APPENDIX

A Detailed Items in the Questionnaire

A.1 Students' Outcome Measures

We construct measures on school bullying behaviors. We also collect multifaceted social-emotional skills. Apart from empathy, we further construct measures of students' mental health status, stress resistance, positive personality traits (positive self-image and perseverance), 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 a 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 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. Students 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 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 cell phone 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, talking/discussing school life, 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 complement 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 (GHQ-12) 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 in 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 the determinants of each student's 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 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 the 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 D7-D11, 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.[5][6] 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 heterogeneous 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 D11.[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; social-emotional 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 a close relationship with the father, indicator of a 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 D11, 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

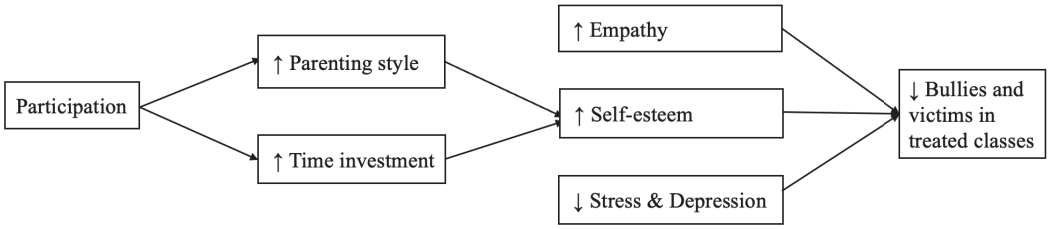
We follow the strategy adopted in studying sexual behavior context by De Paula et al. (2014) and Delavande and Kohler (2016) to correct for misclassification error. We assume that students truthfully report when they do not engage in any bullying incidents and that there is a constant probability α_1 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) bully, (ii) victim, and (iii) bully-victim. 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 the 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



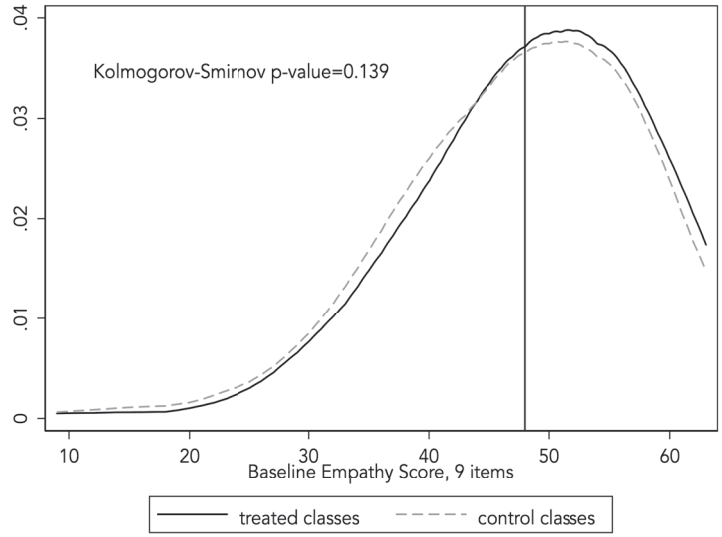
Note. The figure illustrates the theoretical framework, which guides us in implementing and evaluating this parental involvement program.

Figure D2: Program Theme

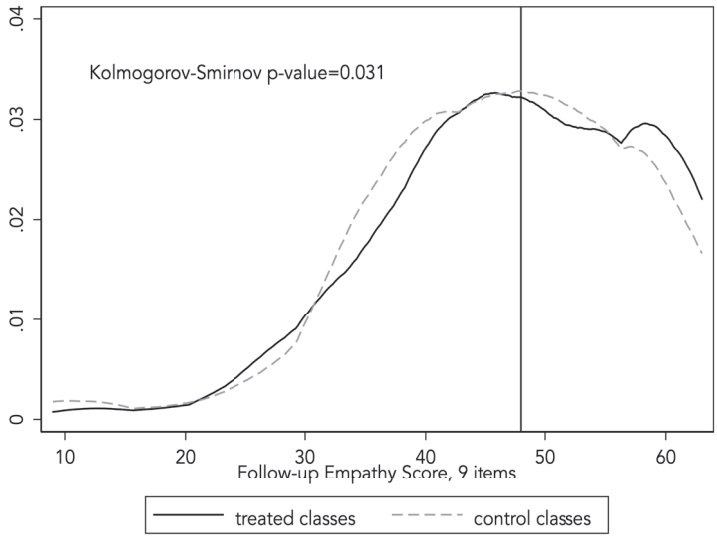
Note. The figure details the theme of our 4-month parental involvement program. We rely on two books, one by A. Ciaramicoli and one by K. Ketcham, as references for these themes and some research papers to relate school bullying and empathy for the last theme. We expand the details of the program in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

Figure D3: Distribution of Empathy Across Treatment and Control Groups at Baseline and Follow-up

Panel A. Baseline



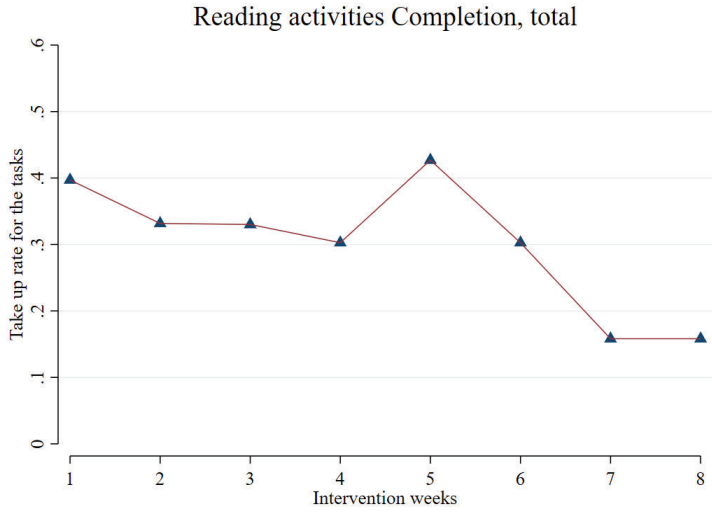
Panel B. Follow-up



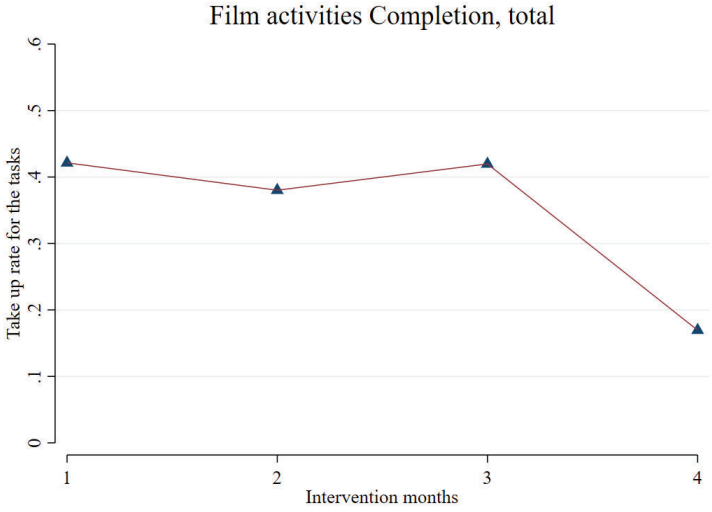
Note: Panel A shows the distribution of the raw empathy score for the treatment and control groups at the baseline. Panel B shows the distribution of the raw empathy score for the treatment and control groups at the follow-up. The P-value for the combined Kolmogorov–Smirnov test for baseline empathy score is 0.139. The P-value for the combined Kolmogorov–Smirnov test for follow-up empathy score is 0.031.

Figure D4: Task Completion Rate by Task

Panel A. Reading



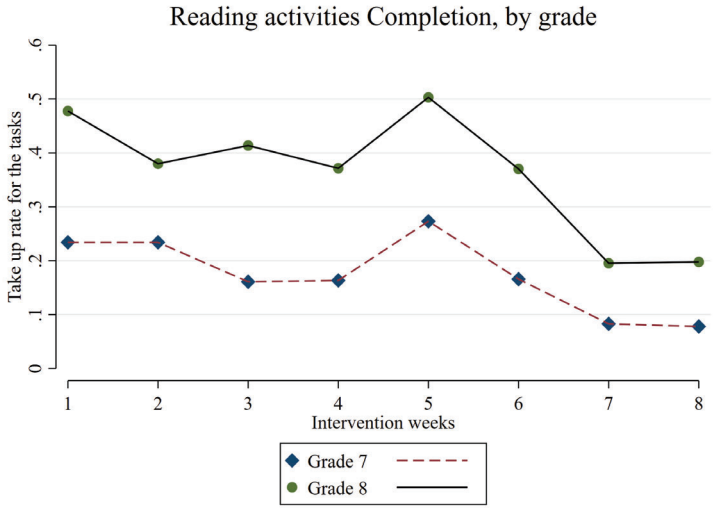
Panel B. Movie



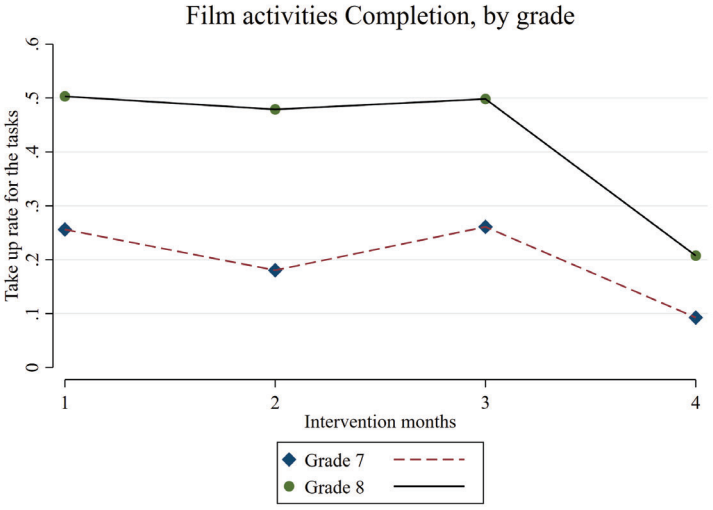
Note: Panel A shows the completion rate of the biweekly reading activities. Panel B shows the completion rate of the monthly movie activities. The numbers are calculated by the total number of those who completed the specific task divided by the total number of those who registered to participate in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

Figure D5: Task Completion Rate by Task and Grade Level

Panel A. Reading

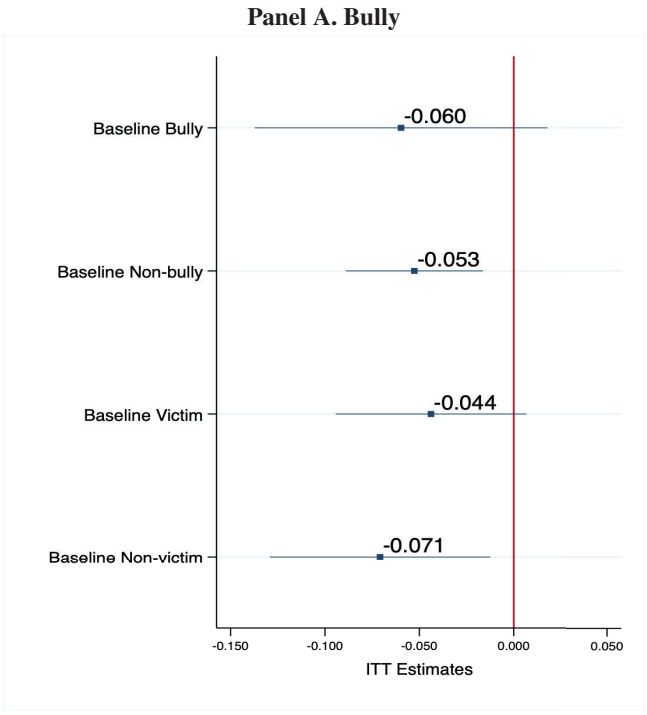


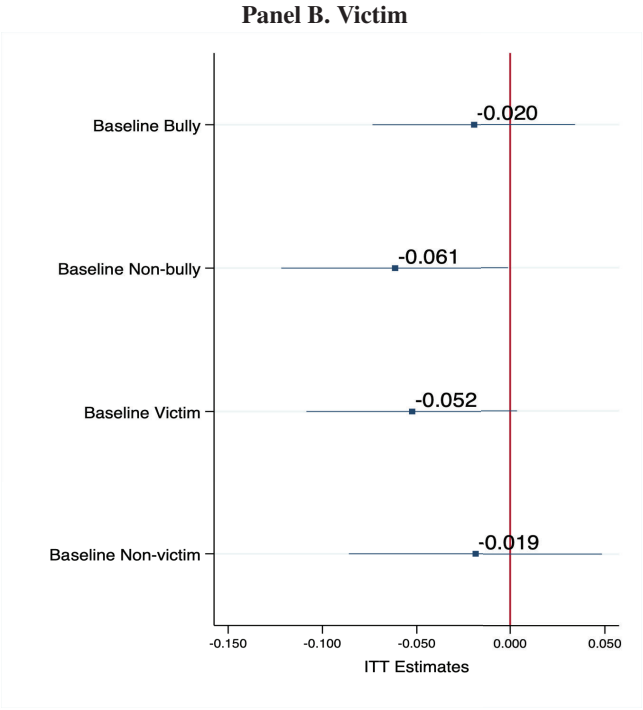
Panel B. Movie



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%.

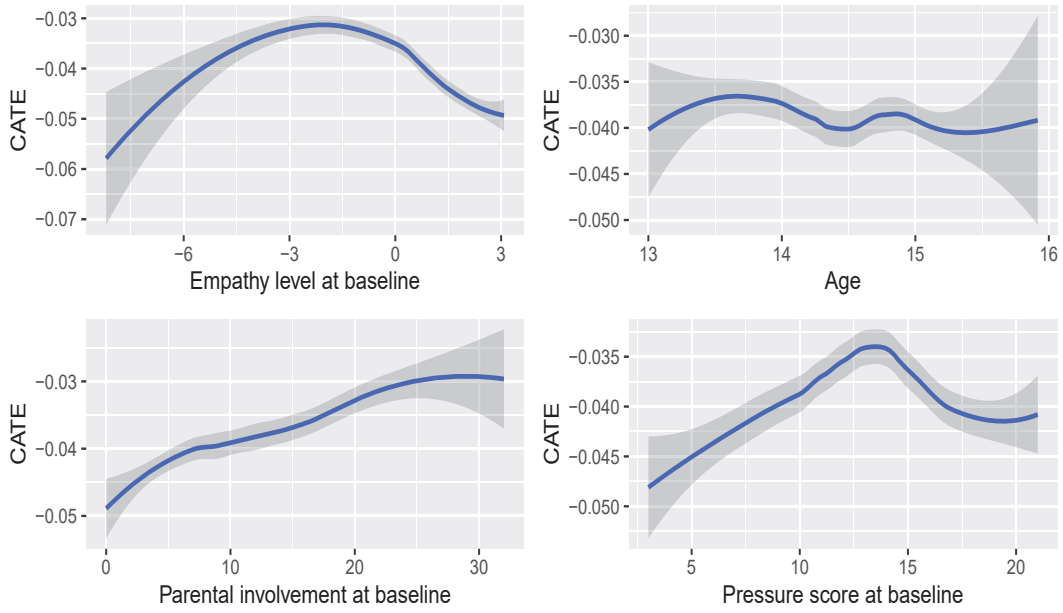
Figure D6: Effects on Bullying by Baseline Bully Status





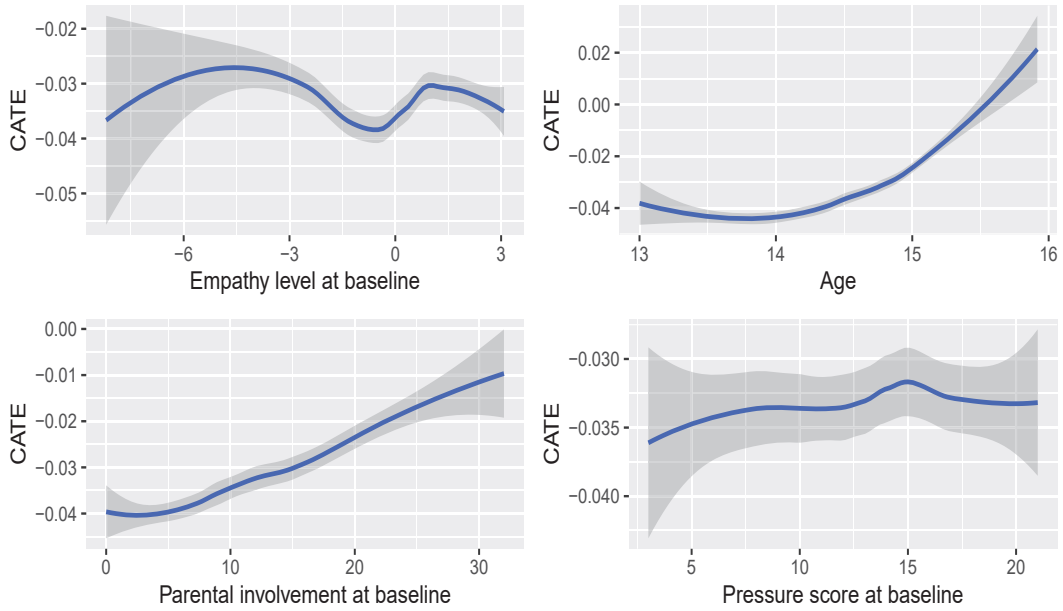
Note: Panel A shows the point estimates and 90% confidence intervals of the program's impacts on bullying involvement by baseline bully status, including being a bully, a nonbully, a victim, and a nonvictim. Panel B shows the point estimates and 90% confidence intervals of the program's impacts on being 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 D7: Out-of-Bag CATE Estimates for Bully from GRF-Trained Algorithm along Observable Characteristics



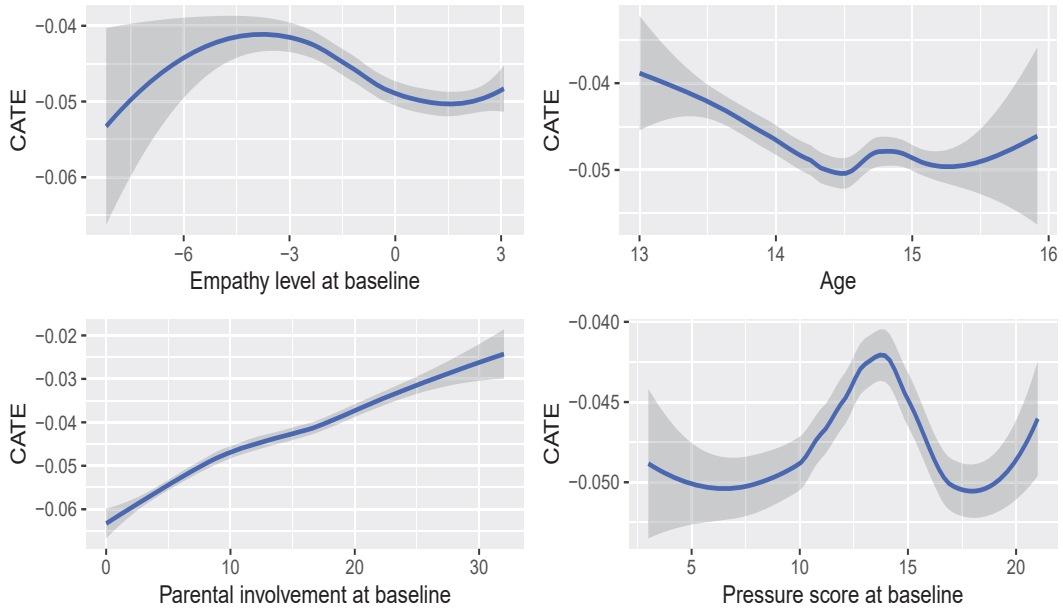
Note: This figure shows the out-of-bag CATE estimates for the bully indicator from the GRF-trained algorithm along the four baseline characteristics. In the case of out-of-bag prediction, the estimated CATEs only consider trees for which the observation is not used as part of the training set.

Figure D8: Out-of-Bag CATE Estimates for Victim from GRF-Trained Algorithm along Observable Characteristics



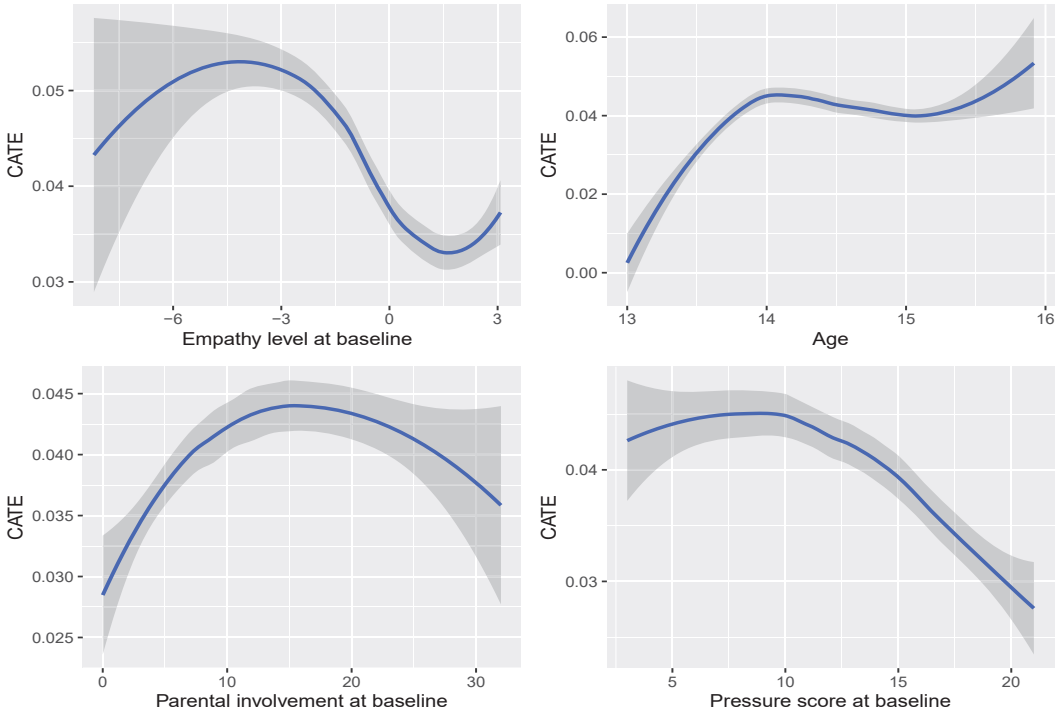
Note: This figure shows the out-of-bag CATE estimates for the victim indicator from the GFR-trained algorithm along the four baseline characteristics. In the case of out-of-bag prediction, the estimated CATEs only consider trees for which the observation is not used as part of the training set.

Figure D9: Out-of-Bag CATE Estimates for Bully-Victim from GRF-Trained Algorithm along Observable Characteristics



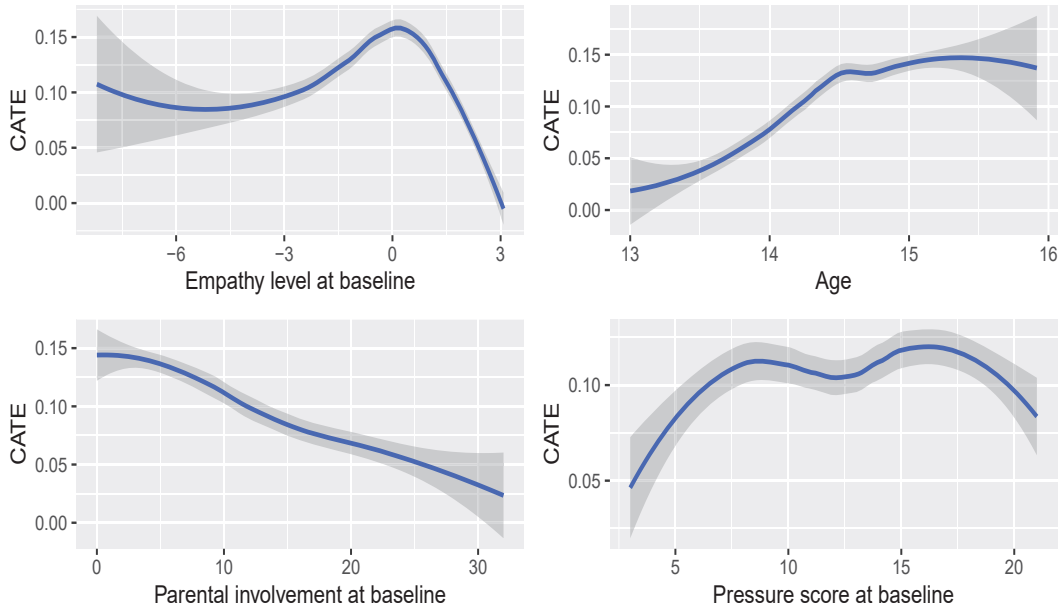
Note: This figure shows the out-of-bag CATE estimates for the bully-victim indicator from the GFR-trained algorithm along the four baseline characteristics. In the case of out-of-bag predictions, the estimated CATEs only consider trees for which the observation is not used as part of the training set.

Figure D10: Out-of-Bag CATE Estimates for Bystander 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 D11: Out-of-Bag CATE Estimates for Empathy from GRF-Trained Algorithm along Observable Characteristics



Note: This figure shows the out-of-bag CATE estimates for students' empathy skills 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 Seventh and Eighth 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 in cm	161.637	8.583	1.368**	(0.551)	-0.766***	(0.288)
Weight in half kilo	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 seventh and eighth 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 the 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: Bullying Behaviors by Type at Baseline

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

Note. (1) This table shows the distribution of bullying behaviors by detailed type at the baseline. 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 semester before the intervention.

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 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 respecting 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 relationships 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 on Friday evenings. 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 on the platform on the first Friday of 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	848	1,004	156	1,852
Number of attrition	181	213	32	394
Attrition rate	0.176 (0.380)	0.175 (0.380)	0.001 (0.016)	

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) Cognitive	(2) 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) T1	(2) T2	(3) T1	(4) T2	(5) T1	(6) 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 bullies, victims or bully-victims. The examined characteristics include demographic variables and social-emotional skills. (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<0.01).*

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<0.01).*

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	0.762	0.154***	0.008	0.012	0.370***
	(0.900)	(0.057)			(0.136)
Read with child per day (h) weekend	1.020	0.158**	0.011	0.013	0.379***
	(1.013)	(0.061)			(0.145)
Help homework per day (h) weekday	0.648	0.148***	0.011	0.011	0.355***
	(0.933)	(0.055)			(0.128)
Help homework per day (h) weekend	0.888	0.163**	0.034	0.042	0.390**
	(1.178)	(0.074)			(0.176)
Play with child per day (h) weekday	0.780	0.063	0.182	0.216	0.150
	(0.861)	(0.047)			(0.111)
Play with child per day (h) weekend	1.263	0.018	0.751	0.752	0.043
	(1.005)	(0.054)			(0.126)
Other education per day (h) weekday	1.536	0.163*	0.043	0.079	0.387*
	(1.418)	(0.084)			(0.198)
Other education per day (h) weekend	2.264	0.081	0.414	0.438	0.193
	(1.660)	(0.096)			(0.229)
N	848	1,852			1,852
Panel B: Details in monetary investment					
Tutoring if friend did	0.268	0.017	0.458	0.424	0.041
	(0.443)	(0.020)			(0.048)
Tutoring if best student did	0.423	-0.001	0.946	0.956	-0.003
	(0.494)	(0.023)			(0.055)
Tutoring if most students did	0.463	-0.006	0.809	0.792	-0.014
	(0.499)	(0.021)			(0.049)
% belief tutoring helps in score	49.467	-0.135	0.894	0.896	-0.322
	(20.203)	(0.985)			(2.323)
% belief tutoring helps in mental health	46.890	-0.649	0.528	0.508	-1.548
	(21.884)	(0.936)			(2.229)
N	848	1,852			1,852

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 of 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 the treatment assignment indicator as the instrument. (5) Classroom-level clustered standard errors are presented in parentheses (* p<0.10, ** p<0.05, and *** p<0.01).*

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<0.01).*

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's effects on students' prosociality, positive traits, and mental health. We use inverse covariance matrix weighting methods to construct the prosociality index (containing variables: return favor and two hypothetical scenarios) and the inverse stress index (containing four different sources of pressure). Panel A reports the program's effects on prosociality. Panel B reports the program's 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 the treatment assignment indicator as the instrument. (3) Classroom-level clustered standard errors are presented in parentheses ($p<0.10$, ** $p<0.05$, and *** $p<0.01$).*

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's 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 *** $p<0.01$).*

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

	(1) Control Mean	(2) ITT	(3) permutation p-value	(4) wcb	(5) TOT
Panel A: Bully					
Threatening	0.147	-0.031	0.152	0.145	-0.042*
	(0.354)	(0.020)			(0.025)
Rumor spreading	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)
Social isolation	0.168	-0.023	0.244	0.278	-0.037
	(0.374)	(0.019)			(0.026)
Cyberbullying	0.114	-0.021	0.278	0.228	-0.034*
	(0.318)	(0.017)			(0.021)
Panel B: Victim					
Threatening	0.308	-0.035	0.192	0.201	-0.070**
	(0.462)	(0.025)			(0.032)
Rumor spreading	0.484	-0.028	0.378	0.424	-0.085**
	(0.500)	(0.030)			(0.041)
Physical bullying	0.388	-0.055**	0.052	0.057	-0.105***
	(0.487)	(0.026)			(0.038)
Social isolation	0.247	-0.045*	0.066	0.062	-0.078**
	(0.431)	(0.022)			(0.032)
Cyberbullying	0.224	-0.035*	0.104	0.090	-0.085***
	(0.417)	(0.020)			(0.026)
Panel C: Bully-victim					
Threatening	0.119	-0.028	0.212	0.175	-0.039*
	(0.323)	(0.019)			(0.024)
Rumor spreading	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)
Social isolation	0.099	-0.014	0.376	0.334	-0.024
	(0.299)	(0.014)			(0.020)
Cyberbullying	0.093	-0.027*	0.170	0.095	-0.040**
	(0.291)	(0.015)			(0.019)

Note. (1) This table shows the program's effects on being bullies, victims, and bully-victims for various domains of school bullying. Panel A reports bullies, Panel B reports 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 the treatment assignment indicator 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<0.01).*

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

	(1) Control Mean	(2) ITT	(3) permutation p-value	(4) wcb	(5) TOT
			Panel A: Bully		
Threatening	1.296	-0.042	0.486	0.442	-0.104
	(0.828)	(0.050)			(0.121)
Rumor spreading	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)
Social isolation	1.34	-0.053	0.324	0.394	-0.134
	(0.890)	(0.054)			(0.129)
Cyberbullying	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: Victim		
Threatening	1.646	-0.126**	0.034	0.047	-0.312**
	(1.150)	(0.058)			(0.135)
Rumor spreading	1.993	-0.045	0.536	0.564	-0.113
	(1.276)	(0.074)			(0.177)
Physical bullying	1.857	-0.150**	0.044	0.049	-0.371**
	(1.294)	(0.070)			(0.162)
Social isolation	1.5	-0.116**	0.048	0.042	-0.288**
	(1.014)	(0.053)			(0.124)
Cyberbullying	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's effects on the intensive margin of bullying behaviors, defined as the frequency of each behavior. Panel A reports the program's effects on bullies, and Panel B reports the program's effects on victims. (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 the treatment assignment indicator 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<0.01).*

Table D15: Heterogeneous Effects on Bullies and Victims

	Panel A. Bully					Panel B. 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 bullies (Panel A) and victims (Panel B) following (3). Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.01).*

Table D16: Heterogeneous Effects on Bully-Victims and Bystanders

	Panel A. Bully-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<0.01).*

Table D17: Heterogeneous Effects on Empathy

	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<0.01).*

Table D18: Effects on Bullying: Accounting for Misreporting

	Bully		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 effects on 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 (i.e., reporting 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 the randomization strata. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.01).*

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

α_1	(1) Bully	(2) Victim	(3) 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 effects on bullying behaviors allowing for misreporting following Hausman et al. (1998). In particular, the model allows for the possibility of false negatives (i.e., reporting no bullying behaviors while being involved in bullying). We vary the level of misreporting α_1. All analyses additionally control for individual demographics and the randomization strata. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.01).*

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

	(1) Threatening	(2) Rumor spreading	(3) Physical bullying	(4) Social isolation	(5) Cyberbullying
α_1			Panel A: Bully		
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: 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: Bully-victim		
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 bullying behaviors by detailed type allowing for misreporting following Hausman et al. (1998). In particular, the model allows for the possibility of false negatives (i.e., reporting no bullying behaviors while being involved in bullying). We vary the level of misreporting α_1. All analyses additionally control for individual demographics and the randomization strata. Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.01).*

Table D21: Bullying Incidents Reported by Parents

	(1) Control Mean	(2) ITT	(3) permutation p-value	(4) wcb	(5) TOT
			Victim		
Victim (overall)	0.236	-0.031	0.136	0.135	-0.074
	(0.425)	(0.020)			(0.048)
Threatening	0.057	-0.011	0.381	0.310	-0.027
	(0.231)	(0.010)			(0.024)
Rumor spreading	0.132	-0.027*	0.086	0.088	-0.064*
	(0.339)	(0.014)			(0.035)
Physical bullying	0.125	-0.016	0.361	0.363	-0.039
	(0.331)	(0.017)			(0.039)
Social isolation	0.080	-0.019*	0.100	0.102	-0.045*
	(0.272)	(0.011)			(0.026)
Cyberbullying	0.037	-0.006	0.463	0.463	-0.015
	(0.188)	(0.008)			(0.019)
N	848	1,852			1,852

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 victims, and the first variable, "Victim (overall)," 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 the treatment assignment indicator as the instrument. (5) Classroom-level clustered standard errors are presented in parentheses (p<0.10, ** p<0.05, and *** p<0.01).*

Table D22: Robustness Analysis For Bullying Behaviors

	(1)	(2)	(3)	(4)	(5)	(6)
	Main	Demographic	SDS	Survey time	Pooled	Entropy Balance
Bullying involvements:						
Bully	-0.053**	-0.053*	-0.056**	-0.051**	-0.033***	-0.078**
	(0.025)	(0.028)	(0.026)	(0.023)	(0.009)	(0.038)
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)
Willing to help 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 the 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<0.01).*

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 in cm	162.8	161.8	162.8	163.3	14.1	0.7	-5.7
Weight in half kilo	100.0	101.6	100.0	100.5	-7.7	0.2	-2.3
Victim	0.7	0.7	0.7	0.7	-5.4	0.0	0.6
Bully	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
Perseverance	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 per 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 $SD_s = 100 \cdot (\overline{s_1} - \overline{s_0}) / \sqrt{0.5 \cdot (\sigma^2_{s1} + \sigma^2_{s0})}$*, where* $\bar{s}_1$ *and* s_0 *are the means of treated and controls, respectively, and* σ^2_{s1} *and* σ^2_{s0} *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)
	Bully	Victim	Bully-victim	Bully	Victim	Bully-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 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 indicator 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<0.01).*

Table D25: Baseline Characteristic Importance Ranking Predicted by GRF Analysis

	Variable importance rank				
	Bully	Victim	Bully-victim	Willing to 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.