-0.005*	-0.003	0.054**	0.009
	(0.022)	(0.003)	(0.002)	(0.021)	(0.011)
share of male	-0.512	0.110	-0.187	-1.567	0.985
	(2.063)	(0.177)	(0.217)	(2.614)	(1.661)
time spent on study	0.016	-0.001	0.003	0.012	0.007
	(0.015)	(0.002)	(0.002)	(0.010)	(0.012)
time spent on leisure	-0.038***	0.001	-0.005***	-0.022**	0.011
	(0.013)	(0.002)	(0.002)	(0.009)	(0.009)
positive personality index	0.064	0.008	0.012	0.049	0.041
	(0.056)	(0.010)	(0.007)	(0.039)	(0.040)
stress index	0.275***	-0.011	0.030***	0.144***	-0.024
	(0.073)	(0.010)	(0.009)	(0.047)	(0.052)
Strata FE	Y	Y	Y	Y	Y
R^2	0.045	0.013	0.035	0.061	0.019
N	2,231	2,231	2,231	2,231	2,231

Note. This table shows the association between empathy and students' social status. I quantify the social status using several network statistics, including *in-degree*, *isolation*, *eigenvector centrality*, *reciprocal link*, and *unilateral link*. The network statistics are defined in Appendix Section A.1. The numbers are the regression results of the network statistics over the variables listed in the first column using the baseline data. For each regression, I also control for strata fixed effects. Standard errors are clustered at the classroom level and presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table I5: Who Makes Friends with Who? – Alternative Bully Definition

	(1) Control	(2) Intention to treat	(3) WCB p-value	(4) N
Bully once (baseline)				
# peers are bullies	0.974	-0.033	0.733	849
	(1.038)	(0.093)		
# peers are non-bullies	1.833	0.262**	0.040	849
	(1.422)	(0.127)		
Never bully (baseline)				
# peers are bullies	0.943	-0.061	0.509	1,382
	(1.004)	(0.090)		
# peers are non-bullies	2.075	0.221**	0.030	1,382
	(1.453)	(0.097)		
Homophily				
Bully homophily	-0.023	-0.004	0.945	48
	(0.177)	(0.060)		
Non-bully homophily	0.054	0.040	0.180	48
	(0.105)	(0.031)		

Note. This table shows the intention-to-treat (ITT) estimates for the subgroup analyses of the program's impact on network structure. For this set of results, I use an alternative definition of being a bully (at least once). I report the ITT estimates on whether friends at the follow-up are bullies and non-bullies for baseline bullies and non-bullies separately. At the classroom level, I also report the ITT estimates on the bully and non-bully homophily indices. I also report the wild cluster bootstrapped (wcb) p-values after 9,999 resampling. Standard errors are clustered at the classroom level and presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table I6: Bullies and Non-Bullies in Friend Network

	(1) Control	(2) Intention to treat	(3) WCB p-value	(4) N
Panel A. Bullies				
# friends are bullies	0.457	-0.067	0.390	2,231
	(0.738)	(0.076)		
# old friends are bullies	0.267	-0.044	0.553	2,231
	(0.553)	(0.050)		
# new friends are bullies	0.189	-0.023	0.473	2,231
	(0.473)	(0.036)		
Panel B. Non-bullies				
# friends are non-bullies	2.482	0.275**	0.033	2,231
	(1.649)	(0.119)		
# old friends are non-bullies	1.397	0.174	0.133	2,231
	(1.355)	(0.106)		
# new friends are non-bullies	1.085	0.101	0.292	2,231
	(1.221)	(0.093)		

Note. This table shows the intention-to-treat (ITT) estimates on various measures of the number of friends who are bullies/non-bullies at the follow-up. I define bullies after accounting for the repetition of each of the five events. Panel A reports results for the number of bully friends for the whole sample at the follow-up, while Panel B reports results for the number of non-bully friends. I further divide these friends based on whether they are new or old friends compared to the baseline friendship links. I also report the wild cluster bootstrapped (wcb) p-values after 9,999 resampling. Standard errors are clustered at the classroom level and presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table I7: Who Makes Friends with Who? A Subgroup Analysis on Victims

	(1) Control	(2) Intention to treat	(3) WCB p-value	(4) N
Victim (baseline)				
# peers are victims	1.153	-0.091	0.530	1,280
	(1.105)	(0.132)		
# peers are non-victims	1.656	0.269**	0.041	1,280
	(1.399)	(0.121)		
Non-victim (baseline)				
# peers are victims	1.082	-0.077	0.474	951
	(1.050)	(0.102)		
# peers are non-victims	2.030	0.264**	0.044	951
	(1.466)	(0.123)		
Homophily				
Victim homophily	0.018	0.005	0.876	48
	(0.170)	(0.034)		
Non-victim homophily	0.041	0.012	0.784	48
	(0.178)	(0.042)		

Note. This table shows the intention-to-treat (ITT) estimates for the subgroup analyses of the program's impact on network structure for victims. I define victims after accounting for the repetition of each of the five events. I report the ITT estimates on whether friends at the follow-up are victims and non-victims for baseline victims and non-victims separately. At the classroom level, I also report the ITT estimates on the victim and non-victim homophily index. I also report the wild cluster bootstrapped (wcb) p-values after 9,999 resampling. Standard errors are clustered at the classroom level and presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table I8: Victims and Non-Victims in Friend Network

	(1) Control	(2) Intention to treat	(3) WCB p-value	(4) N
Panel A. Victims				
# friends are victims	1.123	-0.081	0.494	2,231
	(1.082)	(0.110)		
# old friends are victims	0.671	-0.055	0.527	2,231
	(0.850)	(0.080)		
# new friends are victims	0.452	-0.026	0.622	2,231
	(0.743)	(0.051)		
Panel B. Non-victims				
# friends are non-victims	1.816	0.289**	0.038	2,231
	(1.439)	(0.129)		
# old friends are non-victims	0.993	0.185*	0.075	2,231
	(1.155)	(0.098)		
# new friends are non-victims	0.822	0.105	0.231	2,231
	(1.053)	(0.082)		

Note. This table shows the intention-to-treat (ITT) estimates on various measures of the number of friends who are victims/non-victims. I define victims after accounting for the repetition of each of the five events. Panels A and B report results for the number of victim friends and non-victim friends, respectively. I further divide these friends based on whether they are new or old friends compared to the baseline friendship links. I also report the wild cluster bootstrapped (wcb) p-values after 9,999 resampling. Standard errors are clustered at the classroom level and presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table I9: Program Effects on Homophily Structure

		(1) Treat	(2) S.E	(3) N
Gender	Overall	0.031	(0.032)	1,803
	Male	0.053	(0.038)	926
	Female	0.004	(0.038)	877
Academic skill	Overall	1.662	(1.150)	1,794
	High-achiever	0.154	(0.245)	492
	Low-achiever	2.308	(1.571)	1,302
Hukou status	Overall	-0.028	(0.023)	1,803
	Urban hukou	-0.006	(0.030)	836
	Rural hukou	-0.047	(0.035)	967
College aspiration	Overall	0.089	(0.071)	1,803
	High college aspiration	0.126	(0.075)	1,550
	Low college aspiration	-0.191	(0.146)	253
Self-esteem	Overall	-0.007	(0.042)	1,766
	High self-esteem	-0.069	(0.065)	718
	Low self-esteem	0.037	(0.049)	1,048
Perseverance	Overall	0.052	(0.042)	1,766
	High perseverance	0.049	(0.053)	727
	Low perseverance	0.056	(0.052)	1,039
Self-confidence	Overall	0.009	(0.037)	1,721
	High self-confidence	-0.070	(0.051)	678
	Low self-confidence	0.059	(0.044)	1,043
Empathy	Overall	0.021	(0.032)	1,766
	High empathy	-0.004	(0.041)	899
	Low empathy	0.043	(0.046)	867

Note. This table shows the program's intention-to-treat (ITT) effects on the homophily structure of students' friendship networks. The outcome variable for each column is the homophily index for the corresponding characteristic constructed following Currarini et al. (2009). Low/High is relative to the class average. The number of observations is distinct from 2,231 due to the existence of students nominating zero friends or missing the key variables. For each regression, I also control for baseline outcome measures and strata fixed effects. Classroom-level clustered standard errors are presented in parentheses (* $p<0.10$, ** $p<0.05$, *** $p<0.01$).

Table I10: Empirical Test of the Sequential Independence Assumption

	(1) In-degree	(2) Isolation	(3) Eigenvector centrality	(4) Reciprocal link	(5) Unilateral link
Panel A. Empathy (follow-up)					
empathy index	0.093*	-0.019**	0.012*	0.095**	0.156***
	(0.049)	(0.007)	(0.007)	(0.039)	(0.040)
treat	0.158	-0.016	-0.025	0.019	0.142*
	(0.096)	(0.011)	(0.018)	(0.098)	(0.079)
Panel B. Empathy (baseline)					
empathy index	0.156**	-0.033***	0.008	0.137***	0.028
	(0.063)	(0.010)	(0.008)	(0.040)	(0.048)
treat	0.181*	-0.019*	-0.022	0.037	0.159*
	(0.101)	(0.011)	(0.017)	(0.100)	(0.079)
Baseline network statistics	Y	Y	Y	Y	Y
Follow-up noncognitive skills	Y	Y	Y	Y	Y
Individual characteristics	Y	Y	Y	Y	Y
N	2,231	2,231	2,231	2,231	2,231

Note. This table shows the results of empirically testing the sequential independence assumption. In Panel A, I show the results when regressing the corresponding follow-up network statistics over the follow-up empathy index and the treatment indicator. In Panel B, I show the results when regressing the corresponding follow-up network statistics over the baseline empathy index and the treatment indicator. For all regressions, I control for the corresponding baseline network statistics. I also control for the same set of individual characteristics as in Appendix Table I4 and the follow-up noncognitive skills, including the positive personality index and the stress index. I omit the estimated coefficients as those are all insignificant. Standard errors are clustered at the classroom level and presented in parentheses (* $p<0.10$, ** $p<0.05$, *** $p<0.01$).

Table I11: Probability of Forming a Friendship Link at Follow-up (no baseline network)

	(1) Value	(2) Std.Err
$\mathbb{1}\left[x_i = x_j\right]$		
gender	0.148***	(0.002)
hukou	0.011***	(0.002)
only child	0.005**	(0.002)
bully at baseline	0.004**	(0.002)
victim at baseline	0.005***	(0.002)
$\left\|x_i - x_j\right\|$		
age	-0.010***	(0.002)
empathy index	-0.004***	(0.001)
height	-0.023***	(0.001)
time spent on studying	-0.003**	(0.001)
time spent on leisure	-0.007***	(0.001)
pocket money	-0.006***	(0.001)
test score	-0.010***	(0.001)
x_j		
male	-0.006***	(0.002)
only child	0.002	(0.002)
empathy index	0.004***	(0.001)
test score rank	-0.003***	(0.001)
R^2	0.100	
N	102,054	
# students	2,231	

Note. This table shows the estimation results of Eq (15) when not conditional on the baseline network w_{0ij}. I model the directed links. The dependent variable w_{1ij} takes on value 1 if i nominated j (and zero otherwise, even in the event that j nominated i, i.e., $w_{1ji} = 1$). I estimate fixed-effect logit regressions. I include agent i (i.e., sender) fixed effects. It is worth noting that only the empathy index is obtained from the follow-up measure, while all other variables are measured at the baseline. The number of observations (i.e., potential links) in all regression is 102,054, which stems from a sample of 2,231 unique observations. Standard errors are two-way clustered by nominating and nominated students and presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table I12: Robustness Checks of Network Formation and Bullying Involvement Estimation
Censored Network

	Model		Truncated 5th friend		Truncated 4th friend		Truncated 3rd friend	
	Value	Std. Err.	Value	Std.Err.	Value	Std.Err.	Value	Std.Err.
Panel A. Link formation								
friends at baseline	0.463***	(0.016)	0.455***	(0.015)	0.441***	(0.015)	0.421***	(0.014)
$\mathbb{1}[x_i = x_j]$								
bully at baseline	0.002	(0.001)	0.002*	(0.001)	0.002	(0.001)	0.003**	(0.001)
victim at baseline	0.003	(0.002)	0.002	(0.002)	0.003	(0.002)	0.002	(0.002)
$\lvert x_i - x_j \rvert$								
age	-0.007**	(0.003)	-0.006*	(0.003)	-0.006**	(0.003)	-0.007**	(0.003)
empathy index	-0.001	(0.001)	-0.001	(0.001)	-0.001	(0.001)	-0.001	(0.001)
height	-0.002***	(0.000)	-0.002***	(0.000)	-0.002***	(0.000)	-0.002***	(0.000)
time spent on studying	-0.001	(0.001)	-0.001	(0.001)	-0.001	(0.001)	-0.001	(0.001)
time spent on leisure	-0.002*	(0.001)	-0.002**	(0.001)	-0.003**	(0.001)	-0.003***	(0.001)
pocket money	-0.002	(0.001)	-0.002*	(0.001)	-0.002**	(0.001)	-0.002**	(0.001)
test score	-0.001	(0.002)	-0.000	(0.002)	-0.001	(0.002)	-0.002	(0.002)
x_j								
male	-0.002	(0.002)	-0.001	(0.001)	-0.001	(0.001)	-0.000	(0.002)
only child	-0.001	(0.002)	-0.001	(0.002)	-0.001	(0.001)	0.001	(0.002)
empathy index	0.003***	(0.001)	0.003***	(0.001)	0.002***	(0.001)	0.003***	(0.001)
test score rank	-0.001	(0.001)	-0.002**	(0.001)	-0.001	(0.001)	-0.001	(0.001)
R^2	0.241		0.230		0.217		0.201	
N	102,054							
# students	2,231							
Panel B. Peer effects of bullying								
peer bullying score	0.484***	(0.129)	0.477***	(0.128)	0.471***	(0.135)	0.464***	(0.136)
empathy index	-0.139***	(0.024)	-0.139***	(0.024)	-0.139***	(0.024)	-0.138***	(0.024)
male	0.161***	(0.040)	0.163***	(0.039)	0.164***	(0.040)	0.166***	(0.042)
urban hukou	0.082**	(0.038)	0.085**	(0.039)	0.081**	(0.038)	0.083**	(0.038)
bully at baseline	0.562***	(0.074)	0.563***	(0.074)	0.559***	(0.074)	0.562***	(0.074)
victim at baseline	0.162***	(0.042)	0.162***	(0.042)	0.167***	(0.041)	0.162***	(0.042)
F-statistic	20.534		21.190		19.579		19.724	
N	2,231							

Note. This table shows the robustness checks for the estimation results of network formation and bullying involvement to address the censored network concern. The first two columns display the main model estimates from Table 6. In the subsequent columns, I report estimation results after truncating friends at specific rank-order positions. Students were asked to rank their friends by closeness, with the 1st friend being the best friend and the 5th friend being the least close. Bootstrapped standard errors are presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table I13: Robustness Checks of Bullying Involvement Estimation Comparing with Other Estimates

	Bullying score		
	(1)	(2)	(3)
	OLS	2SLS	2SLS (model)
empathy index	-0.150***	-0.142***	-0.139***
	(0.024)	(0.024)	(0.024)
peer bullying score	0.154**	0.322***	0.484***
	(0.072)	(0.116)	(0.129)
Instruments		$W_1{}^2X, W_1{}^3X$	$\widehat{W_1}^2X, \widehat{W_1}^3X$
Cragg-Donald Wald F-statistic		31.065	20.534
Strata FE	Y	Y	Y
N	2,231	2,231	2,231

Note. This table shows the estimation results of an OLS estimation, 2SLS estimation following Bramoullé et al. (2009), and the 2SLS estimation (used in this study). The 2SLS in Column (2) differs from the 2SLS method used in this study reported in Column (3) in that the former does not account for friendship network endogeneity. I include the same control variables for all three specifications as in the main model. Bootstrapped standard errors are presented in parentheses (* p<0.10, ** p<0.05, *** p<0.01).

Table I14: Within-Sample Model Fit

	Model	Data	
	Mean of simulated means	Mean	SE of mean
Proportion of take-up	0.402	0.400	0.014
Follow-up empathy (SD)	0.072	0.076	0.022
In-degree	3.354	3.101	0.048
Change in bully score (SD)	-0.015	-0.027	0.023

Note. This table shows the within-sample model fit. Simulations are based upon 1,000 repetitions.

Table I15: Effects on Bullying Reduction and Empathy for Counterfactual Experiments

	Everyone	Top10% most pop	Bully	Bully's best friends	Bully & best friends	Popular bully	Nonpop bully
	Panel A. Offer treatment						
Unit treated (N)	2,231	480	456	394	761	200	256
Unit take-up (N)	1,044	243	173	180	314	84	89
Effect on bully score (SD)	-0.165	-0.039	-0.024	-0.030	-0.049	-0.012	-0.013
	(0.0023)	(0.0022)	(0.0026)	(0.0035)	(0.0029)	(0.0026)	(0.0025)
PP decrease in bullies	-8.08pp	-1.93pp	-1.19pp	-1.48pp	-2.41pp	-0.58pp	-0.61pp
Effect on empathy (SD)	0.148	0.033	0.027	0.026	0.047	0.012	0.015
	(0.0016)	(0.0015)	(0.0013)	(0.0014)	(0.0017)	(0.0010)	(0.0010)
	Panel B. Assume take-up						
Unit treated (N)	2,231	480	456	394	761	200	256
Unit take-up (N)	2,231	480	456	394	761	200	256
Effect on bully score (SD)	-0.351	-0.081	-0.058	-0.065	-0.110	-0.027	-0.031
	(0.0043)	(0.0042)	(0.0046)	(0.0045)	(0.0047)	(0.0044)	(0.0044)
PP decrease in bullies	-17.16pp	-3.96pp	-2.82pp	-3.17pp	-5.88pp	-1.33pp	-1.50pp
Effect on empathy (SD)	0.318	0.068	0.065	0.056	0.108	0.028	0.036
	(0.0000)	(0.0028)	(0.0027)	(0.0026)	(0.0032)	(0.0019)	(0.0021)

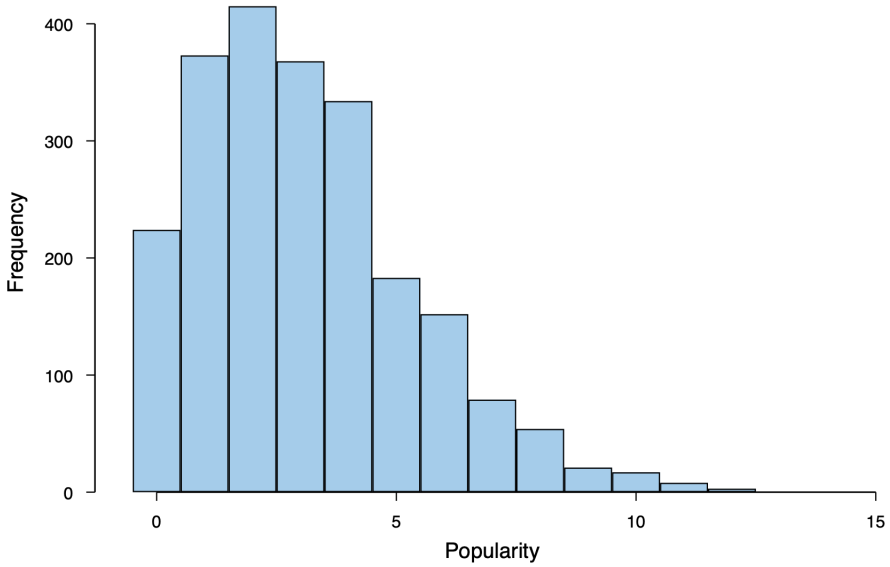
Note. This table reports the simulation results for the following counterfactual scenarios: (i) treating everyone, (ii) treating the top 10% most popular students, (iii) treating bullies, (iv) treating bullies' best friends, (v) treating bullies and best friends, (vi) treating popular bully, and (vii) treating non-popular bully. Popular or non-popular bullies are defined by comparing with the median value of the in-degree measure for bullies. In Panel A, I report the results of offering the treatment. In Panel B, I report the results of assigning take-up, i.e., assuming perfect compliance. I report the changes in bullying scores and in empathy, both measured by the standard deviation (SD). I also convert the bullying score to a percentage-point (pp) decrease in bullies following the evidence that our intervention reduces 4 pp bullies while lowering the bullying score by 0.079 SD in the estimation sample.

Table I16: Testing the Key Player Theorem – Spearman's Rank Correlation Coefficients

	Target order
Panel A. Network statistics	
in-degree	-0.486
eigenvector centrality	-0.030
closeness centrality	-0.067
isolation	-0.357
Panel B. Individual characteristics	
bully score	-0.103
male	0.352
hukou	-0.329
only child	-0.371
positive personality index	0.867
empathy index	0.988
standardized test score	0.648

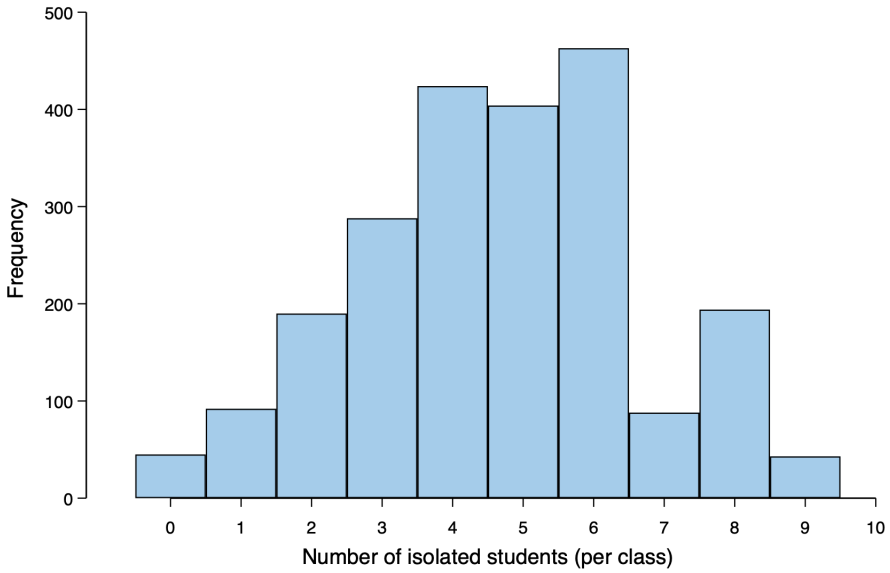
Note. This table reports Spearman's rank correlation coefficients between students' target order predicted by the greedy algorithm and their network statistics as well as individual characteristics at the baseline.

Figure I1: Distribution of the Popularity Measure



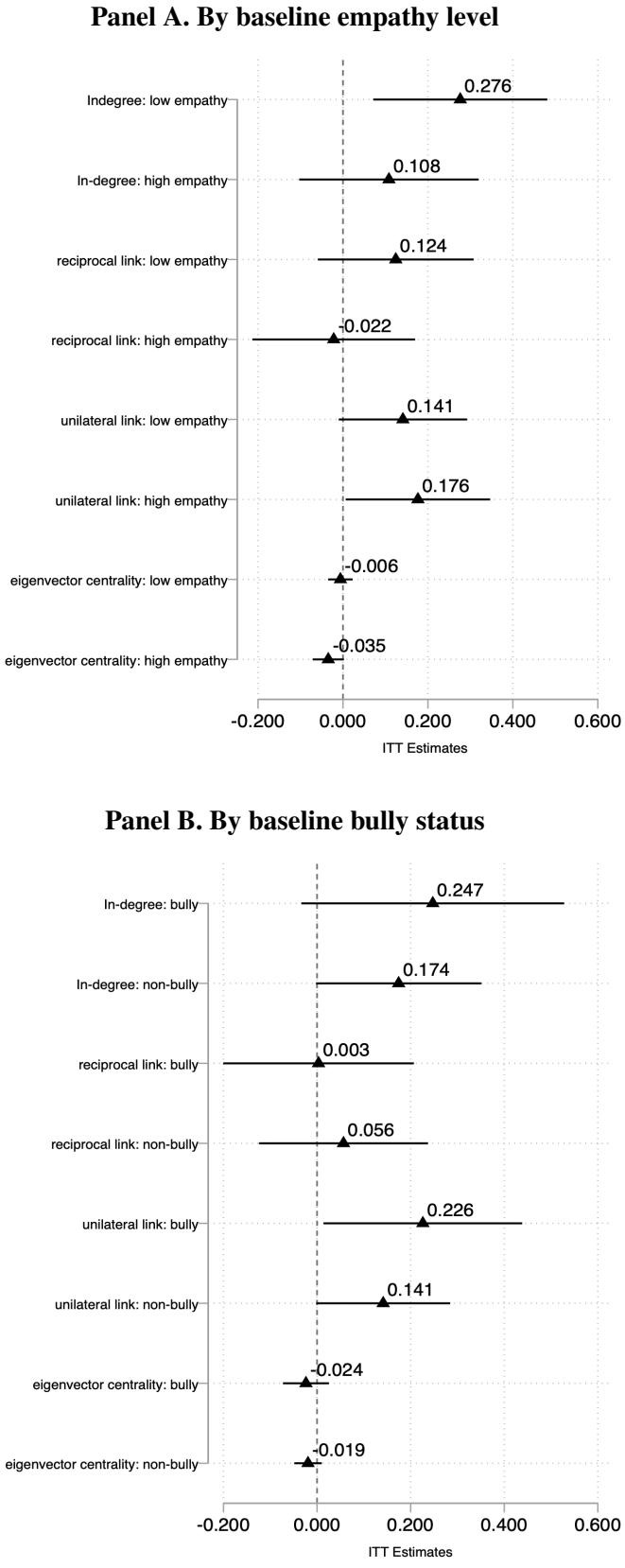
Note. The figure plots the distribution of the popularity measure at the baseline. The popularity measure is constructed by counting the total number of nomination links each student received.

Figure I2: Distribution of the Number of Isolated Students (With No Friend Nominations)



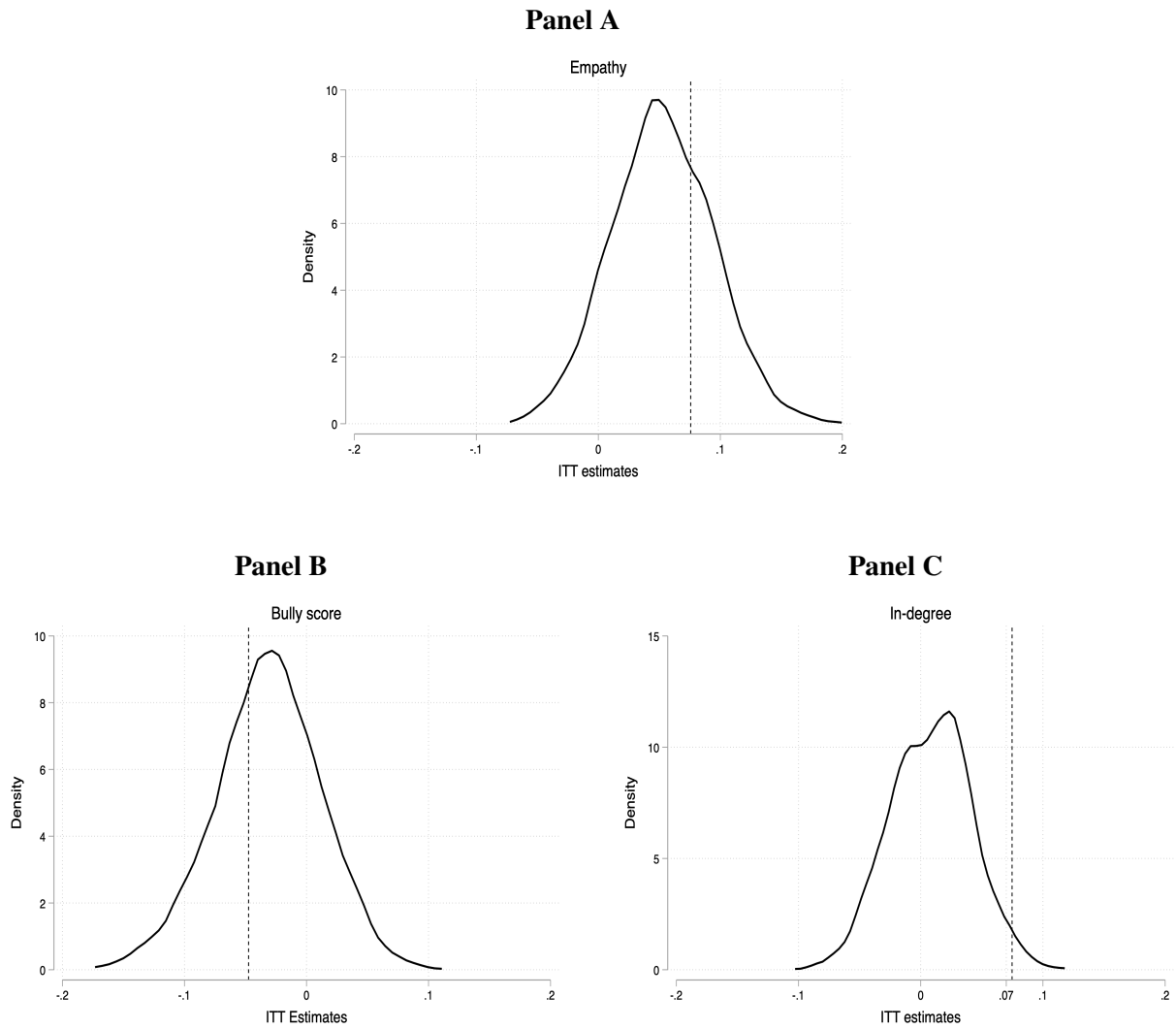
Note. The figure plots the distribution of the number of isolated students with no friend nominations within each class at the baseline.

Figure I3: Heterogeneity of the Program Effects on Network Statistics



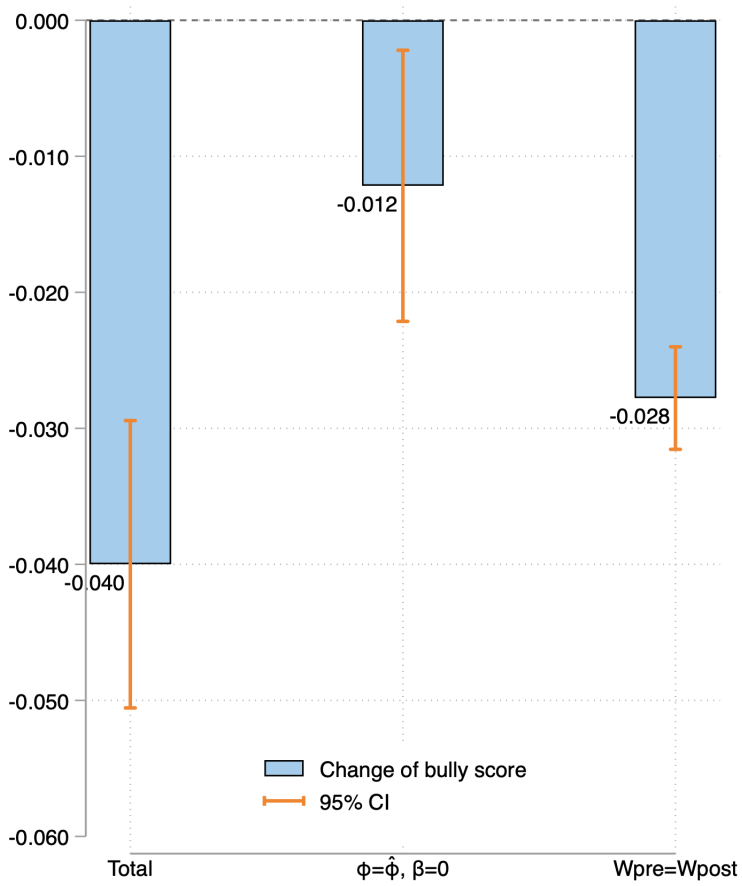
Note. The figures show the heterogeneity effects of program impacts on network statistics by baseline empathy level (Panel A) and by baseline bully status (Panel B). High/low empathy is defined according to the baseline measure of empathy and the classroom-level median empathy: When the individual has a higher empathy than the classroom median, he/she is defined as a high-empathy type and vice versa. I show the point estimates based on estimating Eq (1) as well as 90% confidence intervals. Confidence intervals are calculated based on bootstrapped standard errors clustered at the class level.

Figure I4: Model Validation



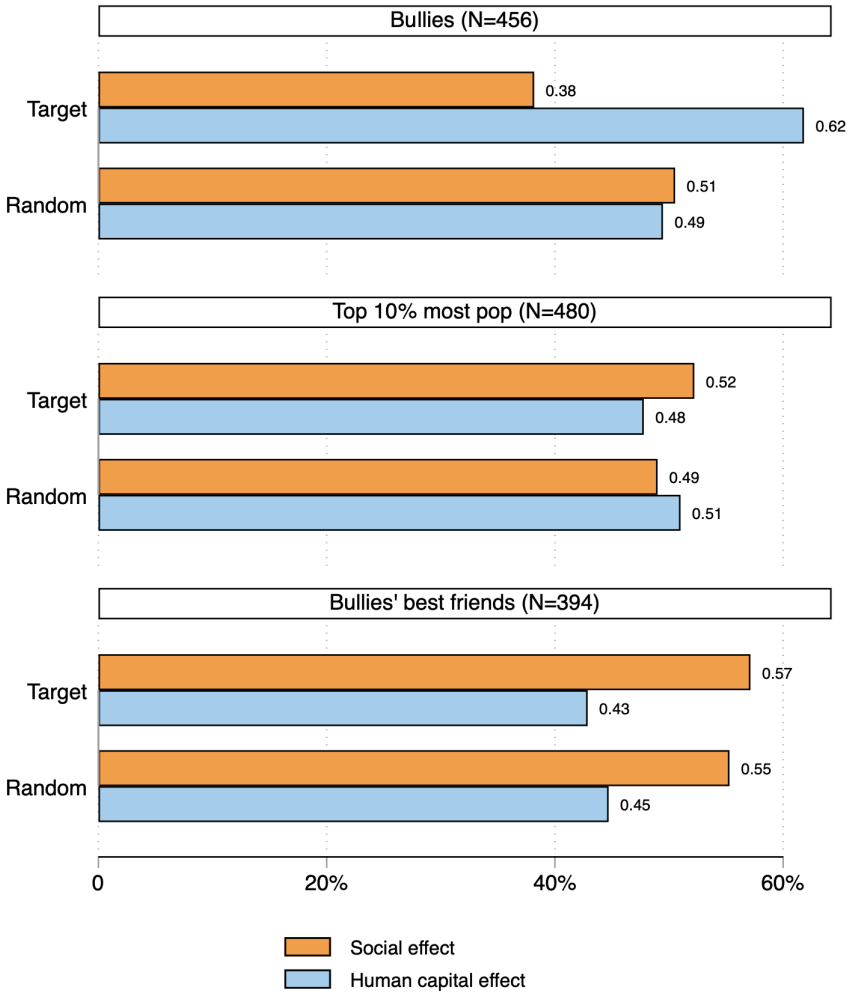
Note. The figures show the model validation results, comparing the predicted impacts under the model to the experimental impact. I present the results for empathy (Panel A), bully score (Panel B), and in-degree (Panel C). The black solid line represents the distribution of the predicted impacts obtained from 1,000 simulations of the 1,206 treatment units, while the black dashed line represents the estimated impacts from the data.

Figure I5: Additional Results of Decomposition Exercise



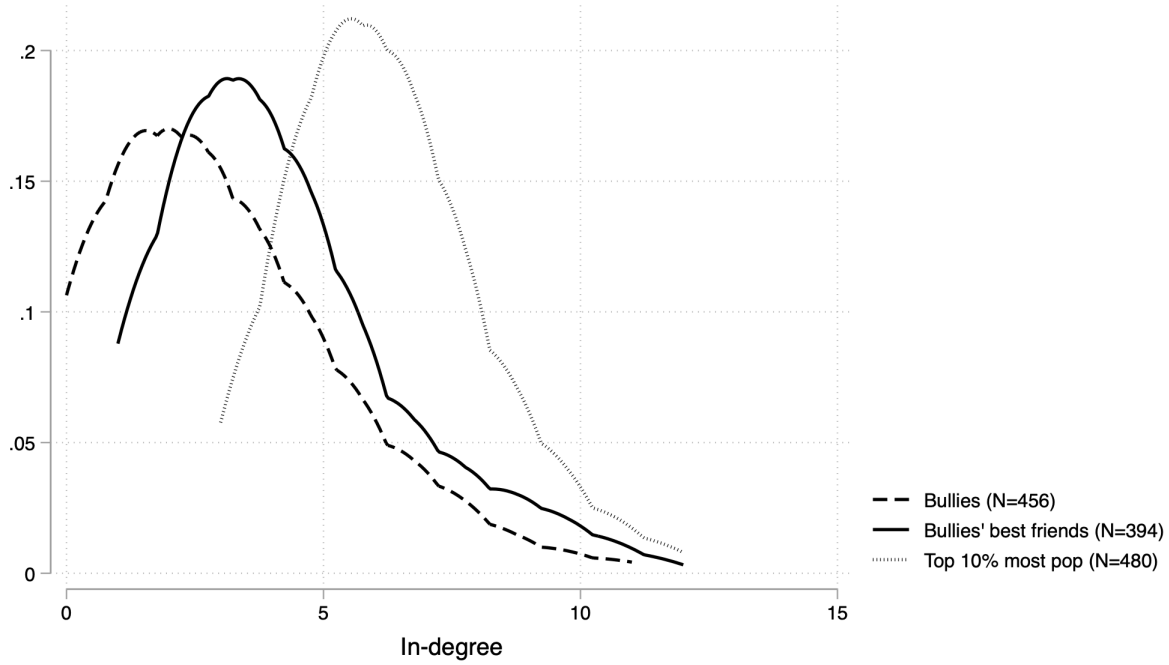
Note. The figure plots the additional results of the decomposition exercise based on Eq (16). The first bar reports the total change in the bullying score. The second bar reports the change in the bullying score when only keeping the estimated peer effect coefficient and setting the empathy effect equal to zero, which is an alternative method to calculate the social effect. The third bar reports the change in the bullying score when assuming the network links at the baseline and at the follow-up are the same. I also report the 95% confidence intervals for the changes using the orange solid line. The social effect obtained from the second bar accounts for 30% of the empathy effect on bullying, which is very close to the results shown in Figure 2.

Figure I6: Decomposition for Counterfactual Experiments



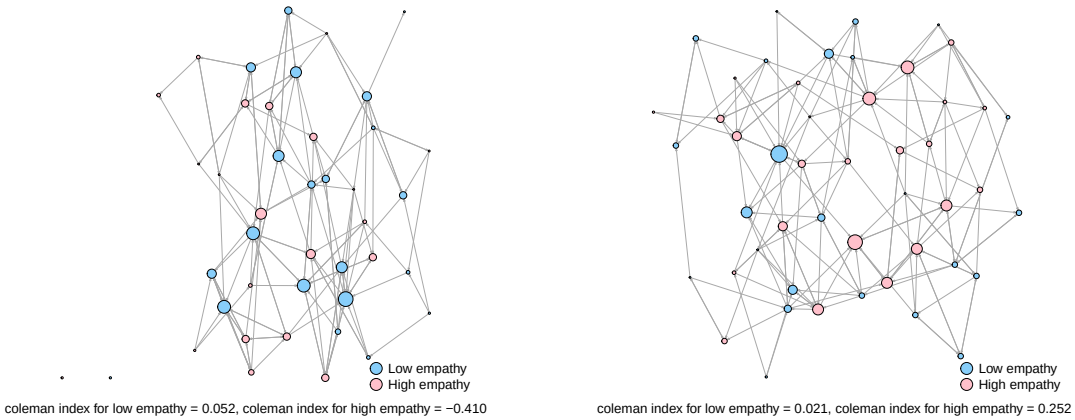
Note. The figure plots the human capital and social effects of empathy on bully reduction for the three counterfactual scenarios: (i) treating bullies, (ii) treating bullies' best friends, and (iii) treating the top 10% most popular students. For each scenario, I present decomposition results for both the targeted interventions and corresponding random assignments with the same number of treated units. The orange-filled box denotes the social effects, defined as the empathy-induced network changes and the associated peer bullying spillovers. The blue-filled box denotes the individual human capital effects, defined as the improvement in own empathy. To obtain the individual human capital effect, I set the peer effect parameter in Eq (9) equal to zero and simulated the new bullying score for the study sample for all three targeting experiments. The percentage is then obtained by calculating the proportion of the simulated change in bullying score over the total change in bully score. The remaining proportion is then defined as the contribution due to the social effect.

Figure I7: Density Plot of the In-degree for Bullies and Bullies' Best Friends



Note. The figure plots the kernel density distribution of the in-degrees for bullies (N = 456) and bullies' best friends (N = 394) at the baseline, respectively. I also plot the density distribution of in-degree for the top 10% most popular students (N = 480) for comparison purposes. In-degree is defined as the total number of nomination links the student received.

Figure I8: Two Examples of Classroom Networks (by Empathy)



Note. The two figures are visualizations of the friendship network of two classes. The blue color denotes students who are low empathetic, while the red color denotes students who are highly empathetic. High/low empathy is defined according to the baseline measure of empathy and the classroom-level median empathy: When the individual has a higher empathy than the classroom median, he/she is defined as a high-empathy type and vice versa. The size of the network node is scaled by in-degree, that is, the nomination links the student received. The figure on the left panel depicts a friendship network with higher segregation for low-empathy students than for high-empathy students. The figure on the right panel depicts a friendship network with higher segregation for high-empathy students compared to low-empathy students.

Appendix References

Alan, Sule, Ceren Baysan, Mert Gumren, and Elif Kubilay, "Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking," *Quarterly Journal of Economics*, 03 2021, *136* (4), 2147–2194. [1, 11]

_ , Elif Kubilay, Elif Bodur, and Ipek Mumcu, "Social Status in Student Networks and Implications for Perceived Social Climate in Schools," 2021. CESifo Working Paper.

_ , Enes Duysak, Elif Kubilay, and Ipek Mumcu, "Social Exclusion and Ethnic Segregation in Schools: The Role of Teacher's Ethnic Prejudice," *Review of Economics and Statistics*, 2021, pp. 1–45. [6]

Coleman, James, "Relational analysis: The study of social organizations with survey methods," *Human Organization*, 1958, *17* (4), 28–36. [11]

Cunha, Flavio, Qinyou Hu, Yiming Xia, and Naibao Zhao, "Reducing Bullying: Evidence from a Parental Involvement Program on Empathy Education," Working Paper 30827, National Bureau of Economic Research 2023. [2, 4, 8, 10, and 12]

Currarini, Sergio, Matthew O Jackson, and Paolo Pin, "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica*, 2009, *77* (4), 1003–1045. [7]

Dahlberg, Linda L, Susan B Toal, Monica Haavisto Swahn, and Christopher B Behrens, "Measuring violence-related attitudes, beliefs, behaviors, and influences among youths: a compendium of assessment tools," Technical Report, Atlanta, GA: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control 2005.

Lee, Lung-Fei, Xiaodong Liu, Eleonora Patacchini, and Yves Zenou, "Who is the Key Player? A Network Analysis of Juvenile Delinquency," *Journal of Business & Economic Statistics*, 2021, *39* (3), 849–857. [7, 19, 20, and 24]

Lin, Xu and Bruce A. Weinberg, "Unrequited friendship? How reciprocity mediates adolescent peer effects," *Regional Science and Urban Economics*, 2014, *48*, 144–153.

Lord, Frederic M, *Applications of item response theory to practical testing problems*, Routledge, 2012. [10, 17]

Vignemont, Frederique De and Tania Singer, "The empathic brain: how, when and why?," *Trends in Cognitive Sciences*, 2006, *10* (10), 435–441.

Yang, Wenhui, Ge Xiong, Luis Eduardo Garrido, John X Zhang, Meng-Cheng Wang, and Chong Wang, "Factor structure and criterion validity across the full scale and ten short forms of the CES-D among Chinese adolescents.," *Psychological Assessment*, 2018, *30* (9), 1186–1198.

Zenou, Yves, "Key players," *Oxford Handbook on the Economics of Networks*, 2016, pp. 244–274. [7, 30]