Reducing Bullying: Evidence from a Parental Involvement Program on Empathy Education[*]

Flavio Cunha Qinyou Hu Yiming Xia Naibao Zhao

June 11, 2022

THIS IS A PRELIMINARY DRAFT.
PLEASE DO NOT CITE OR CIRCULATE WITHOUT THE PERMISSION OF THE AUTHOR.

ABSTRACT:

This article examines a low-cost experiment with the aim of preventing bullying through a four-month parental involvement program on empathy education in Chinese middle schools. Participants included 2,246 7th and 8th graders and their parents. Parents in randomly selected classes were encouraged to finish biweekly family tasks on empathy and parenting with their children. With the monthly actual take-up rate being around 40 percent, the program led to a moderate reduction in school bullying incidents by about 5 percentage points. The program achieved such effectiveness through changes in parenting behaviors and improvement in students' empathy and other noncognitive abilities. Given that over one-third of teenagers worldwide are now involved in bullying incidents, the findings of this work have further implications for the causes of and solutions to school bullying among adolescents.

KEYWORDS: Parental involvement; Bullying; Empathy.
JEL CLASSIFICATION: I20; J13; J24; C93

[*] **Cunha**: Department of Economics, Rice University (e-mail: Flavio.Cunha@rice.edu); **Hu**: Department of Economics, Rice University (e-mail: qinyou.hu@rice.edu); **Xia**: RIEM, Southwestern University of Finance and Economics (e-mail: xiaym@swufe.edu.cn); **Zhao**: RIEM, Southwestern University of Finance and Economics (e-mail: nzhao@swufe.edu.cn). The trial has been registered at the AEA Registry: AEARCTR-0007079.

1 Introduction

School bullying among adolescents is a global challenge.[1] A recent report has estimated that one-third of the world's youth are victims (UNESCO, 2019). Mounting evidence indicates that bullying negatively affects adolescents' mental health, academic skills and abilities and even leads to suicide or suicidal thoughts (Sarzosa, 2021; Sarzosa and Urzúa, 2021; Hinduja and Patchin, 2010). Even for those societies where success in school is evaluated primarily by academic standards, bullying is a major factor hindering academic progress and development (Molcho et al., 2009). Furthermore, the adverse consequences of bullying can persist into adulthood (Copeland et al., 2013). Numerous instances worldwide have shown that daily bullying events can escalate into life-threatening violence and have tragic consequences. Although bullying has been a problem for generations and today's adolescents experience bullying in increasingly intrusive ways, it has garnered limited attention within policy circles, despite the rise in school safety as a top national priority in the U.S. (Federal Commission on School Safety, 2018). The lack of awareness is considerably worse in the developing world, where preventive policies have not been widely adopted. For example, it was not until June 2021 that China's Ministry of Education enacted the Law on Protection of Minors, which legally mandated reporting and protection of suspected child maltreatment or abuse at school (China's Ministry of Education, 2021).

The consequences of bullying are well-known. However, the causes of and solutions to the increase in school bullying remain underexplored. According to the developmental psychology literature, a lack of empathy, or the ability to detect others' emotions, can be a potential driver of school bullying incidents (Jolliffe and Farrington, 2006). Emphatic conception brings individuals closer and facilitates communication in almost every aspect of daily life. Although much research has been carried out on the role of empathy in promoting prosocial behavior and inhibiting antisocial behavior (Eisenberg and Miller, 1987), few studies have looked at the specific causal relationship between empathy and bullying.

To fill in the research gap and inform policy, we design a low-cost, highly scalable, and parent-directed intervention on empathy education aimed at preventing bullying. We design the intervention at the family level, involving children and their parents, motivated by research on parents' central role in children's socioemotional development (Cunha and Heckman, 2008; Cunha et al., 2010; Bono et al., 2016; Del Boca et al., 2017; Attanasio et al., 2020). Our program also borrows insights from the psychology literature that high levels of depression and low levels of empathy may be associated with bullying behaviors (Cook et al., 2010; Jolliffe and Farrington, 2006). The curriculum includes the introduction of concepts and coaching on noncognitive skills and positive parenting skills, The detailed content consists of 8 biweekly parent-child reading tasks and 4 empathy-oriented movies on 4 monthly themes. The intervention urges parents and students to receive education on empathy and encourages them to incorporate empathy and the value of chil-

[1] The research on "bullying" dates back to Olweus (1978), who defined bullying as an aggressive behavior with the feature of repetition and an imbalance of power between the two sides involved.

dren's uniqueness into their daily parenting practices. Parents and students gain access to these materials via a mobile app that also records their progress.[2]

We conducted this experiment in two middle schools in Yongkang, a county in southern China. The study sample consists of 7th and 8th graders from these two schools, which have 48 classes and 2,246 students in total. We target middle schools because school bullying tends to peak in early middle school or middle adolescence. To alleviate classification errors, we use multiple questions embedded with specific examples for five types of bullying behaviors: threatening, physical bullying, rumor spreading, social isolation, and cyberbullying. We find that around 70 percent of the students in the survey sample have been involved in at least one bullying incident during the past semester and that about 38 percent of the students admitted that they have been bullied by others at least once.

The intervention has a mulitfaceted impact on students and their parents. First, it significantly increases parental involvement on empathy-related activities among compliers. Learning together with children also significantly improves the empathy level of the treated parents. These parents are more likely to incorporate empathy-related activities into their parenting behaviors with a 3.4 percentage point increase in the likelihood of adopting a democratic parenting style. Second, it changes parenting behavior in general; it nudges parents to invest more time without crowding out monetary investments. Third, it leads to an improvement in students' empathy index with a 0.1 standard deviation. Students in treated classes are more likely to have prosociality and empathetic concern toward others. All of these factors may contribute to improving students' positive personality, such as self-esteem and mental health, and reducing students' stress, which help further prevent bullying. The finding that students become less stressed is highly consistent with those of the previous literature that bullying is positively correlated with depression (Olweus, 1991; Neary and Joseph, 1994; Slee, 1995).[3]

The intervention is effective in preventing bullying, which leads to a 4.4-percentage-point decrease in self-reported bullying victims, a 5.3-percentage-point decrease in bullying perpetration, and a 6.5-percentage-point decrease in the number of bullying victims. In treated classes, students are 6.1 percentage points less likely to witness bullying incidents and 5.2 percentage points less likely to be bystanders. The findings are robust to alternative specifications.[4] We also confirm the finding by looking at the effects on the bullying incidents reported by parents, which serve as a cross-reported check. Exploring different types of bullying behaviors, we find that the reduction in physical bullying is the most salient and leads to a 4.2-percentage-point decrease in bullying

[2] We design a mobile app that can be easily embedded into the most prevalent social media platform in China, *WeChat*, so that participants can easily access all the materials. More importantly, the app comes with a check-in feature and can automatically record the take-up status of participants. Furthermore, this feature helps researchers track the potential "leakage" of the intervention.

[3] Depression can be considered a way to internalize psychopathology. Namely, bullying can be an expression of difficulties or distress. Swearer et al. (2001) suggested that bully-victims may be at a higher risk for depression and anxiety. Thus, our study adds to the scarce literature showing that improving students' mental health can lead to reduction in bullying.

[4] We detect a similar pattern in alternative specifications—the model allows for misreporting and models with different sets of controls: (1) demographics, (2) social desirability scale, and (3) survey completion time.

perpetration, which is about 23 percent less than the prevalence in the control group (18% are bullies). Additionally, a reduction in physical bullying leads to a 5.5-percentage-point decrease in the number of victims of physical bullying.

This paper is among the first to meticulously examine the role of parental involvement in adolescents' empathy skill formation and school bullying prevention and also presents the real-world impact by providing insights into the design of effective antibullying programs. The family intervention design may be more cost-effective than costly interventions at either the school or classroom level. Our intervention indirectly targets potential bullies and victims with family education as a soft cushion. Beyond the purpose of tackling school bullying, empathy skills are closely associated with prosocial behaviors in a broader context.

Specifically, our study contributes to the following three strands of the literature. First, our paper contributes to the literature on the formation of bullying behavior and prevention. Most studies focus on the impacts of bullying on students' (mostly victims') outcomes, such as mental health, social exclusion and educational outcomes (Brown and Taylor, 2008; Eriksen et al., 2014; Ponzo, 2013).[5] However, the development of bullying behaviors has been less examined.[6] Our paper is among the first few experiments to analyze the potential causal pathways of bullying behaviors among teenagers and also compliments developmental psychology and other fields of social science that have found large correlations between bullying perpetration and various individual or contextual characteristics (Álvarez-García et al., 2015), including low empathy (Farrington and Baldry, 2010; Cook et al., 2010), a distant relationship between children and parents (Li et al., 2019), and even the online course format due to the COVID-19 pandemic (Bacher-Hicks et al., 2021).[7] Our paper is closely related to that of Sarzosa and Urzúa (2021) who built and estimated an empirical model of endogenous bullying with multiple outcomes and latent skills. However, due to data limitations, their analysis only focused on those being bullied, whereas our paper provides more rich evidence on various types of bullying behaviors among both bullying victims and bullying perpetrators.[8]

Regarding the literature on bullying prevention programs, tackling bullying through family education has been narrowly explored both in the literature and in the field. Compared to school-based programs, which are usually costly and whose effects usually fade as the programs are terminated, our intervention has the additional advantage of being low cost and more likely to

[5] Brown and Taylor (2008) and Eriksen et al. (2012) found a negative correlation between bullying and educational, behavioral and wage outcomes. In recent studies, researchers have attempted to estimate the causal impacts of being bullied on educational outcomes and later-life outcomes. Eriksen et al. (2014) employed data from Denmark and suggested that being bullied significantly decreases the academic outcomes of 9th graders in the short run. Likewise, Ponzo (2013) used data from Italy to show that bullying victimization in students in elementary and middle school results in them achieving lower grades.

[6] Xia (2019) suggested that the inequality of family socioeconomic status contributes to being a bully.

[7] Bacher-Hicks et al. (2021) used real-time Google search data during the COVID-19 pandemic in the U.S. and found that there is a positive correlation between online schooling and bullying incidents, especially cyberbullying incidents.

[8] We measure 5 types of bullying behaviors including both direct and indirect bullying. Our paper also complements the research on cyberbullying behaviors in Nikolaou (2017), who relied on the exogenous state-year variation in the implementation of anticyberbullying laws to causally show the strong impact of cyberbullying on youth suicidal behaviors.

have long-run effects, as it changes students' skills and peer relationships (Farrington et al., 2017; Castillo-Eito et al., 2020).[9]

Second, our paper is strongly connected to the literature on parental involvement and parenting. The experimental evidence of returns on time investment on noncognitive abilities is scarce. Existing studies on parental involvement have tended to focus more on improving cognitive and noncognitive abilities in early childhood (Cunha and Heckman, 2008; Cunha et al., 2010; Bono et al., 2016; Del Boca et al., 2017; Attanasio et al., 2020).[10] This paper complements the literature by providing evidence of the importance of parental involvement on children's development in later years, i.e., adolescence.[11] The interventions are designed with specific content on empathy skills rather than being nonsystematic. Moreover, the empirical literature on parental involvement has tended to highlight the positive impact on children's cognitive outcomes. For instance, Fiorini and Keane (2014) analyzed time use diaries from a large sample of children in the Longitudinal Study of Australian Children and found that children's time spent on educational activities with parents is a highly productive input for cognitive skills. Villena-Roldan and Ríos-Aguilar (2012) used various instruments for parents' time with their children and found that maternal educational time with a child has a direct causal effect on their children's math scores. Our paper adds to the literature by showing that noncognitive skills can also be cultivated and that parents who are more involved in the upbringing of their children have children with more favorable attitudes and traits. In contrast to papers relying on observational data to explore the role of parental involvement in children's socioeconomic skills (Moroni et al., 2019; Zumbuehl et al., 2021), we are able to provide causal evidence from a randomized experiment in the developing country context.

Third, our research also builds on the strand of literature on empathy skills, particularly prosocial behaviors.[12] Due to the fact that it is usually difficult to observe and measure empathy or prosocial skills, most studies have reled on experimental methods to explore the formation of prosociality and its further implications. Kosse et al. (2020) conducted a nonprofit mentor program targeting elementary school students for one year to study how prosociality forms and identified prosocial attachment figures and intense social interactions as the two main drivers of child prosociality, which is measured as altruism, trust, and other-regarding behavior in everyday life. Our research complements their study by exploring the other potential channel, i.e., the role of parental involvement, in guiding children's behaviors and focusing on the opposite of prosociality, i.e., bullying behaviors. Our intervention also adapts elements from Alan et al. (2021), who imple-

[9] Farrington et al. (2017) and Castillo-Eito et al. (2020) provided a summary on bullying prevention programs and found that most of the results from randomized control trials show modest or no effects, while before and after comparisons show large short-run effects.

[10] One can refer to Cunha et al. (2021) for a more systematic review and Attanasio et al. (2020) for the most recent analysis.

[11] Our paper is also distinct from Barrera-Osorio et al. (2020), who studied parental involvement programs across 430 public schools in Mexico and found no effects on students' educational outcomes. Through the targeting of older children, our paper emphasizes the cultivation of children's noncognitive skills.

[12] Zaki (2020) suggested that high levels of empathy improve prosocial skills. Studies have shown that empathy also helps reduce racial discrimination (Boisjoly et al., 2006).

mented an educational program in the context of a massive influx of refugee children in Turkey.[13] They found that the intervention increased social cohesion between local and refugee students and enhanced prosocial behaviors. Although both aim at improving children's empathy skills, our research is distinct from theirs in terms of the targeted age group, the detailed intervention, and the delivery format.[14]

This experiment and the results are quite generalizable across society where parental time investment is low or less valued. The discussion follows the four transparent SANS conditions (List, 2020): First, we select 7th and 8th graders in two schools (one public and one private) in a county (small city). Students in counties account for half of the total students in China. Our sample reflects that students are under high levels of stress from education and receive little time from their parents. Second, we have almost no attrition rate from the student side.[15] Third, our parental involvement treatment is quite natural to parents, as many of them often receive homework tasks from teachers. Reading and talking to children are also considered part of the daily routine. The experimental sessions took place in family and school settings, and all the decisions were real. Finally, the intervention costs almost nothing, only requiring teachers' effort in notifying parents. Participants did not receive any monetary incentives for taking part in the intervention, which helped enlarge the possibility of a scalable intervention. The low cost makes it easy to scale up, even though the effect may be moderate or small in some settings that are different from our sample. A comprehensive investigation of local situations of school bullying and parental involvement is needed to understand firsthand whether our results can be extended to other locations, especially in other developing countries.

The remainder of the paper is organized as follows. Section 2 introduces the background of adolescents in Chinese counties and the intervention. Section 3 presents the measures used, experimental design, sample distributions, and empirical strategy for the treatment effect analysis. Section 4 shows the results, as well as the heterogeneous effects. Section 5 presents the additional analysis. Section 6 concludes the paper.

2 Background and the Program

[13] Boucher et al. (2021) also analyzed the causal effect of a mixed ethnic program in Turkey targeting 5-year-old children; they built and structurally estimated a friendship network formation model to uncover the underlying mechanisms and found that interethnic exposure and language speaking skills are the two main drivers affecting social cohesion.

[14] First, their program mainly targets elementary school children, while ours tackles school bullying behavior and low levels of empathy among middle-school students. Second, our education program is parent-directed and emphasizes the interaction between parents and children rather than peer group interactions within the classroom. Third, our intervention develops other desirable personality traits among adolescents, such as self-esteem, that go beyond empathy skills and prosociality.

[15] The main attrition comes from parents' nonresponse, with around a 15% attrition rate that is quite balanced across the treatment and control groups.

2.1 Adolescents in Chinese Counties

Students in the early grades of middle school, aged between 13 and 15 years, are in a transition period between early and middle adolescence. They are considered rebellious and emotionally unstable. They are also often involved in school bullying. While most of the existing interventions in China focus on the school level (Chan and Wong, 2015), parents play a nonnegligible role in helping students navigate adolescence. Parents are overwhelmed by the educational arms race and often neglect students' socioemotional development and the school environment; the lack of socioemotional development may contribute to school bullying incidents. Compared with major cities, the situation is even worse in small counties, which is the focus of our study.

The study sample includes both 7th and 8th graders in one public school and one private school. The two participating schools are located in Yongkang, a county that belongs to Jinhua city, Zhejiang Province. Yongkang has managed to accommodate about 0.9 million residents and reached 103,163 RMB (about 15,000 USD) GDP per capita in 2020 (National Bureau of Statistics of China). Households in Yongkang are relatively wealthy compared to those in other counties in China. However, when looking at the education system in general, the students there still receive a test-oriented education, as do those in most regions of China. More importantly, baseline survey data show that parental time investment is, on average, low: 39% of students reported that their parents spent 0 hours checking their homework, while 41 % of students claimed that their parents spent 0 hours on outdoor activities on a typical weekend. Since parental time investment is a key input of children's future development, especially on noncognitive skills, it is urgent that parents be encouraged to get more involved into their children's education.

Our intervention targeting secondary schools in counties in a developing country has additional implications. Existing studies have tended to focus on schools located in either the most deprived area or the most accessible area. Studying secondary schools at the county level has the below additional value.

First, secondary schools in counties educate about half of China's students, but no single specific study has ever drawn on them.[16]

Second, schools in counties play an important role in the reduction in the rural-urban education division. In Table D1, we compare student composition and parents' characteristics in schools located in small counties with their counterparts located in cities and in rural villages/towns. There are significant differences along multiple dimensions. In general, compared with those in counties and rural areas, students in cities have a more advantaged background in terms of parental support. They also have higher levels of cognitive and noncognitive skills, creating another layer of inequality. In China, county-level cities or counties are ranked as the third level of the administrative hierarchy. There are thousands of counties in China and counties that accommodate half of the country's total population, linking the provincial- or prefectural-level cities with rural China

[16] County-level cities, or counties, are in the third level of the administrative hierarchy. There are thousands of counties, accommodating half of the country's population. These counties link the provincial- or prefectural-level cities with rural China.

and also contributing to reducing education inequality.

Third, schools in counties usually cannot retain talented teachers and have limited infrastructure. Parents in counties are more exam-oriented and focus heavily on preparing their children for all levels of entrance exams. Students suffer from dual pressures, as they often need to internalize both the limited resource constraints and extremely high parental expectations, resulting in an increased likelihood of experiencing stress. This dilemma leads to higher levels of depression and school bullying among students in schools located in small counties. Table D2 illustrates the school bullying situation at the baseline. Verbal bullying (threatening and spreading rumors) and physical bullying are very common; approximately 50% of students consider themselves victims and almost 20% classify themselves as perpetrators.

2.2 The Intervention

Our intervention is low cost, highly scalable, and parent directed. Figure D1 illustrates the theoretical framework of the intervention. The intervention content is aimed at helping parents learn empathy and positive parenting skills. Positive parenting encourages parents to spend more time with their children, especially regarding their mental health, personality trait development, and correction of misbehavior. We expect the intervention to improve parents' empathy and mental health and reduce their stress. Furthermore, following simulation theory from psychology (Preston and De Waal, 2002; De Vignemont and Singer, 2006), adolescents can develop their emotional skills, particularly empathy, by observing and communicating with their parents. In addition, students can cultivate these skills by watching and discussing the vivid examples in the tasks.

Based on this framework, we design the intervention to cover four monthly themes, as illustrated in Figure D2. The themes are empathy, perspective taking, the value of various personalities, and the role of socioemotional skills in maintaining relationships with others. Table D3 summarizes individual family involvement tasks.[17] For each monthly theme, we delivered two biweekly articles. The first biweekly article emphasized understanding the concept and the importance of the topic, while the second short article provided examples from the real world and detailed "*how-to*" procedures to educate students and parents.[18] In terms of the movie tasks, at the beginning of each treatment month, parents received the link delivered by the class teachers to access the movie and watch it together with their children.[19] All the movies are related to positive parenting or empathy.

Regarding program delivery, we incorporate the program into a platform with a special check-in feature to avoid and trace potential spillovers. The platform was embedded in *WeChat* to make

[17] The content of reading mainly relies on two books by American psychologists as references: *"The Power of Empathy: A Practical Guide to Creating Intimacy, Self-understanding, and Lasting Love in Your Life citation"* by A. Ciaramicoli and K. Ketcham and *"The Stress Solution: Using Empathy and Cognitive Behavioral Therapy to Reduce Anxiety and Develop Resilience"* by A. Ciaramicoli.

[18] For instance, the week 1 article introduced the concept of empathy and the potential benefits of being empathetic, while the week 3 article taught parents to incorporate empathy into their parenting and positive parenting techniques.

[19] All the links and materials were available on the platform.

it more convenient to use.[20] All the tasks and materials were uploaded and delivered on the platform, which accurately recorded parents' enrollment in the intervention with the check-in feature. Parents were asked to register on the platform using their children' student IDs, after which they could start carrying out the tasks. Successfully registering on the platform enabled parents to access the biweekly articles and monthly movies during the intervention. Parents and students were asked to read the specific articles and watch the assigned movies online together. Once a task was finished, students and parents were encouraged to submit a short reflection essay via the platform with no word limit.

3 Methods

3.1 Recruitment

We recruited students through principals and class teachers. Participating classes included both 7th and 8th graders in the selected public school and only 8th graders in the selected private school. The number of eligible classes was 48 in total with more than 2,200 students. The average class size was 45 in the public school and 51 in the private school. We prepared the consent form with a brief program introduction. School principals informed the class teachers, who recruited students and parents on our behalf. Teachers distributed the consent forms and explained the program to students and their parents in a parent meeting in early January 2021. As a part of the recruitment process, teachers illustrated the potential risks and the time required to participate in the intervention and presented the program introduction written on the consent form. Taking the survey and participating in the intervention were purely voluntary.

As the tasks are embedded in daily activities with no complications, almost all potential participants gave their consent to participate in this program.[21] Finally, 2,246 students and parents agreed to participate in the experiment.

3.2 Experimental Design

Figure 1 shows the timeline of the experiment starting in January 2021. We collected the baseline survey in mid-January after the parent meeting. Class teachers gathered students in computer lab rooms to complete the survey. In February, we randomly distributed half of the 48 classes into the treatment group and the other half into the control group. Specifically, each class is a cluster in this study. We used a stratified cluster randomization design. The study samples have four strata: 7th graders and 8th graders in the public school and "top classes" and standard classes (8th grader) in the private school. Within each stratum, classes were randomly assigned into the treatment and

[20] *WeChat*, like Facebook or WhatsApp in the United States, is the most popular social networking platform in China. Class teachers use *WeChat* groups to communicate with parents in all the classes of the study sample.

[21] Parents understood that even if they agreed to participate, they did not necessarily need to complete all tasks. In practice, almost 30% of them did not enroll in the intervention. Nevertheless, almost all students completed the two surveys, and more than 80% of parents completed the follow-up surveys.

control groups. In the beginning of the spring semester, we delivered the notification messages about the tasks for the treatment group to the two principals and they gathered the teachers in the treatment classes for a meeting. The teachers in the treatment classes learned about the intervention and agreed to deliver the notification messages to parents through *WeChat* messages during the intervention period. The teachers expected no extra workload except for forwarding the reminder messages of the tasks prepared by our research team to the *WeChat* group.

In early March, parents in the 26 treatment classes received a message inviting them to register on the platform and start the first month of tasks.In the following 4 months, each task message was delivered biweekly via the *WeChat* group of the treatment classes by the teacher at 7:30pm on Friday. In addition, we also asked the teachers of the treatment classes to send out a reminder message at the beginning of the third month. Students and parents in the control classes received no information during the intervention except for an invitation to the follow-up survey in late June. The follow-up survey was conducted right after the end of the intervention.

3.3 Measurement and Data

We collected measures on students' socioemotional skills and bullying behaviors in the baseline and follow-up surveys. To improve the quality of these surveys, students completed them using computer assisted self-interviewing (CASI) with the help of class teachers in the computer lab on campus. To elicit unbiased answers, students were informed of the importance of the confidentiality but were not informed of the actual purpose of the surveys.

We used the enrollment records to trace parents' actual take-up of the parent-child activity tasks of the program, which are biweekly data exported from the platform. We used school administrative data on students' test scores to measure students' academic performance covering two exams: one before the intervention and the other right at the end of the intervention.

Our data collection led to almost no attrition in the student follow-up survey, except for very few mistakes made regarding student ID or name, which were excluded from the sample (less than 0.5%). The main attrition of our study comes from the nonresponse by parents. Table D4 shows that the nonresponse rate by parents accounts for about 15.5% of the student sample and that the attrition rates are almost identical between the treatment (15.3%) and the control groups (15.6%). In total, the number of student samples and matched parent-child pairs were 2,246 and 1,899, respectively.

The primary outcome is school bullying behavior. The secondary outcomes include empathy, parental involvement, mental health and stress, and personality traits.

Bullying Behaviors To alleviate the concern of misreporting in bullying behaviors caused by misunderstanding, we use multiple questions with specific examples to cover five bullying 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 for another dimension of social bullying, and (5) abusive or hurtful texts online for cyberbullying.[22]

Comparing results across different domains also gave us more confidence in handling the Hawthorne reporting effects. One may expect a systematic improvement if the Hawthorne effects exist. Additionally, we include spectators' questions in the student survey[23] and parents' knowledge about their children being victims of the five bullying incidents in the parent survey.[24] From the third-person point of view, we expect the Hawthorne reporting effect to be negligible.

For empirical analysis, we collapse the five questions related to bullying behaviors into three indicator variables—bullying perpetrator, bullying victim and bully-victims—to have a general sense of overall bullying.[25] In addition, we also constructed accumulative measures for bully and victims separately by counting the total number of events that happened, as the indicator may mask the degree of severity among the different types of bullying events.

The use of self-reported bullying behaviors raises some concerns including those related to validity problems and memory, which have been well discussed in criminology research, in which self-reports have been widely used and have become a valuable method for measuring criminal involvement (Hindelang et al., 1979). The concerns are much less severe in self-reported bullying behaviors, as such bullying is generally not punishable by law and not always unacceptable or morally condemned among youths. Hence, the issues raised in crime research may be milder in the study of bullying. Additionally, Junger-Tas and Marshall (1999) suggested that the self-reported method is more valid and reliable with young adolescents than with adults.

Empathy Skill We use a self-reported instrument for students' empathy skill and follow Alan et al. (2021) to measure the two dimensions at the baseline: *perspective taking* and *empathetic concern*.[26] However, our sample students are on average about 69 months older than students in Alan et al. (2021). Therefore, we added another dimension, *prosociality*, in the followup, to construct a more valid empathy measure for adolescents. The new measure is closer to the modified *Interpersonal Reactivity Index* in the psychology literature (Davis, 1983). To overcome the potential multiple hypothesis testing issue, we follow Anderson (2008) to construct an inverse covariance weighting empathy index including the three subcomponents.

Parental Investment Time investment is measured by time spent (hours) on average on parent-child activities per day including reading, checking homework, playing and conducting general education with kids on weekdays and weekends, respectively, over the past week. For monetary investment, we asked parents their per-month education-related activity investment as a proportion of their total income, which we then divided into five categories: 5% or less, 5-10%,

[22] We restrict the focus to bullying incidents that happened within the class since it is the norm that kids in Chinese middle schools tend to stay in the same class for consecutive years until graduation.

[23] "Have you witnessed school bullying in the last semester?" and "What would you do if you saw someone being bullied?"

[24] For example, we asked, in the parents survey, "Has your child ever mentioned being threatened at school?" Similarly, we did the same for the other four types of bullying incidents.

[25] A student was classified as a bullying perpetrator/victim as long as she was involved in at least one of the five events. When a student was classified as both a perpetrator and a victim, she was automatically classified as a bully-victim.

[26] Kamas and Preston (2021) discussed different types of measures of empathy and concluded that a self-reported survey is considered a valid way to measure empathy.

10-25%, 25-50% and greater than 50% of total income.

Mental Health and Stress Mental health is measured using the 10-item Center for the Epidemiological Studies of Depression Short Form, or CES-D-10, which is a 10-item Likert scale questionnaire (Yang et al., 2018).[27] The depression indicator is generated with a threshold value of 12. The *inverse CESD index* is constructed by subtracting 30 from the CESD score so that a higher score indicates a better mental health status. We measure happiness using a 7-Likert scale with 7 being the happiest. In addition, we measure students' stress according to its source: (i) studies at school, (ii) peer relationships, (iii) rank/test scores in the class, and (iv) family background. 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 a *mental health index* consisting of these measurements so that a higher score indicates that students are happier and less stressed.

Positive Traits To better understand the mechanisms beyond the treatment effect, we collect other socioemotional measurements on positive traits including four aspects of positive self-image (self-satisfaction, self-worth, self-confidence, and self-esteem) and perseverance. We then construct a *positive trait index* following Anderson (2008).

Other Outcomes We also collect rich information on parents' inputs and empathy skills as well as parenting styles. Tables D5 and D6 summarize the detailed measures that we developed. Appendix Section A presents the survey questions on these outcomes in detail.

Baseline Characteristics Table 1 shows the descriptive statistics and tests the baseline balance between treatment and control groups. Panel A reports the demographic characteristics of students, and Panel B shows the baseline outcome variables, such as school bullying behaviors, depression and social emotional skills. Column (1) reports the summary statistics of the baseline outcome variables, while Column (3) shows the differences in students' characteristics between the control and treatment groups.

As shown in Panel A, students are, on average, 14.5 years old, and 53% of them are male. On average, urban residents consist of 46% of the sample. Having siblings is common in this sample, with 70% of students reporting that they have at least one sibling. The average height and weight of the sample are 161.9 and 50.5 kilograms, which is almost identical to the students in nationally representative China Education Panel Survey (CEPS) sample (height 161.6 and weight 49.9).[28] Compared with the summary statistics shown in Table D1, our sample has 6 percent more urban residents, and 10 percent less only-child families.

Students have interactions with peers in the classroom. Many of them indicated five good friends in the friendship network question; the mean number of good friends is 4.[29] In terms of school bullying behaviors in the intervened semester, 70.6% of students reported at least one bullying victimization experience, while 37.9% reported at least one type of bullying perpetration.

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

[28] The China Education Panel Survey (CEPS) is a large-scale, nationally representative, longitudinal survey starting with two cohorts of 7th and 9th graders in the 2013-2014 academic year. For more details, one can refer to http://ceps.ruc.edu.cn/English/Home.htm.

[29] The number may be underestimated, as the survey question was capped with nominating at most 5 friends.

Among the five types of school bullying behaviors, spreading rumors and physical bullying were the two top concerns.[30] Table D7 shows that male students and students with low empathy skills and poor mental health are more likely to become bullies.

The total empathy score ranges between 0 and 54, and students on average score 39.[31] Most of them express being somewhat empathetic in the survey. However, slightly more than 12% of them have a score less than half or 27.There is no standard threshold value for empathy measures, and we choose the medium value of the corresponding total score. In terms of other personalities, students on average have a positive self-image. The average scores of feeling self-satisfied, having self-worth, being self-confident and having self-esteem are all above 4 on a 7-point Likert scale. They are somewhat above average in perserverence, with a mean of 4.73 on the 7-point Likert scale. The average score on the final exam was 73% (460 out of 620), which they took the semester before the intervention.

Even though students were quite positive, many students reported feeling stressed at the baseline. The stress scores ranged from 0 to 18. Students on average scored 10.2, and 56% of them had a score above 9. The average CES-D score was slightly below 9. Using 10 as the threshold in CES-D score for depression, 30% of students were classified as being depressed, while 65% of students reported feelings of happiness in the past week, with an average score of 5.

We also found that parents and students were not close enough at the baseline; the total number of parent-child interactions within a normal week was 10.6. We report detailed categories of parent-child interactions in Column 1 in Table D8. The weekly average number of interactions indicated that students only dined with their parents 3 times and discussed their school lives with their parents slightly less than 4 times per week. Parents took students to play outdoor activities about 1.5 times and checked their homework assignments about 2 times.

Balancing Checks In Column 3 of Table 1, we test for imbalances in individual characteristics by reporting the mean differences between the treatment and control groups. We include both demographics (Panel A) and pretreatment outcomes (Panel B). We then report standard errors for item-wise regressions using the variables listed in Column 1 as the dependent variables on the treatment indicator. The standard errors are clustered at the unit of randomization, the classroom, and reported in Column 4. All 22 estimates of individual regressions are nonsignificant, and these estimates are in line with what one would expect under random classroom assignment.

Actual Take-up As participation was purely voluntary, we only relied on the biweekly task messages to nudge the parents.[32] Our platform records the actual registration and take-up of the tasks, and we use them to calculate biweekly task completion rates. In total, about 71% of eligible

[30] Table D2 reports the exact proportions for each behavior: i. threatening other students (34% and 13%); ii. spreading rumors (53% and 17%); iii. physical bullying (45% and 18%); iv. social isolation (18% and 8%); and v. cyberbullying (25% and 13%). Moreover, 63% of students have an exclusive, small group of friends.

[31] The follow-up raw empathy score ranges from 0 to 74.

[32] Teachers in treatment classes forwarded our messages to the *WeChat* group biweekly. We also reminded teachers not to forward the messages to other parents. The data verified that there was no accidental message leakage. The messages briefly introduced the topic of the month and encouraged parents to participate in the program.

parents (N=872) registered in the platform and enrolled in the program.

Figure D3 shows the task completion rates during the intervention. Overall, the task completion rates gradually decreased. The completion rates experienced a significant increase around the middle exam after the teachers forwarded the midterm reminder. It then plummeted to about 15% in the final month, as it was close to the time of preparation for final exams.

It is a four-month intervention, and a majority of parents were not consistent enough to complete the program. If parents and students completed at least half the movies or readings, then we considered them as having participated in the empathy education, and we defined the rest of the parents as having not participated in the empathy education. This classification allowed us to study the effect of empathy education on students' bullying behaviors. We also summed the number of movies and readings to study the dose effects of taking up empathy education.

The record allows us to verify potential spillover effects. Due to technology limitations, we were unable to make the link exclusive. It is possible that treated parents forwarded the link to others or that those in the control groups accidentally obtained access to the platform. We checked the enrollment rates and found that this spillover problem is negligible. The take-up rate is 0 in the control classes.[33] Thus, we are confident that randomization at the class level is robust to spillover, as there is almost no communication between students in different classes. Our experiment meets the noninterference assumption, as there is no spillover detected between classes. We discuss potential spillovers within classrooms in more detail in the Appendix section B.1.

3.4 Estimation of Program Effects

First, we estimate the intent to treat effect (ITT) by comparing outcomes across classes invited to participate in the education program (treatment) and other (control) classes. We follow the empirical specification as follows:[34]

$$Y_{ic1} = \alpha + \beta_1 T_c + \beta_2 Y_{ic0} + \tau_s + \epsilon_{ic}, \tag{1}$$

where Y_{ic1} is a vector of outcome variables for individual i in class c at time 2 (follow-up), T_c is the treatment group indicator for class c, which was assigned at time 1(baseline), Y_{ic0} is the baseline measures of the vector of outcome variables for individual i in class c, and τ_s is a set of strata fixed effects. In particular, we analyze the program's effects on school bullying behaviors. The richness of our data also allows us to explore program effects on detailed types of bullying behaviors, including threatening, spreading gossip, physical bullying, social bullying, and cyberbullying. For all regressions, we cluster the standard error at the class level using the Liang-Zeger estimator. Given 48 clusters, which is marginally greater than the rule of thumb, we complement it with Cameron et al. (2008)'s wild cluster bootstrap (WCB) p-values using 9,999 resampling. To

[33] Only 4 students (0.3%) in the control groups enrolled in the program, but they did not start any reading or movie tasks, thus, the actual take-up rate of the control groups was 0.

[34] McKenzie (2012) showed that in experiments with a single baseline and one follow-up survey, power is maximized when an end-line outcome is regressed on the treatment measure conditional on the baseline value of the outcome.

utilize the experimental nature, we also use a permutation test (Young, 2019) with 2,000 stratified clustered resampling.[35]

Second, individuals in treatment classes were encouraged to participate in the program intervention, but they did so to varying degrees, as illustrated in Section 3.3. Highly motivated students, for instance, may fare well, regardless of the program, but they may have a higher probability than other students of engaging in the program. Hence, we also estimate treatment-on-treated (TOT) models that account for individual students' participation. We consider whether students and their parents actually "took up" the intervention, which is defined as having engaged in at least half of the reading or movie tasks. We estimate the TOT models using a two-stage least-squares (2SLS) design wherein the endogenous variable (the "take up") is instrumented by the treatment assignment indicator with the actual take-up data exported from our platform. To fully use the actual take-up records, we also conduct a TOT-dosage analysis using the number of readings and movie activities for which participants registered. Following Sylvia et al. (2021), we estimate the 2SLS using the number of completed tasks as the regressor and control function approach, which we explain in detail in the Appendix section C.1.

To better understand the effects of the intervention, we further explore the heterogeneous effects. We follow the generalized random forest (GRF) method proposed by Athey et al. (2019) to capture heterogeneity in a more disciplined way.[36] The advantage of the GRF method is that it can be easily applied to deal with high-dimensional baseline characteristics and predicts an importance rank for each variable based on its contribution to the heterogeneous treatment effects. GRF extends the random forest approach by introducing and estimating a weighting function instead of simply averaging across all trees of a forest and using these weights to solve local moment equations. Once the importance rank is obtained, we then use these predictions as a guide and conduct the traditional heterogeneity analysis based on the selected baseline characteristics. The details of the method are discussed in the Appendix section C.2.

4 Results

The program evaluation follows the theoretical framework shown in Figure D1. We start with the direct effect of the program on parents' involvement and their parenting skills, which further affects students' empathy skills, positive traits and mental health. Improved students' socioemotional skills lead to a better classroom climate with fewer bullies, victims and bystanders in the treated classes.

[35] The computation of permutation p-value uses the exact small sample distribution of the test statistic. The method accounts for the complexity of the randomization and secures robust inference in small sample sizes (Walker et al., 2021). To alleviate the concern of relatively small clusters (48), we compute the wild cluster bootstrap (WCB) p-values.

[36] For a more technical explanation of the GRF algorithm, one can refer to Athey et al. (2019), and for a less technical explanation and examples of the application of the GRF algorithm to policy impact evaluations, one can refer to Davis and Heller (2017). We apply the 'grf' package in R to implement the algorithm: https://cran.r-project.org/web/packages/grf/grf.pdf.

4.1 Impacts on Parental Involvement and Outcomes

Taking up the program nudges parents to increase their time investment and educates them on improving their parenting skills. According to the general human capital production function of Cunha and Heckman (2007):[37]

$$h_{t+1} = f_{t+1}\left(h_t, M_{t+1}, T_{t+1}, P_{t+1}, X_t, \eta_{t+1}\right).$$

The program improves students' socioemotional skills through increasing these inputs in this function, such as M_{t+1}, T_{t+1}, and P_{t+1}. Therefore, we examine the program impacts on parents' inputs prior to its impact on students' skills.

We first explore the effects on different categories of time investment, including reading and talking with children, checking kids' homework, playing with kids, and educating them during the week and on weekends. We complement the study with the potential effects of the monetary investment (Del Boca et al., 2014). Parental involvement is particularly important for children's development. The changes in parents' outcomes have significant effects on children's outcomes. We also study the program effects on parents' outcomes, including empathy skills and parenting style.[38]

Panel A in Table 2 reports the treatment effects on parents' investments. The estimates show that parents in the treated group significantly increased their time investment during weekdays, with no significant effects on time investment on the weekends or on monetary investment. We continue to test the effects on detailed measures of specific categories of parent-child activities on weekdays and weekends. First, the program led to a significant increase in time investment in children on weekdays. When looking at specific types of parent-child interactions, there was a significant increase in reading with children and checking children's homework, as reported in Panel A in Table D9. This finding is consistent with a recent report by Doepke and Zilibotti (2021), which shows a global increasing trend in intensive parenting, especially in involvement in education-oriented activities. Although not significant on other types of parent-child activities, all the coefficients have positive signs. Additionally, Panel B in Table D9 shows that the program neither crowds out any monetary investment in after school tutorials nor changes their beliefs about afterschool tutorials.

The increased time investment is also due to an increase in the take-up of empathy-related reading and activities. Table D10 presents the estimated ITT effects on the engagement of empathy-related parent-child activities. There is ample evidence that the program significantly increased the engagement of empathy-related activities, including watching empathy-oriented movies and read-

[37] In the equation, h_{t+1} and h_t are child skills at endline and at baseline, respectively, M_{t+1}, T_{t+1}, and P_{t+1} are the caregiver's monetary investments, time investments, and parenting skills during the intervention period, X_t is baseline household characteristics, and η_{t+1} is a random shock to child skill development.

[38] According to simulation theory in psychology that empathic response is rather automatic and can be stimulated by the observation of imagination of another person's affective state (Preston and De Waal, 2002; De Vignemont and Singer, 2006), we expect that the interaction between parents and children with the specific aim of developing empathy and prosociality may also affect the formation of the empathy skills of both.

ing short articles on empathy. We measured the frequency of engaging in the activity at least once and engaging in the activity monthly during the semester and detected that the greatest increase was in more frequent engagement—at least monthly.

Panel B in Table 2 reports the effects on parents' empathy and parenting skills. The intervention leads to a marginally significant increase of the empathy score of the parents compared to the control group. Furthermore, treated parents are 3.4 percentage point more likely to adjust their parenting style toward being more democratic.

4.2 Impacts on Students' Skills

The goal of parental involvement tasks is to improve students' empathy skills and to increase parents' time investment with better parenting skills, which can further help students develop better socioemotional skills. We expect these improvements to further reduce school bullying for three reasons: i. being empathetic to others and developing perspective taking may increase the psychological costs of bullying; ii. better socioemotional skills and less stress may reduce the bullying motivation; and iii. being prosocial may reduce the problems of bullying perpetration and being a bystander. Hence, we proceed to test the program effects on students' empathy skills, positive traits and mental health.

Panel A in Table 3 reports that the program leads to a 0.9-standard-deviation increase in the overall empathy index. When looking at the three domains, the improvement in prosociality (0.16 SD) and empathetic concerns (0.10 SD) are the main drivers. The TOT estimates suggest that taking up the program (completing at least half of the tasks) leads to a 0.40-standard-deviation increase in prosociality, a 0.24-standard-deviation increase in empathetic concern and a 0.22-standard-deviation increase in perspective taking. The results are consistent with those in Table D11, indicating that the program significantly improves students' prosociality measured in hypothetical scenario questions.

Panel B reports that the program leads to a significant improvement in the positive traits index, which summarizes improvements in self-esteem and perseverance. The TOT estimate suggests that taking up the program and finishing more than half of the tasks improve students' positive traits index by 0.34 standard deviation.

Panel C suggests that the program also helps students improve their mental health status, which summarizes measurements of mental health status (CES-D 10 and happiness) and four types of stress students face. The improvement are attributed mainly to the fact that parental involvement greatly reduces student stress.

Table D12 suggests that the program has no effect on students' test scores. Taking up the program did not crowd out students' time for courses and it should not have any short-run negative effects on students' academic outcome. The improvement in noncognitive skills may complement students' academic skills dynamically; however, it takes time for the program to have detectable impacts on test scores if there is dynamic complementarity between socioemotional skills and academic outcomes.

4.3 Impacts on Bullying

We expect that the improvement in students' socioemotional skills will reduce bullying incidents, as these skills are highly correlated with less bullying involvement. Table D7 shows that students' baseline empathy skills, positive traits and mental health are negatively correlated with their likelihood of being bullies and victims of bullying in both baseline and follow-up.[39]

Table 4 reports the program effects on children's bullying behavior. Column 2 reports the ITT estimates and standard errors using (1). The program reduces school bullying with a significant decrease in bullies, victims, bully-victims, witnesses and bystanders among the treatment classes. In the treated classes, we detect a 5.3-percentage-point decrease in students who classified themselves as bullying perpetrators, a 4.4-percentage-point decrease for bullying victims, and a 6.5-percentage-point decrease for bully-victims. Treated students are less likely to witness bullying incidents (6.1%) or to be bystanders (5.2%). Columns 3 and 4 report the permutation and WCB p-values. The inferences based on exact p-values are almost identical as the clustered robust standard errors. As for reporting concerns, the results are robust to various specifications in the robustness test and HAC probit model, allowing for misclassification errors (false negative rate), as discussed in Section 5.

We construct the accumulative bullying measure by taking into account the types of bullying behaviors, which has a value between 0 and 5. The difference between an indicator and the accumulative measure suggests that bullies and victims are often involved with more than one type of bullying behavior. In the control groups, the mean of bullying perpetration and victimization incidents are 0.81 and 1.65, respectively. ITT estimates on the accumulative measure show a more significant reduction in bullying incidents for both bullies and victims.

From the perspective of observers, students in the treatment classes reported being less likely to witness bullying incidents. The intervention also encouraged students to help bullying victims and to be less likely to become bystanders; i.e., it reduces the bystander effect, which is also an important predictor of bullying behavior (Salmivalli, 2010; Polanin et al., 2012).

It is also meaningful to go beyond ITT estimates and study the effects of program take-up on school bullying behaviors. We consider these parents who completed at least half of the movies or half of readings as take-up group. For parents who only complete one of the tasks, they are more likely to just enroll the program and not really invest much time in it. Using the take-up indicator, we estimate the effect of taking up the program on the outcomes using 2SLS in Column 5 of Table 4. 2SLS reports larger effects, as the ITT estimates are generally a lower bound of the effects. The estimates of the TOT-dose regressions can be found in Column 6 of Table 4. The does-response effects are all positive, and we detect significant linear effects of additional tasks on reducing the likelihood of becoming bullies and bully-victims. Students are also less likely to witness bullying incidents and to be bystanders. We detect marginally significant effects on

[39] Furthermore, boys and students with more friends and who are members of exclusive groups are more likely to get involved in school bullying, which suggests that many school bullying incidents are likely to be collective behaviors; further studies incorporating students' networks would be valuable.

reducing bullying victimization.

It is possible that the program may encounter spillover effects via interactions between students or between parents. We discuss this in the Appendix section B.1 and find that the spillover effects of taking up the program is modest. There is a significant negative correlation between the class-level take-up rate and individual bullying behaviors, but the significant effect faded once we instrumented the class-level take-up rate with the random assignment indicator.

Detailed Types of Bullying Incidents Bullying incidents can be categorized into two groups: direct and indirect bullying.[40] Improving empathy and noncognitive skills may help explain the reduction in direct bullying, as with direct bullying, it is easy to identify targets and very likely to imagine the consequences, such as physical bullying or threats, while these improvements may have less impact on indirect bullying, such as social isolation, rumor spreading or cyberbullying. These indirect bullying behaviors are difficult to be witnessed or evoke peers' empathy.

To look beyond the aggregated bullying behaviors, we study the program impacts on bullying behaviors by type. Panel A of Figure 2 presents the levels of the five types of bullying for treatment and control groups at follow-up using histograms. Almost 20% of students admitted to having spread rumors or physically bullying others. The least frequent bullying perpetration behavior is cyberbullying (11%). Students reported a higher rate of being bullying victims compared to being bullying perpetrators. The rate ranges from the highest, with almost 50% of students having experienced being the target of rumors, to the lowest rate of above 20% of students who were victims of cyberbullying and social isolation. The fraction of bully-victims is lowest, as they are defined as being both bullies and victims. In sum, rumor spreading and physical bullying are the two most frequent bullying behaviors in the study sample.

Comparing the levels between the treatment and control groups, treatment classes show lower levels in all the five types of bullying behaviors. The difference between the levels of the types of bullying behaviors are either significant or marginally significant. We analyze these differences by estimating (1) for different types of bullying. Panel B of Figure 2 presents the ITT estimates, for which we report the exact numbers in Table D13. Panel A in Table D13 illustrates the effects on bullying perpetration, Panel B shows the results on bullying victimization, and Panel C shows the results on being a bully-victim. The coefficients are all negative, regardless of the type of bullying behaviors.

We find that physical bullying and rumor spreading experienced the greatest decrease. In the case of physical bullying, there is a 3.8-percentage-point decrease in being a bullying perpetrator, a 5.5-percentage-point decrease in being a bullying victim, and a 3.5-percentage-point decrease in being a bully-victim. Physical bullying is one of the most common types of bullying and causes the greatest problems on campus. The reduction in physical bullying is meaningful for creating safer studying environments. In the case of rumor spreading, although there was no significant

[40] According to van der Wal et al. (2003), direct bullying is a type of behavior that hurts, harms, or humiliates and is overt, obvious, and apparent to anyone witnessing it, whereas indirect bullying is not always immediately acknowledged as bullying. The different definitions often lead to different solutions and causes of these two types of bullying.

decrease in bullying victimization, we detected a 3.8-percentage-point decrease in being a bullying perpetrator and a 3.6-percentage-point decrease in being a bully-victim. Furthermore, the significance levels of estimates for the other types of bullying behaviors vary around 0.1. The results suggest that the reduction in physical bullying and rumor spreading may serve as the main drivers of the reduction in aggregated school bullying. There is also a significant decrease in being a bully-victim in the case of cyberbullying. The effects on social isolation and threatening are less salient. In Table D14, we further test results on the intensive margin of bullying behaviors. Again, we find that the reduction in physical bullying is the most salient in terms of both bullying perpetration and victimization.

4.4 Heterogeneous Treatment Effects on Bullying

In this section, we explore heterogeneity in the treatment effects on school bullying behaviors. We first look at effects on students' bully or victim status at baseline. We then explore the heterogeneous effects informed by GRF.

Baseline Bully Status To better understand the "transition" of the bully status induced by the intervention, we investigate the program impact on bullying by students' bully or victim status at baseline.[41]

Figure D5 shows the program impacts on the bullies (Panel A) and victims (Panel B) by four bully categories: being a bully, being a nonbully, being a victim, and being a nonvictim at baseline. From Panel A, we can see that our program seems to deliver a larger impact on reducing bullying among those who were nonbullies or nonvictims at baseline (although the difference is not significant). The program is more effective in reducing victimization among groups of students who were nonbullies and victims at baseline. Our program successfully helped the baseline victims get out of victimization. However, we find that the program effects on victimization are small for baseline bullies while the program effects on bullying perpetration are smaller for baseline victims, which suggests that bullying perpetration and victimization are highly entangled.

Heterogeneous Treatment Effects Informed by GRF The rich data collected in the baseline allow us to explore the heterogeneous treatment effects over a variety of dimensions, including the different degrees of maternal involvement, preprogram parental involvement, empathy skills, personality traits, ranks of test scores, associated pressures, and other socioemotional characteristics. We apply the GRF algorithm to limit heterogeneity tests while minimizing the probability that important sources of heterogeneity are neglected. The GRF algorithm selects four baseline characteristics with the highest importance rank: empathy skill, age, parental involvement, and pressure score.[42] It is well-established that male students are more likely to be involved in bullying incidents. Thus, we also test the impact of heterogeneity by gender.

[41] One advantage of our data is that we know the exact bully status, i.e., whether or not the student is a bully, at both baseline and follow-up.

[42] For more detailed discussions, one can refer to the Appendix section C.2.

Table D15 shows the heterogeneous effects on the indicator of being a bullying perpetrator (Panel A) and being a bullying victim (Panel B). When looking at the heterogeneous effects on the bullying perpetrator indicator, we find that treatment effects are significantly higher for children who experienced low levels of parental investment before the start of the program. Children in the lowest quartile of the pre-intervention parental investment distribution experienced a decrease in becoming a bullying perpetrator that was 7.8-percentage-points larger than that of children in the top three quartiles of baseline parental investment on average (Column 3). These findings are consistent with the literature on the importance of parental involvement on children's development (Zumbuehl et al., 2021; Attanasio et al., 2020; Attanasio et al., 2020). Similarly, we find that children with low baseline pressure scores benefited significantly more from the program than did other children (Column 4). The average treatment effect on being a bullying perpetrator is 6.8-percentage-points higher for children who had pressure scores in the lowest quartile at the start of the intervention compared with those in the top three quartiles. Studying the heterogeneous effects by pressure score contributes to understanding the unexpected consequences of China's extremely competitive education system on student outcomes (Qu and Guo, 2021; Jia and Li, 2021). We detect no significant heterogeneous effects among the results on being a bullying victim.

We continue to study the heterogeneous effects on being a bully-victim and bystander in Table D16. The results for being a bully-victim appear to deliver a similar pattern as those for being a bullying perpetrator. We find significantly higher treatment effects for those with lower parental involvement and lower pressure scores at the beginning of the intervention. We do not detect any differences along the other dimensions. The results in Panel B show a heterogeneous effect on the willingness to help bullying victims. Compared to female students, we find that male students benefit more from the program, as they are less likely to be bystanders after the intervention. This finding suggests that male students may be less empathetic and, thus, experience higher treatment effects from the empathy education program. We additionally explore the heterogeneous effects on students' empathy skill in Table D17. Although we hardly find any significant differences along the four dimensions, one thing to note here is that compared with female students, male students indeed have a lower level of empathy skills, and the heterogeneous effects by gender is marginally significant, as shown in Column 5.

In summary, the program appears to have a stronger effect on bullying reduction among those children who have less parental involvement and less study pressure at baseline. Male students are less likely to be bystanders after the empathy education program. The marginal return on parental involvement may be higher when baseline-level parental involvement is low. Students with less study pressure may value parental involvement more than may other students .

5 Robustness Analysis

5.1 Misreporting

The self-reported measures of bullying behaviors may suffer from misreporting as bullying behaviors tend to be considered a sensitive subject. In this subsection, we explore the robustness of the previously estimated specification to allow for the misreporting of bullying behaviors.

Misclassification Error As suggested by Hausman et al. (1998) (HAS), we correct for the misclassification error in a binary choice model (i.e., when a response is reported in the wrong category). The details of the methods are illustrated in the Appendix section C.3. First, we additionally control for the baseline social desirability scale on top of the individual demographics to correct the potential bias. The results are shown in Table D18. The odd columns report the treatment coefficients assuming that students truthfully report, while the even columns report the HAS treatment coefficients assuming that there is a false negative reporting rate. Since we control for the baseline social desirability scale, the false negative rate is low. For bullying perpetration and bully-victims, the false negative rate is close to zero, while it is 0.092 for reporting as a bullying victim.

Then, we follow De Paula et al. (2014) to report the treatment effects, assuming different levels of misreporting. The probit estimates are shown in Table D19. We can see that the patterns are quite similar. All the specifications show negative treatment effects on bullying behaviors. We find that misreporting leads to a downward bias of the treatment effect and that higher levels of misreporting lead to higher coefficient magnitudes. We also explore the misreporting on detailed bullying behaviors in Table D20. Again, we find a similar trend.

Hawthorne Effect Although there may be some concern about the Hawthorne effect, there are several reasons why this effect should be negligible. First, if this is the case, then we should detect a significant decrease in all types of bullying outcomes. Furthermore, the spectators' question about whether they "witnessed bullying incidents" also supports the absence of the Hawthorne effect. Second, the program is advertised as a family education and empathy program. All the teachers and participants (students and their parents) were unaware of the actual purpose of the program; they were only notified that there would be a series of parent-child activities.[43] Third, teachers were not monitoring parent-child activities. They were not present during these activities and thus were only responsible for forwarding the text messages to the *WeChat* group of parents and had no access to content. Thus, we expect almost no teacher effect on students' reporting.

Finally, we conduct a cross-check by exploring the bullying incidents reported by parents. We asked parents whether their child ever talked about becoming a victim of any of the five types of bullying behaviors in the parent survey. The ITT estimates shown in Table D21 confirm the program's significant effect on bullying reduction; the program leads to a 4.2-percentage-point decrease for bullying victims. According to parents' answers, rumor spreading, social isolation and physical bullying are the top three bullying behaviors that have the most significant reduction in terms of being a victim of bullying.

[43] Additionally, there is no word related to bullying in the main content of the materials.

5.2 Alternative Specifications and More Evidence

In Table D22, we conduct several robustness checks to further support the main findings. In sum, the promising effects on school bullying reduction are robust to alternative specifications.

Column 1 shows the results from our main specification taken from Column 2 of Table 4 for comparison purpose. Column 2 of Table D22 reports the ITT estimates controlling for demographic characteristics, such as age, age squared, gender, and an indicator of being an only child. Column 3 further controls for survey completion length and its square. Column 4 reports estimates with additionally controlling for the social desirability bias index, following Dhar et al. (2018). Finally, Column 5 shows results from a pooled regression in which we view an individual type of bullying behavior as an observation and pool them together, which generates a sample that is five times larger. We estimate the effects on bullies, victims and bully-victims with strata and type fixed effects and adjust standard errors clustered at the class and type levels. The results are similar across different specifications except for two cases. The ITT estimates for bullying victims become marginally significant once the social desirability scale is controlled for and when we conduct the pooled regression.

In Column 6, we also implement the entropy balancing (EB) method of Hainmueller (2012) to estimate TOT as a comparison.[44] EB is a multivariate reweighting method that makes the control group data match the covariate moments of the treatment group.[45] Similarly to matching methods, EB can deal with selection on observables. However, EB has been shown to achieve a significantly higher level of covariate balancing than common propensity score approaches (Hainmueller, 2012).[46] From the results, we can see that EB gives more conservative estimates that lie within the range of ITT and TOT.

We also investigate the treatment effects on detailed time use reported by the students, as a form of robust evidence for parental time investment results in Table 2. The results are shown in Table D8. We can see that the program significantly leads to students in the treated groups being more likely to talk with and have their homework checked by their parents. Students in the treated groups also reported a significantly higher frequency of engaging in outdoor activities with their parents.

[44] This part of the analysis was conducted using the "`ebalance`" command in Stata following the instruction of Hainmueller and Xu (2013) and applying the default tolerance level of 0.015.

[45] The entropy balancing scheme assigns a scalar weight to observations in the control group such that the control group's distributions of all selected covariates match the treatment group's covariate distributions on the first and second moment. This strategy produces a sample in which the means and variances of all selected control variables are the same in the treatment and control groups. Of all the possible weighting schemes that fulfill these balancing requirements, entropy balancing chooses the one where all weights are nonnegative and that deviate the least from uniform weights.

[46] In Table D23, we further show summary statistics for the variables used for EB. As a comparison, we also report the standardized difference in means and compare the results with the traditional propensity score matching (PSM) method. The table highlights the differences between the two methods. After entropy balancing, the standardized difference in the means of all covariates is below 5% (Column 7), the criterion for successful matching proposed by Caliendo and Kopeinig (2008), and performs much better than PSM.

6 Conclusions

To tackle the issues of increasing school bullying and low socioemotional skills among adolescents, we conducted an experiment of a directed parental involvement program in two large middle schools in a Chinese county, targeting 2,246 7th and 8th graders and their parents. Motivated by the research evidence that indicates bullying behavior is usually associated with a lack of parental care and low levels of empathy, our intervention comprises biweekly activities that lasts for four months to develop a closer relationship between parents and children. We encourage parents in the treatment groups to watch monthly empathy-oriented movies and reading biweekly articles on empathy and positive parenting with their children.

With the monthly take-up rate being approximately 40%, the program is still moderately effective in reducing school bullying incidents and improving students' empathy. Additionally, the treated students are more stress-resilient, self-satisfied and less likely to be depressed than are the untreated students.

The program also improves the treated parents' empathy levels, which they manage to incorporate into their parenting behaviors; parents in the treated groups are found to experience a 3.4-percentage-point increase in adopting a democratic parenting style.

Finally, the modest take-up rate provides a direction for future improvement. The program can be easily scaled up for several reasons: (i) the intervention is simple and easily regulated by a smartphone application, which incurs little cost; (ii) the program is purely voluntary and is not likely to affect parent-teacher relationships; (iii) the intervention only lasts for a semester; and (iv) the increased time spent on the program successfully nudges parents to spend more time with their children without crowding out monetary investments. The incorporation of an encouragement mechanism into this type of directed parental involvement program may generate larger effects on students' personality development and reduction in school bullying.

References

Alan, S., C. Baysan, M. Gumren, and E. Kubilay (2021). Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking. *The Quarterly Journal of Economics*. Forthcoming.

Álvarez-García, D., T. García, and J. C. Núñez (2015). Predictors of school bullying perpetration in adolescence: A systematic review. *Aggression and Violent Behavior 23*, 126–136.

Anderson, M. L. (2008). 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 103*(484), 1481–1495.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2020). Estimating the production function for human capital: results from a randomized controlled trial in colombia. *American Economic Review 110*(1), 48–85.

Attanasio, O., C. Meghir, and E. Nix (2020). Human capital development and parental investment in india. *The Review of Economic Studies 87*(6), 2511–2541.

Bacher-Hicks, A., J. Goodman, J. G. Green, and M. K. Holt (2021). The covid-19 pandemic disrupted both school bullying and cyberbullying. Technical report, Annenberg Institute for School Reform at Brown University.

Barrera-Osorio, F., P. Gertler, N. Nakajima, and H. Patrinos (2020). Promoting parental involvement in schools: Evidence from two randomized experiments. Technical report, National Bureau of Economic Research.

Boisjoly, J., G. J. Duncan, M. Kremer, D. M. Levy, and J. Eccles (2006). Empathy or antipathy? the impact of diversity. *American Economic Review 96*(5), 1890–1905.

Bono, E. D., M. Francesconi, Y. Kelly, and A. Sacker (2016). Early maternal time investment and early child outcomes. *The Economic Journal 126*(596), F96–F135.

Boucher, V., S. Tumen, M. Vlassopoulos, J. Wahba, and Y. Zenou (2021). Ethnic mixing in early childhood: Evidence from a randomized field experiment and a structural model. *IZA Discussion Paper*.

Brown, S. and K. Taylor (2008). Bullying, education and earnings: evidence from the national child development study. *Economics of Education Review 27*(4), 387–401.

Caliendo, M. and S. Kopeinig (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys 22*(1), 31–72.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics 90*(3), 414–427.

Castillo-Eito, L., C. J. Armitage, P. Norman, M. R. Day, O. C. Dogru, and R. Rowe (2020). How can adolescent aggression be reduced? a multi-level meta-analysis. *Clinical Psychology Review 78*, 101853.

Chan, D. W. (1985). The chinese version of the general health questionnaire: does language make a difference? *Psychological Medicine 15*(1), 147–155.

Chan, H. C. O. and D. S. Wong (2015). Traditional school bullying and cyberbullying in chinese societies: Prevalence and a review of the whole-school intervention approach. *Aggression and Violent Behavior 23*, 98–108.

Cook, C. R., K. R. Williams, N. G. Guerra, T. E. Kim, and S. Sadek (2010). Predictors of bullying and victimization in childhood and adolescence: a meta-analytic investigation. *School Psychology Quarterly 25*(2), 65.

Copeland, W. E., D. Wolke, A. Angold, and E. J. Costello (2013). Adult psychiatric outcomes of bullying and being bullied by peers in childhood and adolescence. *JAMA Psychiatry 70*(4), 419–426.

Crowne, D. P. and D. Marlowe (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology 24*(4), 349.

Cunha, F. and J. Heckman (2007). The technology of skill formation. *American Economic Review 97*(2), 31–47.

Cunha, F. and J. J. Heckman (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources 43*(4), 738–782.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Cunha, F., E. Nielsen, and B. Williams (2021). The econometrics of early childhood human capital and investments. *Annual Review of Economics 13*, 487–513.

Dahlberg, L. L., S. B. Toal, M. H. Swahn, and C. B. Behrens (2005). 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.

Davis, J. and S. B. Heller (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review 107*(5), 546–50.

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

De Paula, Á., G. Shapira, and P. E. Todd (2014). How beliefs about hiv status affect risky behaviors: Evidence from malawi. *Journal of Applied Econometrics 29*(6), 944–964.

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

Del Boca, D., C. Flinn, and M. Wiswall (2014). Household choices and child development. *Review of Economic Studies 81*(1), 137–185.

Del Boca, D., C. Monfardini, and C. Nicoletti (2017). Parental and child time investments and the cognitive development of adolescents. *Journal of Labor Economics 35*(2), 565–608.

Delavande, A. and H.-P. Kohler (2016). Hiv/aids-related expectations and risky sexual behaviour in malawi. *The Review of Economic Studies 83*(1), 118–164.

Dhar, D., T. Jain, and S. Jayachandran (2018). Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in india. Technical report, National Bureau of Economic Research.

Doepke, M. and F. Zilibotti (2021). Do rising returns to education justify "helicopter" parenting? *IZA World of Labor 487*.

Eisenberg, N. and P. A. Miller (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin 101*(1), 91.

Eriksen, T. L., H. S. Nielsen, and M. Simonsen (2012). The effects of bullying in elementary school. *IZA Discussion Paper*.

Eriksen, T. L. M., H. S. Nielsen, and M. Simonsen (2014). Bullying in elementary school. *Journal of Human Resources 49*(4), 839–871.

Falk, A., A. Becker, T. Dohmen, B. Enke, D. Huffman, and U. Sunde (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics 133*(4), 1645–1692.

Farrington, D. and A. Baldry (2010). Individual risk factors for school bullying. *Journal of Aggression, Conflict and Peace Research 2*(1), 4–16.

Farrington, D. P., H. Gaffney, F. Lösel, and M. M. Ttofi (2017). Systematic reviews of the effectiveness of developmental prevention programs in reducing delinquency, aggression, and bullying. *Aggression and Violent Behavior 33*, 91–106.

Fiorini, M. and M. P. Keane (2014). How the allocation of children's time affects cognitive and noncognitive development. *Journal of Labor Economics 32*(4), 787–836.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis 20*(1), 25–46.

Hainmueller, J. and Y. Xu (2013). Ebalance: A stata package for entropy balancing. *Journal of Statistical Software 54*(7).

Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics 87*(2), 239–269.

Hindelang, M. J., T. Hirschi, and J. G. Weis (1979). Correlates of delinquency: The illusion of discrepancy between self-report and official measures. *American Sociological Review*, 995–1014.

Hinduja, S. and J. W. Patchin (2010). Bullying, cyberbullying, and suicide. *Archives of Suicide Research 14*(3), 206–221.

Jia, R. and H. Li (2021). Just above the exam cutoff score: Elite college admission and wages in china. *Journal of Public Economics 196*, 104371.

Jolliffe, D. and D. P. Farrington (2006). Examining the relationship between low empathy and bullying. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression 32*(6), 540–550.

Junger-Tas, J. and I. H. Marshall (1999). The self-report methodology in crime research. *Crime and Justice 25*, 291–367.

Kamas, L. and A. Preston (2021). Empathy, gender, and prosocial behavior. *Journal of Behavioral and Experimental Economics 92*, 101654.

Kosse, F., T. Deckers, P. Pinger, H. Schildberg-Hörisch, and A. Falk (2020). The formation of prosociality: causal evidence on the role of social environment. *Journal of Political Economy 128*(2), 434–467.

Li, J., A. M. Sidibe, X. Shen, and T. Hesketh (2019). Incidence, risk factors and psychosomatic symptoms for traditional bullying and cyberbullying in chinese adolescents. *Children and Youth Services Review 107*, 104511.

List, J. A. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. Working Paper 27535, National Bureau of Economic Research.

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more t in experiments. *Journal of Development Economics 99*(2), 210–221.

Miller, P. H., S. D. Baxter, J. A. Royer, D. B. Hitchcock, A. F. Smith, K. L. Collins, C. H. Guinn, A. L. Smith, M. P. Puryear, K. K. Vaadi, et al. (2015). Children's social desirability: Effects of test assessment mode. *Personality and Individual Differences 83*, 85–90.

Molcho, M., W. Craig, P. Due, W. Pickett, Y. Harel-Fisch, and M. Overpeck (2009). Cross-national time trends in bullying behaviour 1994–2006: findings from europe and north america. *International Journal of Public Health 54*(2), 225–234.

Moroni, G., C. Nicoletti, and E. Tominey (2019). Child socio-emotional skills: The role of parental inputs. *IZA Discussion Paper*.

Neary, A. and S. Joseph (1994). Peer victimization and its relationship to self-concept and depression among schoolgirls. *Personality and individual differences 16*(1), 183–186.

Nikolaou, D. (2017). Does cyberbullying impact youth suicidal behaviors? *Journal of Health Economics 56*, 30–46.

Olweus, D. (1978). *Aggression in the schools: Bullies and whipping boys.* Hemisphere.

Olweus, D. (1991). Bully/victim problems among schoolchildren: Basic facts and effects of a school based intervention program. *The development and treatment of childhood aggression 17*(17), 411–448.

Polanin, J. R., D. L. Espelage, and T. D. Pigott (2012). A meta-analysis of school-based bullying prevention programs' effects on bystander intervention behavior. *School Psychology Review 41*(1), 47–65.

Ponzo, M. (2013). Does bullying reduce educational achievement? an evaluation using matching estimators. *Journal of Policy Modeling 35*(6), 1057–1078.

Preston, S. D. and F. B. De Waal (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences 25*(1), 1–20.

Qu, X. and J. Guo (2021). Competition in parental investments. *Available at SSRN 3562806*.

Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and violent behavior 15*(2), 112–120.

Sarzosa, M. (2021). Victimization and skill accumulation: The case of school bullying. *Journal of Human Resources*, 0819–10371R2.

Sarzosa, M. and S. Urzúa (2021). Bullying among adolescents: The role of skills. *Quantitative Economics 12*(3), 945–980.

Slee, P. T. (1995). Peer victimization and its relationship to depression among australian primary school students. *Personality and individual differences 18*(1), 57–62.

Swearer, S. M., S. Y. Song, P. T. Cary, J. W. Eagle, and W. T. Mickelson (2001). Psychosocial correlates in bullying and victimization: The relationship between depression, anxiety, and bully/victim status. *Journal of Emotional Abuse 2*(2-3), 95–121.

Sylvia, S., N. Warrinnier, R. Luo, A. Yue, O. Attanasio, A. Medina, and S. Rozelle (2021). From Quantity to Quality: Delivering a Home-Based Parenting Intervention Through China's Family Planning Cadres. *The Economic Journal 131*(635), 1365–1400.

UNESCO (2019). *Behind the numbers: Ending school violence and bullying*. UNICEF.

van der Wal, M. F., C. A. M. de Wit, and R. A. Hirasing (2003, 06). Psychosocial Health Among Young Victims and Offenders of Direct and Indirect Bullying. *Pediatrics 111*(6), 1312–1317.

Villena-Roldan, B. and C. Ríos-Aguilar (2012). Causal effects of maternal time-investment on children's cognitive outcomes. *Center for Applied Economics, Working Paper 285*.

Walker, S. P., S. M. Chang, A. S. Wright, R. Pinto, J. J. Heckman, and S. M. Grantham-McGregor (2021). Cognitive, psychosocial, and behaviour gains at age 31 years from the Jamaica early childhood stimulation trial. *Journal of Child Psychology and Psychiatry*. Forthcoming.

Xia, Y. (2019). What makes a bully? *Available at SSRN 3683897*.

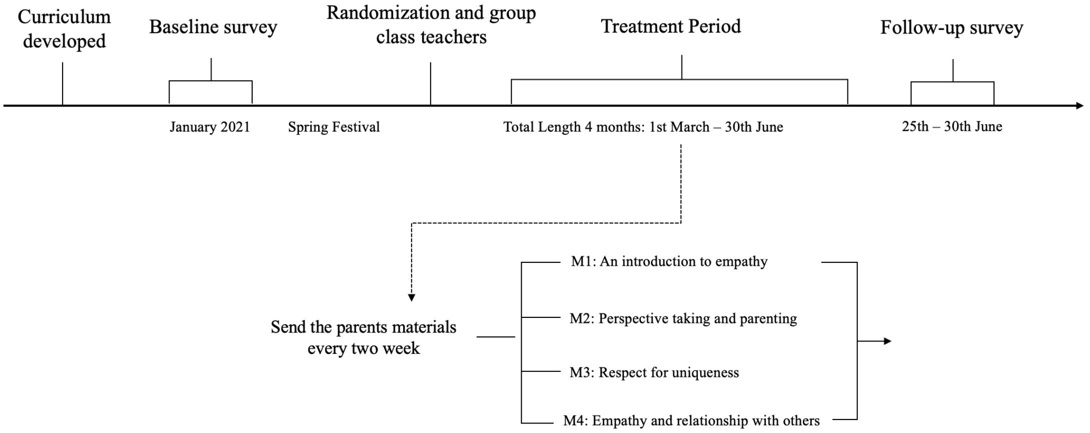
Yang, W., G. Xiong, L. E. Garrido, J. X. Zhang, M.-C. Wang, and C. Wang (2018). Factor structure and criterion validity across the full scale and ten short forms of the ces-d among chinese adolescents. *Psychological Assessment 30*(9), 1186.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics 134*(2), 557–598.

Zaki, J. (2020). Integrating empathy and interpersonal emotion regulation. *Annual Review of Psychology 71*, 517–540.

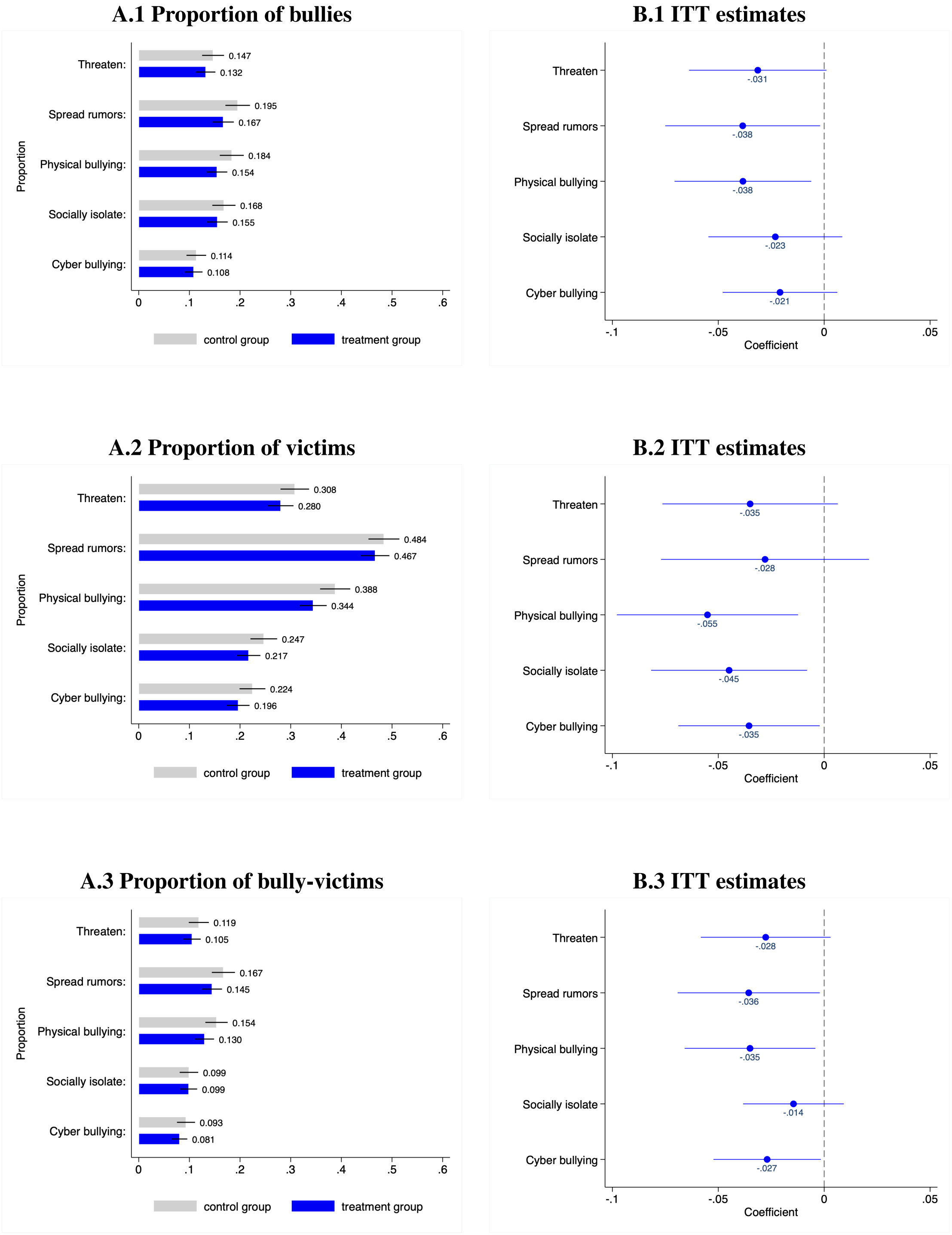
Zumbuehl, M., T. Dohmen, and G. Pfann (2021). Parental involvement and the intergenerational transmission of economic preferences, attitudes and personality traits. *The Economic Journal 131*(638), 2642–2670.

Figure 1: Timeline of the Experiment


Curriculum developed
Baseline survey
Randomization and group class teachers
Treatment Period
Follow-up survey
January 2021
Spring Festival
Total Length 4 months: 1st March – 30th June
25th – 30th June
Send the parents materials every two week
M1: An introduction to empathy
M2: Perspective taking and parenting
M3: Respect for uniqueness
M4: Empathy and relationship with others

Note. The top figure shows the timeline of the experiment. The intervention lasted for 4 months starting from the March 1 until June 30. We list the monthly theme of the program during the treatment period.

Figure 2: Treatment Effects on School Bullying Behavior by Type



A.1 Proportion of bullies

B.1 ITT estimates

A.2 Proportion of victims

B.2 ITT estimates

A.3 Proportion of bully-victims

B.3 ITT estimates

Note. (1) The figures on the left show the fractions of school bullying behaviors, including bullies in Panel A.1, victims in Panel A.2 and bully-victims in Panel A.3, for control and treatment groups at follow-up in detail. We document five detailed types of school bullying: i. threatening other students, ii. spreading rumors, iii. physical bullying (kick), iv. social isolation, and v. cyber bullying. The numbers are calculated from the dummy variable of whether or not one was involved in certain types of bullying behavior in the past semester. (2) The figures on the right show the point estimates and 90% confidence intervals of the program impacts on the bullying behaviors in detail. The estimated effects are ITT estimates based on (1)*. The coefficient estimates are reported in Table D13. Confidence intervals are calculated based on robust standard errors clustered at the class level.*

Table 1: Balance Test

VARIABLES	(1) Mean	(2) Std.dev	(3) Difference T-C	(4) S.E
Panel A. Demographics				
Age	14.463	0.578	-0.018	(0.136)
Male	0.530	0.499	0.008	(0.015)
Urban hukou	0.460	0.499	-0.028	(0.033)
Onlychild	0.298	0.458	0.007	(0.020)
Height in cm	161.885	7.799	0.182	(0.714)
Weight in half kilo	100.974	21.371	-1.171	(1.403)
Panel B. Outcomes at baseline				
Bullying perpetrator	0.379	0.485	-0.000	(0.025)
Bullying victim	0.706	0.456	0.003	(0.023)
Number of friends	3.991	1.371	0.050	(0.130)
Member of exclusive group	0.628	0.483	0.015	(0.026)
Empathy score	48.120	9.636	0.935	(0.619)
Social desirability scale	6.193	2.017	0.128	(0.122)
Self-satisfied	4.461	1.884	0.112	(0.095)
Self-worth	4.776	1.752	0.108	(0.090)
Self-confident	5.000	1.692	0.064	(0.083)
Self-esteem	4.670	1.847	0.049	(0.089)
Perseverance	4.728	1.830	0.017	(0.086)
Total test score (620)	460.553	88.228	1.257	(8.529)
Stress score	13.229	4.149	-0.090	(0.260)
CESD 10-item	8.619	5.686	-0.547	(0.369)
Happiness score	5.067	1.735	0.134	(0.125)
Weekly interaction with parents	10.663	6.884	0.015	(0.769)

Note. (1) This table shows basic regressions attempting to verify the randomization of classroom assignments. Panel A reports demographic variables, while Panel B reports outcome variables at the baseline. (2) Columns 1 and 2 report the summary statistics for the whole sample. Column 3 reports the differences in means for each variable between treatment and control groups. Column 4 reports the standard errors for item-wise regressions using the variables labelled in the first column as the dependent variables. (3) Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table 2: Effects on Parents' Outcomes

	(1) Control Mean	(2) ITT	(3) Permutation p-value	(4) WCB	(5) TOT	(6) Dosage
			Panel A: Time and monetary investment			
Time invest weekday	3.781	0.412**	0.059	0.041	0.970**	0.105**
	(3.303)	(0.190)			(0.446)	(0.049)
Time invest weekend	5.457	0.334	0.145	0.139	0.786	0.085
	(3.711)	(0.211)			(0.495)	(0.054)
Monetary invest	3.414	0.033	0.567	0.571	0.077	0.008
	(1.141)	(0.053)			(0.122)	(0.013)
N	868	1,899			1,899	1,899
			Panel B: Parents' outcomes			
Very willing to invest in empathy education	0.590	0.039	0.140	0.141	0.091	0.010
	(0.492)	(0.024)			(0.058)	(0.006)
Empathy index (short)	-0.035	0.077	0.145	0.138	0.182	0.020
	(1.007)	(0.048)			(0.114)	(0.012)
Democratic parenting	0.790	0.034*	0.079	0.086	0.080*	0.009*
	(0.407)	(0.019)			(0.046)	(0.005)
N	868	1,899			1,899	1,899

Note. (1) This table shows the program impacts on parental investments and parents' outcomes using (1), 2SLS, and the control function approach. (2) Column 1 reports the means and the standard deviations for outcomes of control groups. Column 2 reports ITT estimates and standard errors, while Columns 3 and 4 report the associated permutation P-value after 2,000 stratified clustered resampling and wild cluster bootstrap P-value after 9,999 resampling. In Column 5, we report the TOT estimates using completing at least half of the reading or movie tasks as the "take-up" indicator and random assignment as an IV. In Column 6, we report the dosage-response estimates using the number of completed reading and movie tasks and a control function approach. (3) All regressions control for strata fixed effects. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001)*

Table 3: Effects on Students' Outcomes

	(1) ITT	(2) Permutation p-value	(3) WCB	(4) TOT	(5) Dosage
		Panel A: Empathy			
Empathy index	0.094**	0.027	0.053	0.273**	0.030*
	(0.044)			(0.135)	(0.015)
Empathetic concern	0.097	0.124	0.152	0.244*	0.027*
	(0.060)			(0.141)	(0.016)
Perspective taking	0.086	0.180	0.225	0.219	0.024
	(0.064)			(0.150)	(0.017)
Prosociality	0.160**	0.012	0.019	0.402**	0.045**
	(0.067)			(0.161)	(0.018)
		Panel B: Positive traits			
Positive traits index	0.138***	0.003	0.007	0.339**	0.038***
	(0.048)			(0.115)	(0.013)
		Panel C: Stress and mental health			
Mental health index	0.110**	0.015	0.042	0.273**	0.031**
	(0.051)			(0.120)	(0.014)
N	2,246			2,246	2,246

Note. (1) This table shows the estimated effects on students' empathy, positive traits, and mental health using (1)*, 2SLS, and the control function approach. The empathy index contains empathetic concern, perspective taking, and prosociality. The positive trait index contains self-satisfaction, self-worth, self-confidence, self-esteem, and perseverance. The mental health index contains the pressure score, happiness, and inverse CES-D. Refer to Table D11 for the impacts on the detailed components of each index. (2) Column 1 reports the ITT estimates and standard errors, while Columns 2 and 3 report the associated permutation P-value after 2,000 stratified clustered resampling and wild cluster bootstrap P-value after 9,999 resampling. In Column 4, we report the TOT estimates using those who completed at least half of the reading or movie tasks as the "take-up" indicator and a random assignment as an IV. In Column 5, we report the dosage-response estimates using the number of finished reading and movie tasks and a control function approach. (3) All regressions control for strata fixed effects. The default standard errors clustered at class-level are presented in parentheses (* $p<0.10$, ** $p<0.05$, and *** $p<0.001$).*

Table 4: Program Impacts on Bullying

	(1)	(2)	(3)	(4)	(5)	(6)
	Control Mean	ITT	Permutation p-value	WCB	TOT	Dosage
Bullying involvements:						
Indicator (0/1):						
Bullying perpetration	0.328	-0.053**	0.038	0.044	-0.122**	-0.013*
	(0.470)	(0.025)			(0.059)	(0.007)
Bullying victimization	0.619	-0.044	0.146	0.166	-0.107	-0.012
	(0.486)	(0.029)			(0.070)	(0.008)
Bully-victim	0.295	-0.065**	0.004	0.022	-0.138**	-0.015**
	0.456	(0.025)			(0.058)	(0.007)
Accumulative measure (0-5):						
Bullying perpetration	0.808	-0.147*	0.096	0.085	-0.367*	-0.041*
	(1.462)	(0.083)			(0.195)	(0.022)
Bullying victimization	1.651	-0.194**	0.044	0.054	-0.483**	-0.054**
	(1.726)	(0.092)			(0.217)	(0.024)
Spectators:						
Witnessed bullying incidents	0.420	-0.061*	0.092	0.094	-0.149*	-0.016*
	(0.494)	(0.034)			(0.079)	(0.009)
Willing to help bullying victims	0.767	0.052**	0.014	0.009	0.129**	0.014**
	(0.423)	(0.020)			(0.049)	(0.005)
N	1,029	2,246			2,246	2,246

Note. (1) This table shows the program impacts on school bullying using (1), 2SLS, and a control function approach. (2) Column 1 reports the means and the standard deviations for the corresponding outcome variables for students in control groups. Column 2 reports the ITT estimates and standard errors, while Columns 3 and 4 report the associated permutation P-value after 2,000 stratified clustered resampling and wild cluster bootstrap P-value after 9,999 resampling. In Column 5, we report the TOT estimates using completing at least half of the reading or movie tasks as the "take-up" indicator and random assignment as an IV. In Column 6, we report the dosage-response estimates using the number of finished reading and movie tasks and a control function approach. (3) All regressions control for strata fixed effects. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

APPENDIX

A Detailed Items in the Questionnaire

A.1 Students' Outcome Measures

We construct measures on school bullying behaviors. We also collect multifaceted social and emotional skills. Apart from empathy, we further construct measures of students' mental health status, stress resistance, positive personality traits (positive self-image and commitment), and friendships.

- 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. In addition, they were asked whether they ever witnessed school bullying incidents in the follow-up survey. We also asked them whether they would help those bullying victims when they witness a bullying incident.

- Empathy: To avoid a long questionnaire, we use 9-item empathy measurement to explore two dimensions on empathetic concerns and perspective taking, which is also used in Alan et al. (2021). For most items, we use a 7-point Likert scale for both baseline and follow-up surveys, with scores ranging from from completely disagree (1) to completely agree (7). We added another dimension, prosociality, in the follow-up survey. The questions include the hypothetic scenarios about helping other children in difficulties, doing others a favor, helping their mother do housework, becoming a charitable person and rescuing a drowning child adapted from the official guide from the Centers for Disease Control (CDC) (Dahlberg et al., 2005). For each scenario, we asked students whether they have ever imagined the scenarios and ask them to choose (1) never, (2) sometimes, or (3) very frequently.

- Mental health: Mental health is measured using the 10-item Center for the Epidemiological Studies of Depression Short Form, or CES-D-10, which is a 10-item Likert scale questionnaire (Yang et al., 2018).[1] 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 the happiness score using a scale of 1-7, with 7 being the happiest.

- Stress score: We elicit students' stress by four categories of sources: (i) studies at school, (ii) peer relationships, (iii) rank/test scores in the class, and (iv) family background. For each item, we use a 7-point Likert scale for both baseline and follow-up surveys, with scores

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

ranging from the least stressed (1) to the most stressed (7). We then construct the *inverse stress index* consisting of these four components so that a higher score indicates feeling less stressed.

- Positive self-image: 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), and (iv) I am not worse than others and am proud of myself (self-esteem). For each item, we use a 7-point Likert scale for both baseline and follow-up surveys, with scores ranging 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.

- Perseverance: We ask the students whether they agree or disagree with the following 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, with scores ranging from completely disagree (1) to completely agree (7).

- Friendship: We construct the number of friends by counting the number of good friends the students reported (maximum 5 best friends). We also asked the interaction intensity with each friend they nominate (scale 0-5 with 0 if there is no friend nominated and 5 being interacted with the most). We then sum up the intensity into a measure of their total intensity score with friends (scale 0-25). To evaluate their attitudes toward diversity, we also ask the students whether they view themselves as being able to make friends with anyone, regardless of test scores, family background, etc. The scale is a 1-5 scale, with 1 being completely disagree and 5 being completely agree.

- Social desirability bias index: We construct a standardized PCA index using the binary responses to the modified children's version of the Crowne-Marlowe module used at baseline to measure social desirability bias (Miller et al., 2015; Crowne and Marlowe, 1960). The questions we used include the following. (i) Have you ever felt like saying unkind things to a person? (ii) Do you sometimes feel like staying home from school even if you are not sick? (iii) Do you sometimes feel angry when you do not get your way? (iv) Are there some times when you do not like to do what your parents tell you? (v) Do you sometimes get mad when people do not do what you want them to do? (vi) Are you always polite, even to people who are not very nice? (vii) Sometimes, do you do things you have been told not to do? (viii) Do you always listen to your parents?

- Prosociality index: Improvement in empathy skills may lead to more take-up of students' prosocial behaviors. In the questionnaire, we measure prosociality by asking students to make choices about their behaviors in different hypothetical situations. First, we elicit students' willingness to pay for prosociality using the question from the Chinese version of the Global Preference Survey (Falk et al., 2018).[2] The question asks about the following hypothetical scenario. You were helped by a stranger when you got lost, and the stranger spent 16 RMB to help you get to your destination. You have six gifts in hand with different values ranging from 4 RMB to 24 RMB. Student were asked to choose which gift they would give the stranger. They could choose one of the six gifts as well as another option—giving nothing. The variable "return favor" is a dummy variable and was constructed by whether the gift the student would choose to give is more than 16 RMB. Second, we ask two hypothetical questions about donation behaviors to measure prosociality. One of them asks about whether the student would be willing to donate old clothes to poor kids in the western area of China. The dummy variable asks whether the student would be willing to donate his or her favorite cellphone to the really needy, i.e., the left-behind children in the poor western area of China. Both are dummy variables.

- 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 have parents involved in a typical week in the previous semester. The activities include having dinner, taking/discussing about school lives, watching TV, checking homework, and playing outdoor activities.

A.2 Parents' Outcome Measures

To understand the potential mechanisms, we construct various outcome measures for parents. To compliment students' self-reported measures on empathy education and time spent with parents, we also collected data on parents' time use and the take-up of parent-child activities reported by parents.

- Participation in empathy-related activities: The variables measure the take-up of the program. We asked parents whether they watched movies or read short articles on empathy with their kids at least once or at least once per month in the past semester.

- Parents' empathy and prosocial motives: The empathy measure is constructed following the same method as constructing students' outcome measure.

- Parents' mental health status: Following the Chinese Education Panel Study adult survey, we use the 12-item General Health Question (GHQ12) to elicit mental health (Chan, 1985).

- Parenting style: We provide detailed explanations of the four types of parenting styles—authoritative, authoritarian, permissive, and neglecting—and ask the parents to select the type that is the most applicable to them.

- Parental time investment: Time spent (hours) on average on parent-child activities per day including reading, checking homework, playing and conducting general education activities with kids on weekdays and weekends over the past week.

[2] For details, one can refer to https://www.briq-institute.org/global-preferences/home.

- Parental monetary investment: We are also interested in whether the increase of time investment may crowd out monetary investment or change parents' attitudes toward monetary investment. It is insensitive to directly inquire about their monetary investment in their children. Instead, we only ask them about the investment in monthly education-related activities as a proportion of their total income. We divide it into five categories: 5% or less, 5-10%, 10-25%, 25-50% and more than 50% of total income.
- Parental attitudes toward monetary investment after-school tutoring: Parents are asked to select whether they would send their kids to after-school tutoring in three hypothetical settings: (i) when their best friends' children were sent to after-school tutoring, (ii) when the best students in the class were sent to after-school tutoring, and (iii) when most of the students in the class were sent to after-school tutoring. We also elicit the perceived value of cram schools by asking parents to score whether the after-school tutoring is good for students' test scores and whether the after-school tutoring is good for students' mental health for a hypothetically struggling student, on a scale of 1100.
- Relationship with kids: Parents are asked to rank the tension with their child in a 5-point Likert scale question, with 5 being the worst parent-child relationship.

B Additional Analysis

B.1 Spillover Effects

To estimate the spillover effects of taking up the program within the class, one needs to assume that the actual take-up rate of each class is orthogonal to determinants of each students' bullying behaviors. This assumption is not valid since the actual take-up rate of each class is correlated with other cofounders. However, the conditional independence assumption is more credible with a set of rich baseline outcome variables and class level variables. In short, we estimate the linear-in-mean equation as follows:

$$Y_{ic2} = \alpha + \beta\pi_c + \gamma X_{ic} + \epsilon_{ic}, \tag{2}$$

where π_c is the take-up rate of class c; X_{ic} is a vector of individual demographics and baseline outcome which is identical to the specifications are similar to (1). The take-up rate is often correlated with other class level characteristics, which leads to biased estimates. To reduce the bias, we instrument the π_c using the treatment assignment.

The results are reported in Table D24. Panel A studies all samples and Panel B studies the non-take-up sample. The simple linear-in-mean regression reports that the class-level take-up rate is negatively correlated with individual bullying behaviors in the class. However, 2SLS suggests that spillover effects may exist but that the size of these effects is not large enough to be significant.

B.2 Detailed Monetary Investments in Afterschool Tutoring

To test the crowding-out effect of time investments, we report the effects on monetary investments and parents' beliefs about the main education expenditure - afterschool tutoring in Panel B in Table D9. In China today, more than 60% of parents spend zero hours accompanying their kids on a typical weekday, while over 90% of them send their children to attend afterschool tutoring classes, the so-called "cram schools."[3] From the baseline surveys that we collected, about 27% in the sample point out that they send their kids to afterschool programs because other children's parents also send their kids to these classes, they have a fear of their children lagging behind other students given the fierceness of competition at school, and the mentality of feeling obliged to do "whatever everyone else is doing." Furthermore, 11% point out that they send their kids to these schools since they do not have time to accompany their children. Thus, we would like to test whether our directed parental involvement program may affect parents' attitudes toward private tutoring investment decisions. The monetary investment is measured as a percentage of total income. From the estimates shown in Column (3), it appears that the program did not crowd-out the monetary investment and that there was no effect on parents' beliefs in cram schools. The decision to send children to cram schools is larger in scenarios 2 and 3; in scenario 2, the hypothetical case is that the best student in the class takes private tutoring classes, and in scenario 3, we suppose that most students in the class seek extra tutoring.

C Detailed Illustration of Methods

C.1 Control Function Approach and Dosage Effect

In the TOT-dosage analysis, we use the control function approach to fully explore the number of parent-child activities finished as recorded by the platform. For each of the outcome variables, we instrument the number of reading and movie activities with the treatment assignment at the first stage:

$$N_{ict} = \alpha_1 + \beta_1 T_c + \gamma_1 Y_{ic(t-1)} + \tau_s + \xi_{ic},$$

where N_{ict} is the number of completed reading and movie activities for child i in class c at follow-up; T_c is the class-level treatment indicator, $Y_{ic(t-1)}$ is an outcome measure for child i in class c at baseline, and τ_s is a set of strata fixed effects. We again adjust standard errors for clustering at the class level using the Liang Zeger estimator. For the second stage, we add the first-stage predicted residuals $\hat{\xi}_{ic}$:

$$Y_{ict} = \alpha_2 + \beta_2 N_{ict} + \beta_3 \hat{\xi}_{ic} + \gamma_2 Y_{ic(t-1)} + \tau_s + \eta_{ic}$$

where Y_{ict} is an outcome measure for child i in class c at follow-up and $\hat{\xi}_{ic}$ the estimated residual of the first-stage equation. We adjust standard errors for clustering at the class level using the Liang-Zeger estimator.

[3] http://www.chinadaily.com.cn/a/201806/15/WS5b2300a5a310010f8f59d147_1.html.

C.2 GRF and Effect Heterogeneity

Here, we introduce the method to study heterogeneous effects in detail. The first step is to use the GRF method to select which baseline characteristics predict differences in treatment effects of the program. The gist of the GRF method relies on the concept of the conditional average treatment effects (CATE) for different subgroups of the population. Specifically, it is defined as follows:

$$\tau(X) = E[Y(T=1) - Y(T=0) \mid X = x],$$

where Y is the outcome variable, T is the treatment indicator, and X is the observable covariate. We select in total 24 baseline characteristics for the prediction stage.[4] After training the GRF algorithm, we mainly focus on four baseline characteristics: empathy skills, age, parental involvement, and pressure score. In Table D25, we list the corresponding importance rank of each variable predicted by the GRF algorithm. The numbers are obtained based on the percentage of importance each observable characteristic has in the forest in terms of the frequency with which the variable is used as a splitting variable in the forest. The higher the rank is, the better the variable in predicting treatment heterogeneity. Following Sylvia et al. (2021), in Figures D6-D10, we also plot the estimated out-of-bag CATEs from the GRF estimation along the distribution of these four characteristics as a motivation of our heterogeneity analysis.[56] There is indeed much heterogeneity along the distribution of observable characteristics, as shown in the figures. Although they lack a clear pattern for all the characteristics, we do find that the treatment effects on four out of the five outcome variables tend to be higher for lower parental involvement at the baseline. Motivated by the algorithm prediction results, we proceed to conduct the traditional heterogenous treatment effect analysis along the four dimensions.

To capture heterogeneity, for each of the four dimensions, we create a dummy variable indicating whether the children were below a certain threshold in the baseline distribution. We define the threshold for each dimension based on how the estimated out-of-bag CATEs from the GRF analysis vary across the baseline distribution of each variable. Since we have multiple outcome variables to check, we report the results by constructing the threshold value following the estimated out-of-bag CATEs for empathy score, as shown in Figure D10.[7] For all the baseline characteristics such as parental involvement, and pressure score, we define an indicator for being in the first quartile

[4] The characteristics include demographic characteristics, such as age, gender, hukou status, and an indicator of being an only child; socioemotional characteristics, such as pressure score, CES-D score, being depressed or not, happiness score, rank pressure score, college aspiration, confidence level, and whether one feels lonely in childhood; and parental and household characteristics, such as an indicator of close relationship with the father, indicator of close relationship with the mother, whether parents have a say in making friends with classmates, pocket money per week, being brought up by their mother before the age of 6 years, and the intensity of parental involvement. We also include baseline outcome variables.

[5] The out-of-bag prediction refers to the estimated CATE's only considering trees for which the observation is not used as part of the training set (Sylvia et al., 2021).

[6] As stated in Sylvia et al. (2021), the plots need to be interpreted with caution as they are not informative for causal inference but are just a way of visualizing the estimated out-of-bag CATEs of the GRF algorithm.

[7] We also check the results by varying the threshold definition based on the estimated out-of-bag CATEs of the other four outcome variables and the pattern of the results remain unchanged.

of the pre-intervention distribution to capture the potential nonlinearity effects. As suggested in Figure D10, we define an indicator for being in the third quarter of the pre-intervention distribution of the empathy score and an indicator for being below the median of the age distribution. Using the new indicator variables, we estimate the ITT effects of the intervention using OLS regression with the following specification:

$$Y_{ict} = \alpha_1 + \beta_1 T_c + \beta_2 T_c Q_{ic(t-1)} + \beta_3 Q_{ic(t-1)} + \tau_s + \epsilon_{ic}, \tag{3}$$

where Y_{ict} is an outcome measure for student i in class c at follow-up, T_c is a dummy variable indicating the treatment assignment of class c, $Q_{ic(t-1)}$ is the relevant indicator defined using the baseline characteristic of interest, $T_c Q_{ic(t-1)}$ is the interaction of treatment assignment with the baseline characteristic indicator, and τ_s is a set of strata fixed effects. We adjust standard errors for clustering at the class level using the Liang-Zeger estimator.

C.3 Misclassification Error in Self-reported Bullying Behaviors

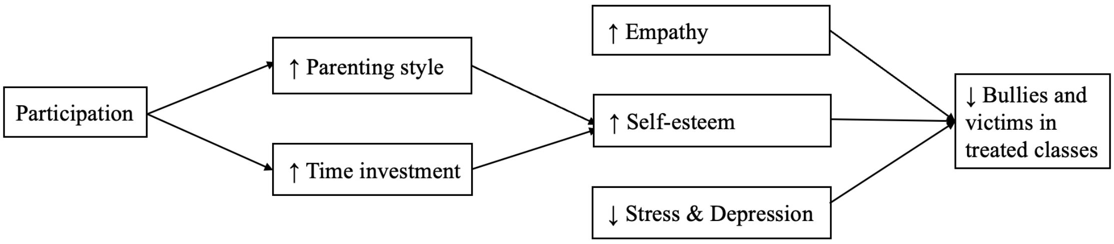
We follow the strategy adopted in studying sexual behavior context by De Paula et al. (2014) and Delavande and Kohler (2016) to correct for misclassification error. We assume that students truthfully report when they do not engage in any bullying incidents and that there is a constant probability α_1 of misreporting bullying behaviors when engaging in bullying. Probability α_1 is estimated together with the other coefficients of the model. Specifically, we estimate the model with maximum likelihood estimation and find the coefficients to minimize the objective function:

$$L(\alpha_1, b) = \frac{1}{n} \sum_{i=1}^{n} y_i \ln\left((1-\alpha_1)\Theta\left(x_i' b\right) + (1-y_i)\ln\left(\alpha_1 \Theta\left(x_i' b\right)\right)\right),$$

which is a modification of the maximum likelihood of a probit with misreporting probability α_1. We conduct a separate analysis for the three main binary indicators of bullying behaviors: (i) bullying perpetrator, (ii) bullying victim, and (iii) bully-victims. We include the same set of controls as those in Column 2 of Table D22. Standard errors are always clustered at the classroom level. When estimating the model, in Table D18, we additionally control for baseline social desirability scale. In Tables D19 and D20, we vary the values of α_1 and report the predicted treatment effects.

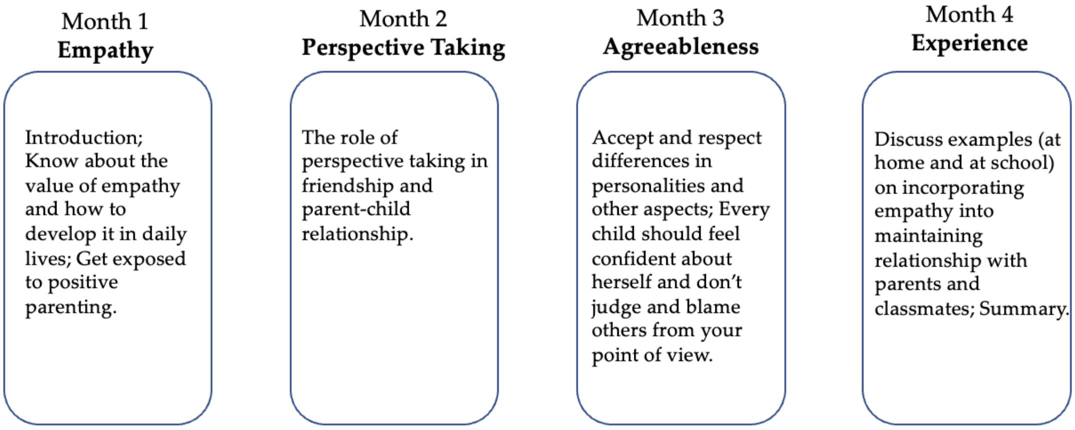
D Tables and Figures

Figure D1: How the Program Works



Note. The figure illustrates the theoretic framework, which guides us in implementing and evaluating this parental involvement program.

Figure D2: Program Theme


Month 1 **Empathy**

Introduction; Know about the value of empathy and how to develop it in daily lives; Get exposed to positive parenting.

Month 2 **Perspective Taking**

The role of perspective taking in friendship and parent-child relationship.

Month 3 **Agreeableness**

Accept and respect differences in personalities and other aspects; Every child should feel confident about herself and don’t judge and blame others from your point of view.

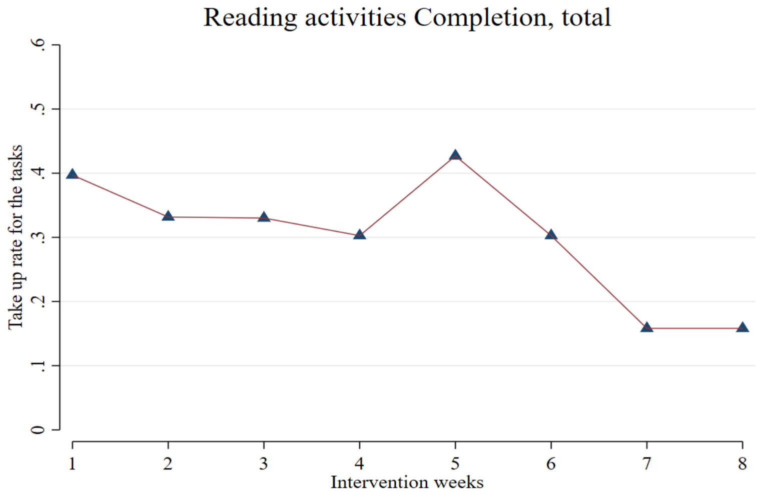
Month 4 **Experience**

Discuss examples (at home and at school) on incorporating empathy into maintaining relationship with parents and classmates; Summary.

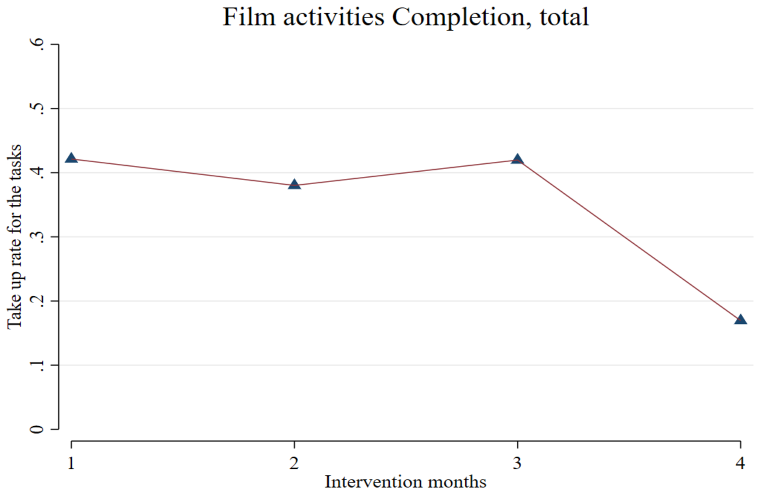
Note. The figure details the theme of our 4-month parental involvement program. We rely on two books, one by A. Ciaramicoli and one by K. Ketcham, as references for these themes and some research papers to relate school bullying and empathy for the last theme. We expand the details of the program in Table D3.

Figure D3: Task Completion Rate by Task

Panel A. Reading



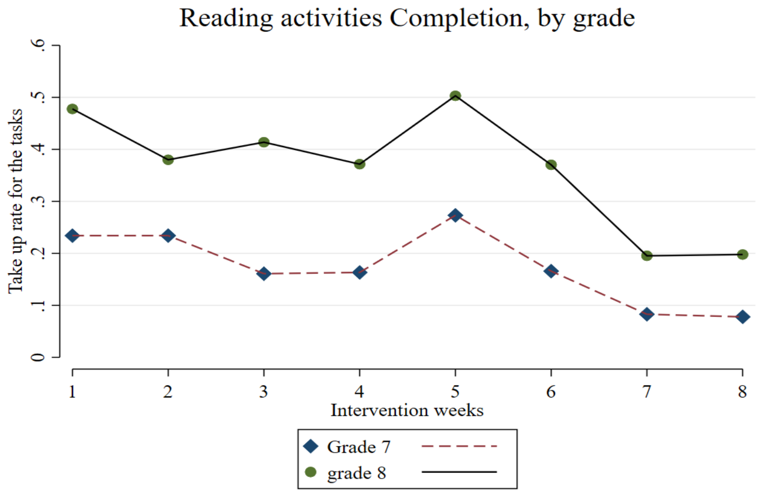
Panel B. Movie



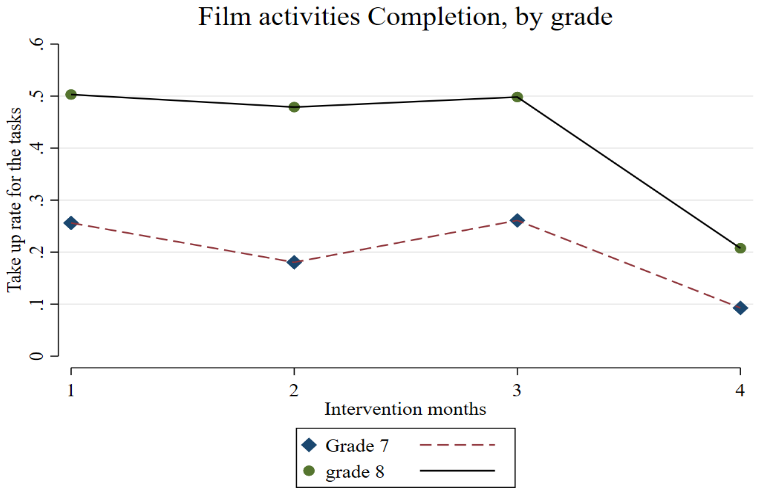
Note: Panel A shows the completion rate of the biweekly reading activities. Panel B shows the completion rate of the monthly movie activities. The numbers are calculated by the total number of those who completed the specific task divided by the total number of those who registered to participate in the program. The total number of registered parents is 872, and the registration rate is 71%.

Figure D4: Task Completion Rate by Task and Grade Level

Panel A. Reading

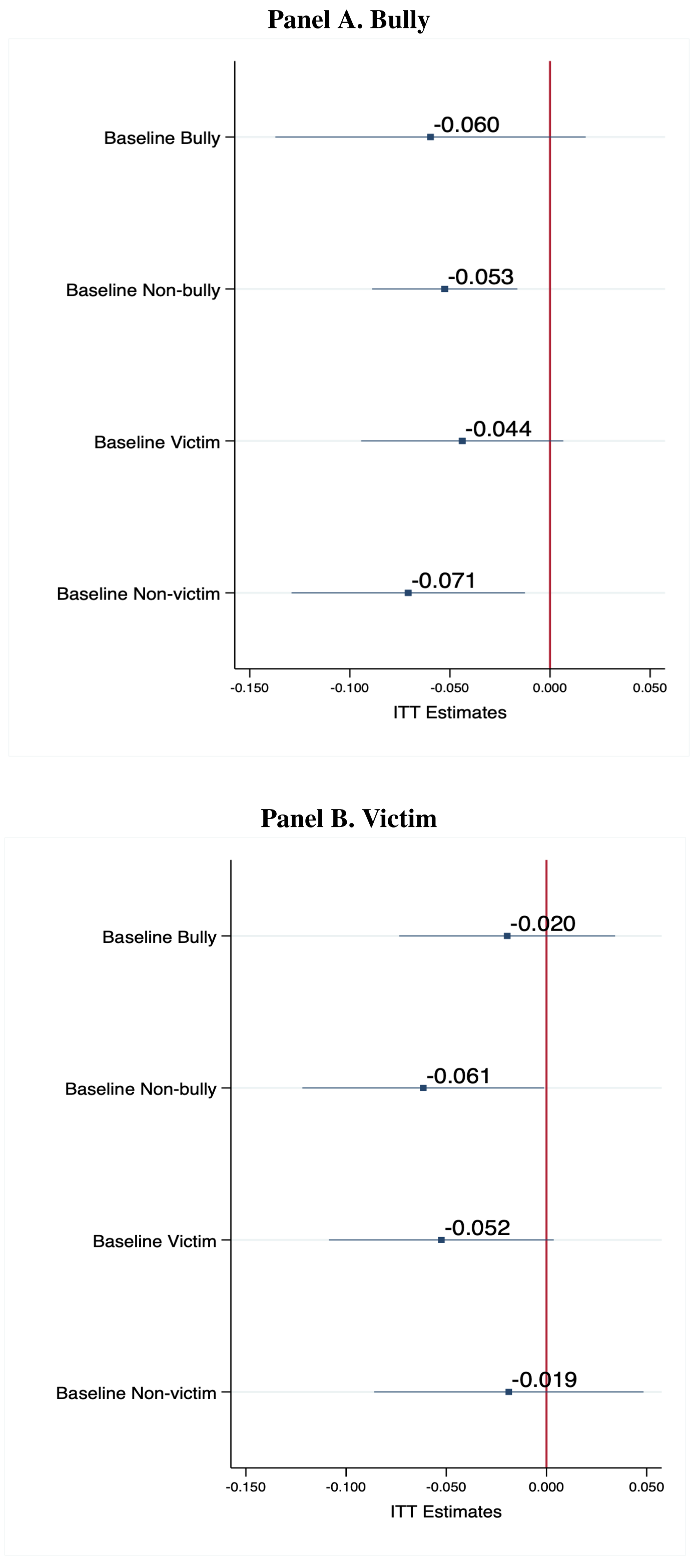


Panel B. Movie



Note: Panel A shows the completion rate of the biweekly reading activities by grade level. Panel B shows the completion rate of the monthly movie activities by grade level. The numbers are calculated by the total number of those who completed the specific task within the grade cohort divided by the total number of those who registered to participate in the program within the grade cohort. The total number of registered parents is 872, and the registration rate is 72%.

Figure D5: Effects on Bullying by Baseline Bully Status

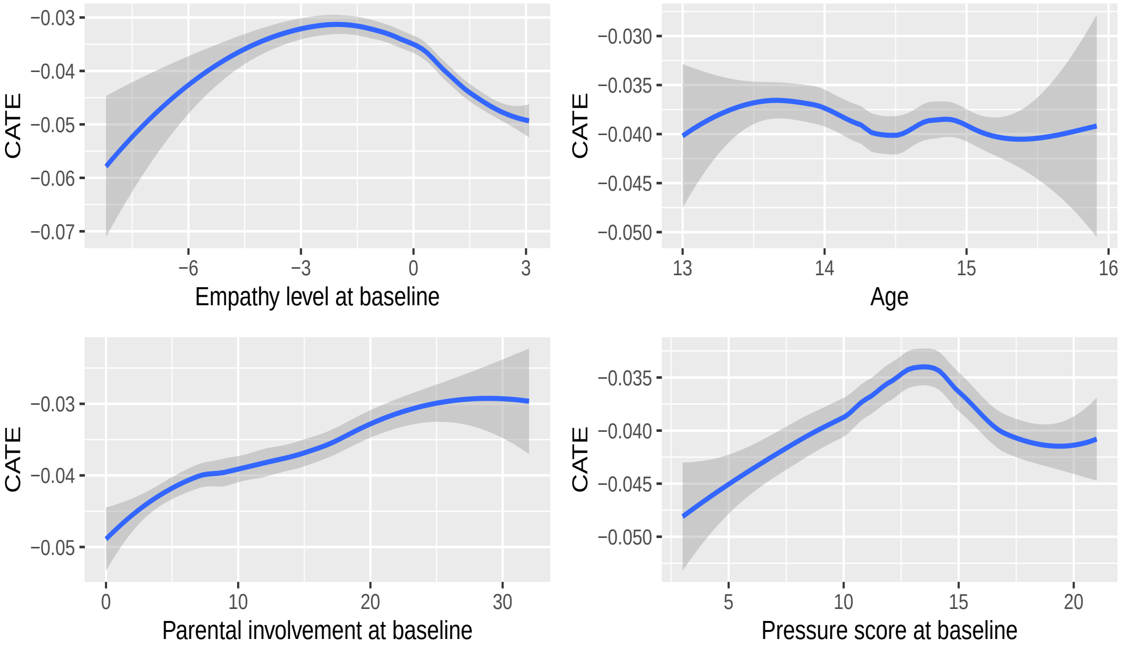


Panel A. Bully

Panel B. Victim

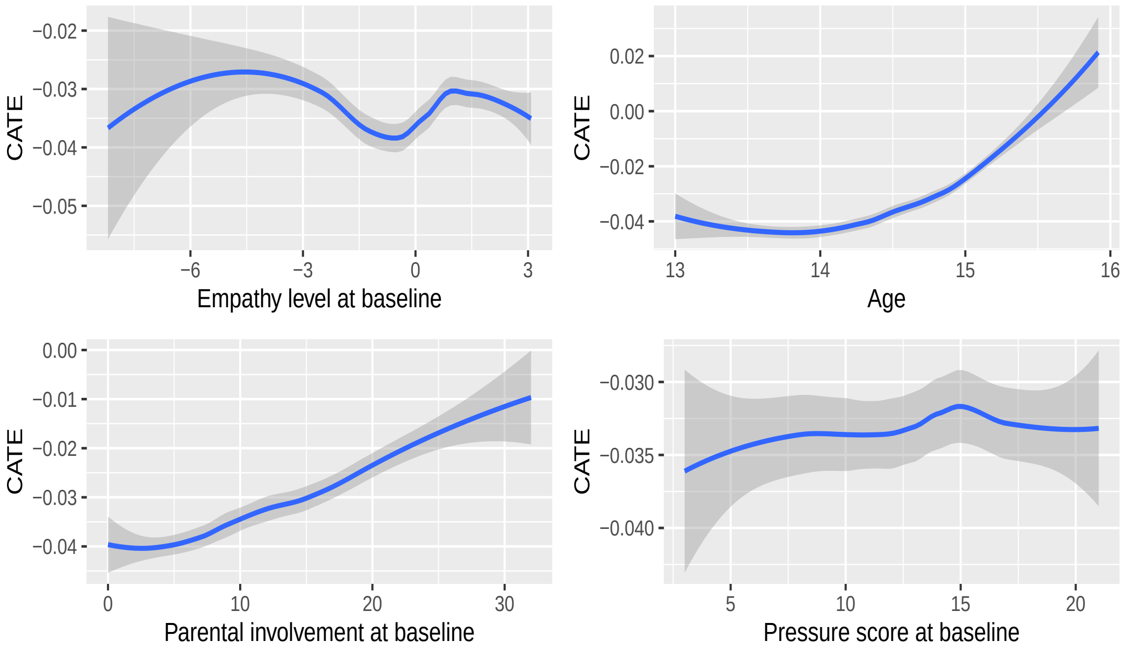
Note: Panel A shows the point estimates and 90% confidence intervals of the program impacts on bullying involvement by baseline bully status, including being a bully, a nonbully, a victim, and a nonvictim. Panel B shows the point estimates and 90% confidence intervals of the program impacts on being bullied by the same four different baseline bully categories. The estimated effects are ITT estimates based on (1). Confidence intervals are calculated based on robust standard errors clustered at the class level.

Figure D6: Out-of-Bag CATE Estimates for Bullying Perpetrator from GRF-Trained Algorithm along Observable Characteristics



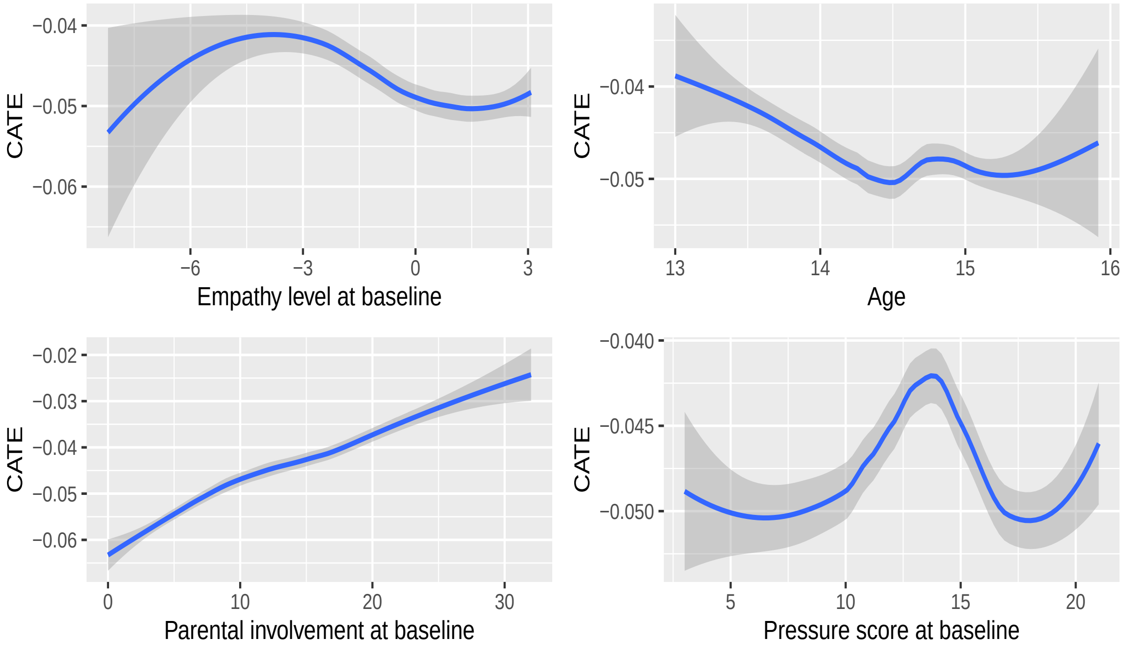
Note: This figure shows the out-of-bag CATE estimates for the bullying perpetrator indicator from GRF-trained algorithm along the four baseline characteristics. In the case of out-of-bag prediction, the estimated CATEs only consider trees for which the observation is not used as part of the training set.

Figure D7: Out-of-Bag CATE Estimates for Bullying Victim from GRF-Trained Algorithm along Observable Characteristics



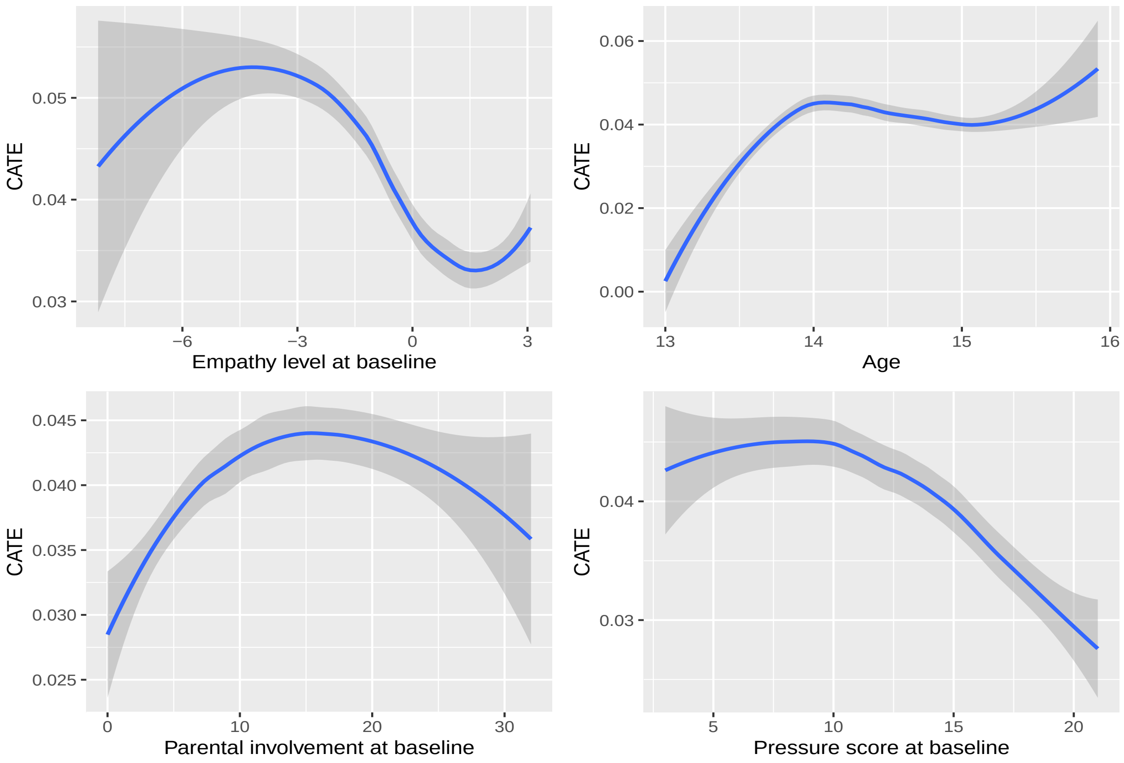
Note: This figure shows the out-of-bag CATE estimates for bullying victim indicator from the GFR-trained algorithm along the four baseline characteristics. In the case of out-of-bag prediction, the estimated CATEs only consider trees for which the observation is not used as part of the training set.

Figure D8: Out-of-Bag CATE Estimates for Bullying Perpetrators and Victims from GRF-Trained Algorithm along Observable Characteristics



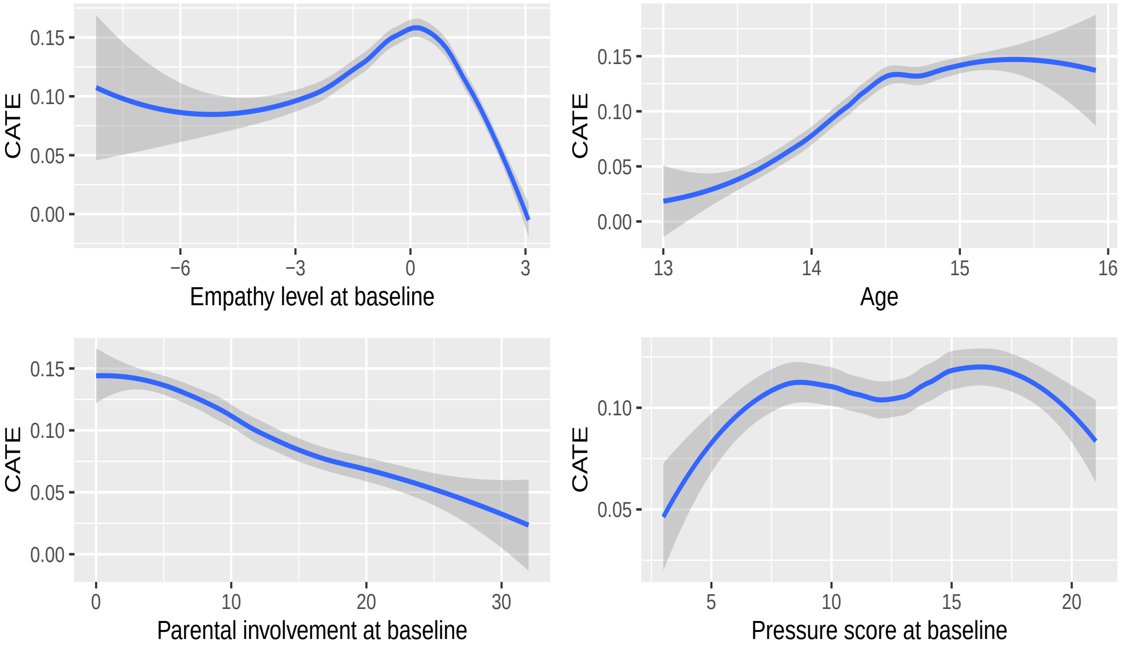
Note: This figure shows the out-of-bag CATE estimates for the bully-victim indicator from the GFR-trained algorithm along the four baseline characteristics. In the case of out-of-bag predictions, the estimated CATEs only consider trees for which the observation is not used as part of the training set.

Figure D9: Out-of-Bag CATE Estimates for Those Willing to Help Victims from GRF-Trained Algorithm along Observable Characteristics



Note: This figure shows the out-of-bag CATE estimates for bystanders from the GFR-trained algorithm along the four baseline characteristics. In the case of out-of-bag predictions, the estimated CATEs only consider trees for which the observation is not used as part of the training set.

Figure D10: Out-of-Bag CATE Estimates for the Empathy Skill from GRF-Trained Algorithm along Observable Characteristics



Note: This figure shows the out-of-bag CATE estimates for students' empathy skill from the GFR-trained algorithm along the four baseline characteristics. In the case of out-of-bag predictions, the estimated CATEs only consider trees for which the observation is not used as part of the training set.

Table D1: Comparison of 7th and 8th Graders in Schools Located in Cities, Counties and Villages

	Counties or suburban		Central area of the city		Towns and rural area	
	(1)	(2)	(3)	(4)	(5)	(6)
	Mean	Std dev	Difference	S.E	Difference	S.E
Panel A: Student demographics						
Male	0.524	0.499	-0.021**	(0.010)	-0.013**	(0.005)
Age	14.073	1.359	-0.148	(0.138)	0.087	(0.070)
Height	161.637	8.583	1.368**	(0.551)	-0.766***	(0.288)
Weight	99.849	22.327	3.056**	(1.214)	-2.009***	(0.609)
Onlychild	0.411	0.492	0.222***	(0.033)	-0.085***	(0.015)
Urban hukou	0.418	0.493	0.289***	(0.029)	-0.094***	(0.013)
College aspiration	0.644	0.479	0.084***	(0.021)	-0.063***	(0.010)
Stressed about parents' expectation	3.045	1.096	-0.076**	(0.034)	0.033*	(0.017)
Cognitive score (standardized)	-0.024	0.969	0.324***	(0.073)	-0.137***	(0.034)
Grit score	0.046	0.961	-0.076*	(0.039)	-0.023	(0.019)
Depression score	-0.043	0.994	0.048	(0.039)	0.033*	(0.018)
Boarding school	0.332	0.471	-0.258***	(0.043)	0.120***	(0.025)
Left behind children	0.227	0.419	-0.051***	(0.018)	0.034***	(0.011)
Panel B: Maternal education and parenting						
Mother at least high school graduate	0.224	0.417	0.252***	(0.023)	-0.039***	(0.009)
Parental time investment index	0.049	1.016	0.196***	(0.064)	-0.172***	(0.033)
Parenting style harsh	0.073	0.988	-0.034	(0.036)	-0.082***	(0.018)

Note. This table compares the differences among students in schools located in cities, counties and villages who are in the 7th and 8th grades. We use the nationally representative data from the China Education Panel Study (CEPS) 2013 wave and follow the same definition of school location type from the survey. We compare students' characteristics by location of the schools in Panel A. Being stressed from parents' expectations is a measure ranging from 1 to 5, with 5 being the most stressed. Cognitive scores are directly obtained from CEPS. The CEPS conducts standardized cognitive ability tests for students in each grade. The grit score is the standardized PCA index from the survey. The depression score is also a standardized PCA index. We also compare whether the school is a boarding school and the proportion of left-behind children in the school. We compare maternal educational attainment and parenting style in Panel B. We study the differences in terms of the proportion of mothers being at least high school graduates, the standardized PCA index of parental time investment, and the standardized PCA index of parenting style being very harsh. Columns 1 and 2 report the summary statistics for those in schools located in counties in suburban areas. Columns 3 and 4 report the differences and standard errors between city students and county students for each variable $X_{city} - X_{county}$*, and Columns 5 and 6 report the same statistics for each variable* $X_{rural} - X_{county}$*. * p<0.10, ** p<0.05, and *** p<0.01. The numbers in parentheses are robust and standard errors for the differences clustered at the class level.*

Table D2: School Bullying by Type at Baseline

	Extensive margin		Intensive margin	
Panel A. Bullying perpetration				
	(1) Mean	(2) Std.dev	(3) Mean	(4) Std.dev
Threatening	0.131	(0.337)	1.224	(0.660)
Spreading rumors	0.166	(0.372)	1.310	(0.799)
Physical bullying	0.175	(0.380)	1.309	(0.764)
Cyber bullying	0.130	(0.336)	1.228	(0.684)
Isolating	0.083	(0.276)	1.150	(0.578)
N	2,246			
Panel B. Bullying victimization				
Threatening	0.337	(0.473)	1.731	(1.182)
Spreading rumors	0.526	(0.499)	2.256	(1.402)
Physical bullying	0.449	(0.498)	2.014	(1.302)
Cyber bullying	0.247	(0.431)	1.527	(1.051)
Isolating	0.183	(0.387)	1.392	(0.937)
N	2,246			

Note. (1) This table shows the distribution of bullying behaviors by detailed type. Panel A reports distributions for bullying perpetration. Panel B reports the distributions for bullying victimization. (2) Columns 1 and 2 report the mean and standard deviation for the extensive margin of bullying behaviors, defined as being involved in the type of behavior. Columns 3 and 4 report the mean and standard deviation for the intensive margin of bullying behaviors, defined as the frequency of involvement in the type of behavior in the past semester.

Table D3: Intervention Content

Time	Tasks
Week 1 (M1)	**Reading task**: read a short article on (i) What is empathy? and (ii) The importance and value of empathy? **Movie of the 1st month**: watch “Looking Up?” together, then discuss about the parenting styles in the movie with your child
Week 3	**Cases and examples**: read a short article on (i) Parents incorporate empathy into parenting styles and (ii) Positive parenting skills
Week 5 (M2)	**Reading task**: read a short article on (i) Perspective taking and (ii) The importance of perspective taking on friendship and parent-child relationship **Movie of the 2nd month**: watch “Wonder” together, then think about the script “Auggie can’t change the way he looks. But maybe we can change how we look at him.”
Week 7	**Cases and examples**: read a short article on (i) Self-centeredness and (ii) How to become less self-centered
Week 9 (M3)	**Reading task**: read a short article on (i) Personality and multiple intelligences and (ii) The importance of being unique and respect each other **Movie of the 3rd month**: watch “Taare Zameen Par” together, then focus on discussing the script ”Every child is like a shining star, we should discover the uniqueness of each child from different perspectives.” with your child
Week 11	**Cases and examples**: read a short article on (i) How to educate your child according to their individual uniqueness and (ii) How to teach your child to embrace others’ uniqueness, especially when they look different.
Week 13 (M4)	**Reading task**: read a short article on (i) Lack of empathy and peer relationship and (ii) How parents can help children to get through poor peer relationship **Movie of the 4th month**: watch “Better Days” together, then focus on discussing the script “why we can’t learn sympathy until becoming an adult?” with your child
Week 15	**Cases and examples**: read a short article on (i) Emotional skills help students improve peer relationship and (ii) Lack of empathy fosters cold and distant relationships with peers, creating more adverse consequences.

Note. (1) This table shows the detailed content of the intervention. Parents are encouraged to discuss and exchange views on the task content with their children and submit a short reflection essay to the platform once they finish the task. (2) The first column shows the times when the tasks were sent. Each task was delivered via the WeChat *group of treated classes by the class teacher at 7:30pm on Friday. (3) The second column summarizes the main components of each task. (3) The short articles of the biweekly reading tasks were uploaded to the platform that we created one day prior to the delivery date. The estimated time for the reading task is about 30-45 minutes. (4) The monthly activities (watching movies) were assigned and announced in the platform at the beginning of each month (7:30pm on the first Friday each month). The estimated time for the movie task is about 90-120 minutes.*

Table D4: Attrition in Parent Survey

	(1) Control	(2) Treatment	(3) Difference	(4) Total number
Number of classes	22	26	4	48
Number of students completed survey	1,029	1,217	188	2,246
Number of parents completed survey	868	1,031	163	1,899
Number of attrition	161	186	25	347
Attrition rate	0.156 (0.365)	0.153 (0.363)	0.004 (0.023)	

Note. (1) Columns 1 and 2 show the completion rate and attrition rates at follow-up in the control and treated groups. (2) Column 3 shows the difference of the completion and attrition rates between the control and treatment groups. (3) Column 4 is the sum of the first and second columns. (4) The numbers in parentheses are robust standard errors for the difference of the attrition rates in Column 3, and the standard deviation for the attrition rate in the control and treated groups in Columns 1 and 2.

Table D5: Students' Skill Measurements

	(1)	(2)
	Cognitive	Noncognitive
Standardized Test Scores		
	Math	
	Language	
Empathy Measure		Perspective taking
		Empathetic concern
		Prosocial fantasy
Mental Health and Stress		CES-D10
		Study life at school
		Peer relationships
		Rank/test scores in the class
		Family background
Positive Personality (1-item)		Self-satisfied
		Self-worth
		Self-confident
		Self-esteem
		Perseverance

Note. (1) This table shows the detailed content of the measurements of students' abilities, including cognitive and noncognitive skills. In total, we measure test scores, empathy, and mental health, as well as positive personality traits.

Table D6: Parents' Input and Skill Measurements

	(1) Investment	(2) Skills
Time Investment in Hours (weekday and weekend)	Read books Help homework Play and Leisure Caring and talk	
Monetary Investment (categorical variable)	5%- 5-10% 10-25% 25-50% 50%+	
Parenting Style (1-item)		Type of parenting style
Empathy Measure		Perspective taking Empathetic concern
Mental Health Measure		GHQ-12

Note. (1) This table shows the detailed contents of the measurements of parental investment and parenting skills. In total, we measure time investment (reported as weekdays and weekends), monetary investment, parenting style, empathy, and mental health status.

Table D7: Students' Characteristics and Bullying Behavior

	Bully		Victim		Bully-victim	
	(1)	(2)	(3)	(4)	(5)	(6)
	T1	T2	T1	T2	T1	T2
Male	0.098***	0.123***	0.103***	0.103***	0.106***	0.124***
	(0.022)	(0.021)	(0.018)	(0.023)	(0.022)	(0.019)
Age in years	-0.004	-0.016	-0.014	-0.018	-0.008	-0.018
	(0.025)	(0.020)	(0.021)	(0.020)	(0.025)	(0.019)
Urban hukou	-0.003	0.043**	-0.021	0.019	-0.002	0.030
	(0.024)	(0.019)	(0.020)	(0.019)	(0.023)	(0.020)
Onlychild	0.021	0.000	0.008	-0.017	0.022	0.001
	(0.022)	(0.022)	(0.019)	(0.021)	(0.022)	(0.021)
Height in cm	0.001	-0.000	0.001	0.000	0.002	-0.001
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Weight in half kilo	0.000	0.001**	0.000	0.000	0.000	0.001*
	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)
Empathy score	-0.044**	-0.045**	0.026	0.031	-0.031*	-0.040**
	(0.017)	(0.017)	(0.017)	(0.019)	(0.017)	(0.017)
Self-esteem index	-0.004	0.005	0.014	0.004	0.007	0.004
	(0.015)	(0.018)	(0.011)	(0.016)	(0.014)	(0.018)
Stress coping index	-0.110***	-0.060***	-0.107***	-0.093***	-0.115***	-0.060***
	(0.022)	(0.018)	(0.019)	(0.023)	(0.019)	(0.018)
Mental health index	-0.061***	-0.053***	-0.101***	-0.091***	-0.074***	-0.053***
	(0.013)	(0.014)	(0.012)	(0.016)	(0.014)	(0.015)
Weekly interaction with parents	-0.003*	-0.000	0.000	0.003	-0.001	0.000
	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Number of friends	0.036***	0.013**	0.016**	0.010	0.030***	0.009
	(0.007)	(0.007)	(0.007)	(0.008)	(0.007)	(0.006)
Member of exclusive group	0.094***	0.056***	0.071***	0.041**	0.094***	0.053***
	(0.024)	(0.016)	(0.021)	(0.019)	(0.022)	(0.015)

Note.(1) This table shows the correlations between students' characteristics and being school bullies, victims or bully-victims. The examined characteristics include demographic variables and socioemotional characteristics. (2) We construct indices for empathy, self-esteem, stress coping skills, and mental health following Anderson (2008). (3) Odd columns report the correlations for baseline bullying behaviors (T1), while even columns report the correlations for follow-up bullying behaviors (T2). (4) All regressions control for strata fixed effects. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D8: Effects on Time Spent with Parents, as Reported by Children

	(1) Control Mean	(2) ITT	(3) permutation p-value	(4) wcb	(5) TOT
Eat with parents	3.038	0.200*	0.096	0.111	0.488*
	(2.370)	(0.113)			(0.271)
Talk with parent	3.488	0.384**	0.018	0.028	0.953**
	(2.739)	(0.171)			(0.396)
Watch TV with parent	1.085	0.052	0.718	0.754	0.123
	(1.750)	(0.118)			(0.280)
Homework checked	1.766	0.375**	0.024	0.028	0.915**
	(2.524)	(0.171)			(0.405)
Outdoor activities	1.438	0.327**	0.014	0.028	0.804**
	(1.919)	(0.144)			(0.344)
N	1,029	2,246			2,246

Note. (1) This table shows the results of the robustness test when we analyze the time spent with parents reported by students. The variables measure the number of particular events with parents during a normal week in the intervened semester (range 0-8, with 8 meaning more than 7 times). (2) Column 1 reports the means and the standard deviations for students in control groups. (3) Column 2 reports the ITT estimates and standard errors, while Columns 3 and 4 report the associated permutation test and WCB p-values. Column 5 shows the 2SLS estimates with the indicator of "take-up" defined as completing at least half of the reading or movie tasks. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D9: Effects on Parents: Detailed Time and Monetary Investment

	(1) Control Mean	(2) ITT	(3) permutation p-value	(4) wcb	(5) TOT
Panel A: Details in time investment					
Read with child per day (h) weekday	0.783	0.146**	0.017	0.015	0.355***
	(0.923)	(0.056)			(0.137)
Read with child per day (h) weekend	1.031	0.149**	0.021	0.023	0.360**
	(1.032)	(0.061)			(0.147)
Help homework per day (h) weekday	0.672	0.093*	0.092	0.055	0.226**
	(0.957)	(0.048)			(0.114)
Help homework per day (h) weekend	0.899	0.157**	0.024	0.023	0.381**
	(1.159)	(0.065)			(0.155)
Play with child per day (h) weekday	0.779	0.062	0.218	0.205	0.149
	(0.857)	(0.047)			(0.112)
Play with child per day (h) weekend	1.258	0.018	0.744	0.741	0.044
	(0.989)	(0.053)			(0.126)
Other education per day (h) weekday	1.552	0.120	0.164	0.154	0.290
	(1.410)	(0.077)			(0.187)
Other education per day (h) weekend	2.295	0.021	0.840	0.830	0.050
	(1.700)	(0.089)			(0.214)
Panel B: Details in monetary investment					
Tutoring if friend did	0.279	0.016	0.509	0.477	0.039
	(0.449)	(0.021)			(0.052)
Tutoring if best student did	0.43	-0.003	0.893	0.883	-0.008
	(0.495)	(0.022)			(0.052)
Tutoring if most students did	0.464	0.007	0.775	0.765	0.017
	(0.499)	(0.022)			(0.053)
% belief tutoring helps in score	49.574	-0.351	0.725	0.684	-0.848
	(20.025)	(0.923)			(2.203)
% belief tutoring helps in mental health	47.01	-0.700	0.487	0.506	-1.690
	(21.683)	(0.947)			(2.287)

Note. (1) This table shows the ITT estimates of (1) *for parental time and monetary investments. Panel A reports the ITT estimates for different categories of time investments over the previous week; Panel B reports parents' attitudes toward cram schools/after-school tutoring. Parents were asked to choose whether they would send their kids to cram schools in three hypothetical settings: Scenario 1 - when their best friends' children went to cram schools; Scenario 2 - when the best students in the class went to cram school; and Scenario 3 - when most of the students in the class went to cram schools. Finally, we elicit the perceived value of cram schools by asking parents to score (scale 1-100) whether the cram school is good for students' test scores and whether it is good for students' mental health for a hypothetically struggling student. (2) Column 1 reports the means and the standard deviations for outcomes for parents in control groups. (3) Column 2 reports the ITT estimates and standard errors, while Columns 3 and 4 report the associated permutation test and WCB p-values. (4) Column 5 shows the TOT estimates using the indicator of "take-up," defined as having completed at least half of the tasks as the main regressor and treatment assignment as the instrument. (5) Classroom-level clustered standard errors are presented in parentheses (* p<0.10, ** p<0.05, and *** p<0.001).*

Table D10: Effects on Self-reported Engagement in Empathy-related Movie/Reading Activities

	(1) Control Mean	(2) ITT	(3) permutation p-value	(4) wcb	(5) TOT
Empathy related movie at least once	0.436	0.128***	0.000	0.000	0.328***
	(0.496)	(0.027)			(0.069)
Empathy related movie at least monthly	0.196	0.121***	0.000	0.001	0.313***
	(0.397)	(0.034)			(0.087)
Empathy related reading at least once	0.517	0.102***	0.002	0.001	0.264***
	(0.500)	(0.028)			(0.074)
Empathy related reading at least monthly	0.273	0.092**	0.002	0.011	0.240***
	(0.446)	(0.035)			(0.088)
N	1,029	2,246			2,246

Note. (1) This table shows the estimated effects on student-reported engagement in empathy-related movie or reading activities in the past semester. (2) Column 1 reports the means and the standard deviations for the corresponding outcomes for students in control groups. (3) Column 2 reports the ITT estimates and standard errors, while Columns 3 and 4 report the associated permutation P-value after 2,000 stratified clustered resampling and wild cluster bootstrap P-value after 9,999 resampling. Column 5 reports the TOT estimates using 2SLS with the indicator of "take-up" defined as completing at least half of the tasks. (4) Classroom-level clustered standard errors are presented in parentheses ($p<0.10$, ** $p<0.05$, and *** $p<0.001$).*

Table D11: Effects on Students' Outcomes (Detailed Components)

	(1) ITT	(2) Permutation p-value	(3) WCB	(4) TOT	(5) Dosage
		Panel A: Prosociality			
Prosociality index	0.080*	0.088	0.096	0.197*	0.022*
	(0.044)			(0.102)	(0.012)
		Panel B: Positive traits			
Self-satisfy	0.110**	0.046	0.059	0.267***	0.030**
	(0.053)			(0.128)	(0.014)
Self-worth	0.106*	0.064	0.088	0.260***	0.029*
	(0.056)			(0.137)	(0.015)
Self-confident	0.143**	0.010	0.018	0.353***	0.040*
	(0.057)			(0.129)	(0.015)
Self-esteem	0.150***	0.004	0.005	0.369***	0.041***
	(0.052)			(0.124)	(0.014)
Perseverance	0.179***	0.004	0.001	0.440***	0.050***
	(0.055)			(0.128)	(0.014)
		Panel C: Stress and mental health			
Inverse CES-D	0.074	0.301	0.293	0.193	0.022
	(0.063)			(0.152)	(0.017)
Feel happy	0.101*	0.082	0.097	0.254*	0.029*
	(0.056)			(0.132)	(0.015)
Inverse stress score	0.151*	0.026	0.075	0.377**	0.042*
	(0.078)			(0.188)	(0.022)
N	2,246			2,246	2,246

Note. (1) This table shows the results of the program effects on students' prosociality, positive traits, and mental health. We use inverse covariance matrix weighting methods to construct the prosociality index (contain variables: return favor and two hypothetical scenarios) and inverse stress index (contain four different sources of pressure). Panel A reports the program effects on prosociality. Panel B reports the program effects on the subcomponents of positive traits: self-satisfaction, self-worth, self-confidence, self-esteem, and perseverance. Panel C reports the program effects on inverse CESD-10 (mental health), happiness score, and the inverse stress index. (2) Column 1 reports the ITT estimates using (1) for these outcomes. Columns 2 and 3 report the associated permutation test and WCB p-values. Column 4 reports the TOT estimates using the indicator of "take-up," defined as having completed at least half of the tasks as the main regressor and treatment assignment as the instrument. (3) Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D12: Program Impacts on Test Scores

	(1) Test score	(2) Grade rank
Panel A. Average effect		
Control Mean	0.0235	-0.0284
	(0.987)	(1.003)
N	1,029	
ITT	-0.009	0.011
	(0.015)	(0.016)
N	2,240	
Panel B. Quantile		
1st Decile	0.010	0.010
	(0.020)	(0.024)
3rd Decile	-0.017	0.008
	(0.016)	(0.015)
Median	-0.012	0.013
	(0.014)	(0.013)
7th Decile	-0.013	0.017
	(0.014)	(0.016)
9th Decile	-0.010	-0.012
	(0.017)	(0.019)
N	2,240	

Note. (1) This table shows the program effects on students' test scores, measured by their scores on the final exam, shown in Column 1, and grade rank, shown in Column 2. Panel A reports the average effects by first reporting the means and standard deviations in the control group and the ITT effects. Panel B reports the estimation results of an unconditional quantile regression. (2) Classroom-level clustered standard errors are presented in parentheses ($p<0.10$, ** $p<0.05$, and *** $p<0.001$).*

Table D13: Program Impacts on School Bullying Behavior by Type (Extensive Margin)

	(1) Control Mean	(2) ITT	(3) permutation p-value	(4) wcb	(5) TOT
Panel A: Bullying perpetrator					
Threaten	0.147	-0.031	0.152	0.145	-0.042*
	(0.354)	(0.020)			(0.025)
Spread rumors	0.195	-0.038*	0.116	0.100	-0.067***
	(0.397)	(0.022)			(0.023)
Physical bullying	0.184	-0.038*	0.070	0.076	-0.054**
	(0.387)	(0.020)			(0.025)
Socially isolate	0.168	-0.023	0.244	0.278	-0.037
	(0.374)	(0.019)			(0.026)
Cyber bullying	0.114	-0.021	0.278	0.228	-0.034*
	(0.318)	(0.017)			(0.021)
Panel B: Bullying victim					
Being threatened	0.308	-0.035	0.192	0.201	-0.070**
	(0.462)	(0.025)			(0.032)
Being spread rumors	0.484	-0.028	0.378	0.424	-0.085**
	(0.500)	(0.030)			(0.041)
Physical bullying victim	0.388	-0.055**	0.052	0.057	-0.105***
	(0.487)	(0.026)			(0.038)
Being socially-isolated	0.247	-0.045*	0.066	0.062	-0.078**
	(0.431)	(0.022)			(0.032)
Being cyber-bullied	0.224	-0.035*	0.104	0.090	-0.085***
	(0.417)	(0.020)			(0.026)
Panel C: Bullying victim & bullying perpetrator					
Threaten	0.119	-0.028	0.212	0.175	-0.039*
	(0.323)	(0.019)			(0.024)
Spread rumors	0.167	-0.036*	0.084	0.099	-0.051**
	(0.373)	(0.020)			(0.022)
Physical bullying	0.154	-0.035*	0.094	0.079	-0.057**
	(0.361)	(0.019)			(0.023)
Socially isolate	0.099	-0.014	0.376	0.334	-0.024
	(0.299)	(0.014)			(0.020)
Cyber bullying	0.093	-0.027*	0.170	0.095	-0.040**
	(0.291)	(0.015)			(0.019)

Note. (1) This table shows the program effects on being school bullies, victims and bully-victims for various domains of school bullying. Panel A reports bullying perpetrators, Panel B reports bullying victims, and Panel C reports bully-victims. Within each panel, each row reports the results for a specific type of bullying. (2) We use Column 1 to report the means and the standard deviations for outcomes for students in control groups. (3) Column 2 reports the ITT estimates and standard errors, while Columns 3 and 4 report the associated permutation P-value after 2,000 stratified clustered resampling and wild cluster bootstrap P-value after 9,999 resampling. Column 5 reports the TOT estimates using the indicator of "take-up," defined as having completed at least half of the tasks as the main regressor and treatment assignment as the instrument. (4) All regressions control for strata fixed effects. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D14: Program Impacts on School Bullying Behavior (Intensive Margin)

	(1) Control Mean	(2) ITT	(3) permutation p-value	(4) wcb	(5) TOT
Panel A: Bullying perpetrator					
Threaten	1.296	-0.042	0.486	0.442	-0.104
	(0.828)	(0.050)			(0.121)
Spread rumors	1.374	-0.074	0.248	0.203	-0.185
	(0.895)	(0.055)			(0.128)
Physical bullying	1.377	-0.088*	0.124	0.103	-0.218*
	(0.928)	(0.051)			(0.119)
Socially isolate	1.34	-0.053	0.324	0.394	-0.134
	(0.890)	(0.054)			(0.129)
Cyber bullying	1.261	-0.041	0.480	0.434	-0.106
	(0.818)	(0.049)			(0.117)
N	1,029	2,246			2,246
Panel B: Bullying victim					
Being threatened	1.646	-0.126**	0.034	0.047	-0.312**
	(1.150)	(0.058)			(0.135)
Being spread rumors	1.993	-0.045	0.536	0.564	-0.113
	(1.276)	(0.074)			(0.177)
Physical bullying victim	1.857	-0.150**	0.044	0.049	-0.371**
	(1.294)	(0.070)			(0.162)
Being socially-isolated	1.5	-0.116**	0.048	0.042	-0.288**
	(1.014)	(0.053)			(0.124)
Being cyber-bullied	1.484	-0.094*	0.122	0.099	-0.238*
	(1.039)	(0.054)			(0.127)
N	1,029	2,246			2,246

Note. (1) This table shows the results of the program effects on the intensive margin of school bullying behaviors, defined as the frequency of each behavior. Panel A reports the program effects on bullying perpetration, and Panel B reports the program effects on bullying victimization. (2) Column 1 reports the means and the standard deviations for students in control groups. (3) Column 2 reports the ITT estimates and standard errors, while Columns 3 and 4 report the associated permutation P-value after 2,000 stratified clustered resampling and wild cluster bootstrap P-value after 9,999 resampling. Column 5 reports the TOT estimates using the indicator of "take-up," defined as having completed at least half of the tasks as the main regressor and treatment assignment as the instrument. (4) All regressions control for strata fixed effects. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D15: Heterogeneous Effects on Bullies and Bullying Victims

	Panel A. Bullying perpetrator					Panel B. Bullying victim				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment	-0.047	-0.048	-0.034	-0.029	-0.044	-0.061*	-0.011	-0.029	-0.021	-0.047
	(0.032)	(0.043)	(0.028)	(0.033)	(0.032)	(0.034)	(0.049)	(0.034)	(0.035)	(0.039)
Third quartile of empathy X treatment	-0.028					0.061				
	(0.051)					(0.039)				
Third quartile of empathy	-0.032					-0.015				
	(0.039)					(0.029)				
Lower than medium age X treatment		-0.011					-0.068			
		(0.047)					(0.047)			
Lower than medium age		-0.003					0.074*			
		(0.039)					(0.037)			
First quartile of parental involvement X treatment			**-0.078***					-0.064		
			(0.045)					(0.055)		
First quartile of parental involvement			0.063**					0.049		
			(0.031)					(0.043)		
First quartile of pressure X treatment				**-0.068***					-0.058	
				(0.038)					(0.041)	
First quartile of pressure				-0.025					-0.081***	
				(0.030)					(0.028)	
Male X treatment					-0.019					0.003
					(0.043)					(0.046)
Male					0.137***					0.076**
					(0.030)					(0.036)

Note. (1) This table shows the heterogeneity of the treatment effects on bullying perpetration (Panel A) and bullying victimization (Panel B) following (3). Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D16: Heterogeneous Effects on Bully-Victims and Bystanders

	Panel A. Bullying perpetrator & victim					Panel B. Willing to help victims				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment	-0.052	-0.052	-0.033	-0.032	-0.057*	0.063**	0.060*	0.064***	0.044*	0.009
	(0.032)	(0.038)	(0.027)	(0.032)	(0.029)	(0.025)	(0.032)	(0.020)	(0.026)	(0.024)
Third quartile of empathy X treatment	-0.031					-0.039				
	(0.049)					(0.036)				
Third quartile of empathy	-0.022					0.054**				
	(0.039)					(0.027)				
Lower than medium age X treatment		-0.014					-0.015			
		(0.042)					(0.034)			
Lower than medium age		-0.005					0.057*			
		(0.038)					(0.033)			
First quartile of parental involvement X treatment			**-0.099****					-0.038		
			(0.045)					(0.042)		
First quartile of parental involvement			0.077**					-0.042		
			(0.029)					(0.038)		
First quartile of pressure X treatment				**-0.079****					0.026	
				(0.035)					(0.046)	
First quartile of pressure				-0.021					-0.008	
				(0.028)					(0.035)	
Male X treatment					-0.005					**0.082****
					(0.039)					**(0.032)**
Male					0.127***					-0.142***
					(0.030)					(0.020)

Note. (1) This table shows the heterogeneity of the treatment effects on bully-victims (Panel A) and the willingness to help bullying victims (Panel B) following (3). Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D17: Heterogeneous Effects on Empathy Skills

	Empathy skill				
	(1)	(2)	(3)	(4)	(5)
Treatment	0.168**	0.181*	0.129**	0.122*	0.103
	(0.072)	(0.101)	(0.058)	(0.068)	(0.070)
Third quartile of empathy X treatment	-0.070				
	(0.083)				
Third quartile of empathy	0.219***				
	(0.066)				
Lower than medium age X treatment		-0.066			
		(0.107)			
Lower than medium age		0.081			
		(0.086)			
First quartile of parental involvement X treatment			0.079		
			(0.090)		
First quartile of parental involvement			-0.203***		
			(0.065)		
First quartile of pressure X treatment				0.077	
				(0.069)	
First quartile of pressure				-0.007	
				(0.048)	
Male X treatment					0.084
					(0.078)
Male					-0.119**
					(0.058)

Note. (1) This table shows the heterogeneity of the treatment effects on empathy skills following (3). Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D18: Effects on School Bullying: Accounting for Misreporting

	Bullying perpetrator		Bullying victim		Bully&victim	
	(1)	(2)	(3)	(4)	(5)	(6)
	Probit	Probit (HAS)	Probit	Probit (HAS)	Probit	Probit (HAS)
Treatment coefficients	-0.155*	-0.155*	-0.104	-0.123	-0.203**	-0.203**
	(0.081)	(0.081)	(0.85)	(0.137)	(0.085)	(0.085)
Marginal effects at the mean	-0.053*	-0.053*	-0.040	-0.044	-0.065**	-0.065**
	(0.027)	(0.027)	(0.033)	(0.043)	(0.027)	(0.027)
Misreporting (false negative rate)		0.000		0.092		0.000
N	2,246	2,246	2,246	2,246	2,246	2,246

Note. (1) The table shows the regression coefficients for the program treatment effect on school bullying behaviors using a probit specification. Columns 1, 3 and 5 are probit models, while Columns 2, 4 and 6 are probit models allowing for misreporting following Hausman et al. (1998). In particular, the model allows for the possibility of false negatives (report no bullying behaviors while being involved in bullying). The row Misreporting shows the estimated probability of misreporting. All analyses additionally control for individual demographics, social desirability scale, and randomization strata. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D19: Effects on School Bullying: Varying Levels of Misreporting (α_1)

	(1)	(2)	(3)
α_1	Bullying perpetrator	Bullying victim	Bully&victim
0	-0.053*	-0.044	-0.059**
	(0.028)	(0.032)	(0.027)
0.1	-0.059*	-0.051	-0.065**
	(0.031)	(0.037)	(0.030)
0.2	-0.066*	-0.060	-0.074**
	(0.034)	(0.044)	(0.033)
0.3	-0.076*	-0.072	-0.084**
	(0.039)	(0.061)	(0.038)
0.4	-0.089*	-0.068	-0.099**
	(0.046)	(0.062)	(0.044)
0.5	-0.105*	-0.049	-0.118**
	(0.055)	(0.000)	(0.052)

Note. (1) The table shows the program treatment effect on school bullying behaviors allowing for misreporting following Hausman et al. (1998). In particular, the model allows for the possibility of false negatives (report no bullying behaviors while being involved in bullying). We vary the level of misreporting α_1. All analyses additionally control for individual demographics and randomization strata. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D20: Effects on School Bullying by Type: Varying Levels of Misreporting (α_1)

	(1) Threaten	(2) Spread rumors	(3) Physical bullying	(4) Socially isolate	(5) Cyber bullying
α_1			Panel A: Bullying perpetrator		
0	-0.030	-0.042*	-0.042**	-0.023	-0.019
	(0.020)	(0.022)	(0.021)	(0.020)	(0.015)
0.1	-0.033	-0.046*	-0.047**	-0.025	-0.021
	(0.022)	(0.024)	(0.023)	(0.022)	(0.017)
0.2	-0.038	-0.052*	-0.053**	-0.028	-0.023
	(0.024)	(0.027)	(0.026)	(0.025)	(0.019)
0.3	-0.043	-0.059*	-0.061**	-0.032	-0.026
	(0.027)	(0.030)	(0.029)	(0.028)	(0.021)
0.4	-0.050	-0.068*	-0.072**	-0.037	-0.031
	(0.031)	(0.035)	(0.034)	(0.032)	(0.024)
0.5	-0.060	-0.081*	-0.087**	-0.043	-0.037
	(0.037)	(0.042)	(0.040)	(0.038)	(0.029)
			Panel B: Bullying victim		
0	-0.045	-0.027	-0.057*	-0.039	-0.040**
	(0.028)	(0.032)	(0.031)	(0.025)	(0.019)
0.1	-0.050	-0.031	-0.064*	-0.043	-0.045**
	(0.031)	(0.036)	(0.035)	(0.027)	(0.021)
0.2	-0.057	-0.035	-0.073*	-0.048	-0.050**
	(0.035)	(0.041)	(0.039)	(0.030)	(0.024)
0.3	-0.065*	-0.042	-0.084*	-0.055	-0.057**
	(0.039)	(0.048)	(0.044)	(0.034)	(0.027)
0.4	-0.076*	-0.052	-0.099**	-0.063	-0.067**
	(0.045)	(0.060)	(0.050)	(0.039)	(0.031)
0.5	-0.091*	-0.076	-0.114**	-0.075	-0.080**
	(0.052)	(0.173)	(0.054)	(0.046)	(0.037)
			Panel C: Bullying victim & bullying perpetrator		
0	-0.028	-0.034*	-0.037*	-0.011	-0.025*
	(0.018)	(0.020)	(0.020)	(0.015)	(0.014)
0.1	-0.031	-0.037*	-0.041*	-0.012	-0.027*
	(0.020)	(0.022)	(0.022)	(0.016)	(0.016)
0.2	-0.035	-0.042*	-0.046*	-0.013	-0.031*
	(0.023)	(0.024)	(0.025)	(0.018)	(0.017)
0.3	-0.041	-0.048*	-0.053*	-0.015	-0.035*
	(0.025)	(0.028)	(0.028)	(0.020)	(0.020)
0.4	-0.047	-0.056*	-0.062*	-0.017	-0.041*
	(0.029)	(0.032)	(0.032)	(0.023)	(0.023)
0.5	-0.057*	-0.067*	-0.075**	-0.019	-0.050*
	(0.034)	(0.038)	(0.038)	(0.027)	(0.027)

Note. (1) The table shows the program treatment effect on school bullying behaviors by detailed type allowing for misreporting following Hausman et al. (1998). In particular, the model allows for the possibility of false negatives (report no bullying behaviors while being involved in bullying). We vary the level of misreporting α_1. All analyses additionally control for individual demographics and randomization strata. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D21: Bullying Incidents Reported by Parents

	(1) Control Mean	(2) ITT	(3) permutation p-value	(4) wcb	(5) TOT
			Bullying victim		
Bullying victim	0.237	-0.042**	0.024	0.043	-0.099**
	(0.426)	(0.020)			(0.047)
Threaten	0.056	-0.009	0.442	0.440	-0.021
	(0.231)	(0.010)			(0.024)
Spread rumors	0.144	-0.038***	0.008	0.004	-0.089***
	(0.351)	(0.013)			(0.032)
Physical bullying	0.121	-0.019	0.236	0.235	-0.045
	(0.326)	(0.015)			(0.035)
Socially isolate	0.089	-0.026**	0.044	0.044	-0.060**
	(0.284)	(0.012)			(0.029)
Cyber bullying	0.037	-0.002	0.842	0.861	-0.004
	(0.189)	(0.008)			(0.019)
N	868	1,899			1,899

Note. (1) This table shows the robustness analysis by exploring bullying incidents reported by parents to complement the main results in Table 4. These are all bullying victims, and the first variable "bullying victim" is an indicator for a victim of any type of bullying behavior. (2) Column 1 reports the means and the standard deviations for control groups. Column 2 reports the ITT estimates using (1) for these outcomes. Columns 3 and 4 report the associated permutation tests and WCB p-values. Column 5 reports the TOT estimates using the indicator of "take-up," defined as having completed at least half of the tasks as the main regressor and treatment assignment as the instrument. (5) Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D22: Robustness Analysis For Bullying Behaviors

	(1)	(2)	(3)	(4)	(5)	(6)
	Main	Demographic	SDS	Survey time	Pooled	Entropy Balance
School bullying involvements:						
Bullying perpetrator	-0.053**	-0.053*	-0.056**	-0.051**	-0.033***	-0.078**
	(0.025)	(0.028)	(0.026)	(0.023)	(0.009)	(0.038)
Bullying victim	-0.044	-0.046	-0.052*	-0.042	-0.041***	-0.088**
	(0.029)	(0.032)	(0.030)	(0.026)	(0.013)	(0.044)
Bully & victim	-0.065**	-0.059**	-0.062**	-0.062***	-0.029***	-0.084**
	(0.025)	(0.036)	(0.025)	(0.023)	(0.009)	(0.037)
Spectators:						
Witnessed bullying incidents	-0.061*	-0.061*	-0.065*	-0.061*		-0.102***
	(0.034)	(0.034)	(0.033)	(0.032)		(0.036)
Help bullying victims	0.052**	0.052**	0.052**	0.051***		-0.048
	(0.020)	(0.021)	(0.020)	(0.019)		(0.029)
Baseline outcomes	Yes	Yes	Yes	Yes	-	Yes
Demographics	-	Yes	Yes	Yes	-	Yes
Social desirability scale	-	-	Yes	Yes	-	Yes
Survey time	-	-	-	Yes	-	Yes
Bullying type fixed effects	-	-	-	-	Yes	-
N	2,246	2,246	2,246	2,246	11,125	1,522

Note. (1) This table shows the robustness analysis to complement the main results in Table 4. (2) Column 1 shows the results from our main specification. Column 2 reports the estimates controlling for individual demographics. Column 3 additionally controls for survey completion time and square of the completion time. Column 4 additionally controls for the social desirability scale measured at baseline. Columns 5 and 6 estimate the impact effects using different models. Column 5 pools all types of bullying behaviors and estimates them with type fixed effects. Column 6 uses program take-up as independent variable and estimates the effect of taking up the program using the entropy balancing (EB) method. EB gives more conservative estimates that lie within the range of ITT and TOT. (3) Classroom-level clustered standard errors are presented in parentheses ($p<0.10$, ** $p<0.05$, and *** $p<0.001$).*

Table D23: Descriptive Statistics after EB and PSM: Means and Standardized Bias

Variables	Means tookup	Means control			Standardized Bias (%)		
		Raw	EB	PSM	Raw	EB	PSM
Age	14.6	14.5	14.6	14.6	27.8	1.0	-2.5
Male	0.6	0.5	0.6	0.6	5.8	0.0	-2.7
Urban hukou	0.5	0.5	0.5	0.5	-3.5	0.0	-0.9
Onlychild	0.3	0.3	0.3	0.3	1.2	0.0	-1.4
Height	162.8	161.8	162.8	163.3	14.1	0.7	-5.7
Weight	100.0	101.6	100.0	100.5	-7.7	0.2	-2.3
Bullying victim	0.7	0.7	0.7	0.7	-5.4	0.0	0.6
Bullying perpetrator	0.3	0.4	0.3	0.3	-5.0	0.0	-1.1
Empathy score	49.5	47.6	49.4	49.6	19.6	0.2	-1.8
Self-satisfied	4.6	4.4	4.6	4.7	11.7	0.1	-2.1
Self-worth	4.9	4.7	4.9	5.0	10.8	0.1	-5.7
Self-confident	5.1	5.0	5.1	5.2	10.5	0.1	-3.5
Self-esteem	4.8	4.6	4.8	4.9	9.2	0.1	-5.9
Preseverance	4.9	4.7	4.9	4.9	7.5	0.1	-2.1
Pressure score	13.1	13.3	13.1	13.1	-3.8	0.1	1.3
CESD 10-item	7.9	8.9	7.9	7.7	-18.2	0.0	3.6
Depressed	0.2	0.3	0.2	0.2	-11.5	-0.0	1.8
Happiness score	5.3	5.0	5.3	5.3	16.1	0.1	-2.2
Weekly interaction with parents	10.3	10.7	10.3	10.6	-4.9	0.0	-3.9
Rank pressure	4.6	4.6	4.6	4.5	-3.9	0.1	2.7
College aspiration	6.7	6.5	6.7	6.7	16.9	0.2	1.5
Optimistic about future	5.6	5.4	5.6	5.7	12.0	0.1	-3.6
Return favor	0.7	0.6	0.7	0.7	9.3	0.1	2.8
Feel lonely during childhood	0.5	0.5	0.5	0.5	-6.7	0.0	-0.8
Have dinner with parents	2.9	3.3	2.9	3.0	-14.7	0.0	-1.8
Chat about school lives with parents	3.9	4.0	3.9	4.0	-2.0	0.0	-3.3
Watch TV with parents	1.1	1.1	1.1	1.1	-1.8	0.0	-2.6
Homework checked	2.4	2.3	2.4	2.5	3.8	0.0	-3.1
Outdoor activities with parents	1.5	1.6	1.5	1.6	-5.6	0.0	-4.0
Feel close to dad	2.1	2.1	2.1	2.1	0.8	0.1	-2.9
Feel close to mom	2.5	2.5	2.5	2.5	2.5	0.1	-3.2
Brought up by mother (before age 6)	0.4	0.5	0.4	0.4	-7.5	0.0	0.9
Parents control friendship	0.2	0.2	0.2	0.2	3.9	0.0	-0.8
Pocket money peer week	2.2	2.3	2.2	2.2	-3.6	0.1	0.9
N	495	1,027					

Note. This table shows the comparison results between entropy balancing (EB) and propensity score matching (PSM). The pretreatment means of the variables used in the two methods for the take-up and control groups are in the first and second columns, respectively. The "take-up" group is defined as completing at least half of the tasks. The means of the reweighted control group using entropy balancing weights and using PSM are in the third and fourth columns, respectively. The last three columns make up the standardized difference in means, a matching quality indicator. The standardized difference in means for each control variable s is defined as $SD_s = 100 \cdot \left(\overline{s_1} - \overline{s_0}\right) / \sqrt{0.5 \cdot \left(\sigma_{s1}^2 + \sigma_{s0}^2\right)}$*, where* $\bar{s}_1$ *and* s_0 *are the means of treated and controls, respectively, and* σ_{s1}^2 *and* σ_{s0}^2 *are the corresponding variances. The mean represents a percentage share.*

Table D24: Spillover Effect from Compliers

	Panel A. Whole sample (N=2,246)			Panel B. Not-take-up sample (N=1,751)		
	(1)	(2)	(3)	(4)	(5)	(6)
	Bullying perpetrator	Bullying victim	Bullying&victim	Bullying perpetrator	Bullying victim	Bullying&victim
Linear-in-mean:						
Take up ratio	-0.164**	-0.047	-0.168***	-0.167**	-0.027	-0.167**
	(0.064)	(0.073)	(0.057)	(0.073)	(0.069)	-0.066
2SLS:						
Take up ratio	-0.090	-0.018	-0.100	-0.090	-0.013	-0.101
	(0.077)	(0.078)	(0.071)	(0.078)	(0.076)	(0.073)
1st stage:						
Estimate	0.355***	0.355***	0.355***	0.349***	0.349***	0.349***
	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)
F-statistics	106	106	106	103	103	103

Note. (1) This table shows the estimates for spillover effects on school bullying behaviors using a linear-in-mean specification ((2)). Panel A reports the ITT estimates for the whole sample; Panel B reports the ITT estimates for the non-take-up sample. (2) We first report the correlates using a linear-in-mean model. We then report 2SLS estimates using the treatment assignment as the instrument for the take-up rate and first-stage statistics. (3) Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.001).*

Table D25: Baseline Characteristic Importance Ranking Predicted by GRF Analysis

	Variable importance rank				
	Bullying perpetrator	Bullying victim	Bullying perpetrator & victim	Help victims	Empathy skill
Baseline characteristics					
empathy skill	1	2	1	1	1
age	5	1	4	2	2
parental involvement	2	3	2	4	3
pressure score	4	6	3	3	4

Note. This table shows the variable importance rank predicted by the generalized random forest (GRF). The four variables are selected and ranked out of 24 baseline characteristics that we used in the prediction algorithm.