Journal of School Psychology 99 (2023) 101226

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Effects of the KiVa anti-bullying program on defending behavior: Investigating individual-level mechanisms of change☆

Claire F. Garandeau [a,*], Tiina Turunen [a], Silja Saarento-Zaprudin [a], Christina Salmivalli [a,b]

[a] INVEST Flagship, University of Turku, Finland
[b] Department of Psychology, Shandong Normal University, Jinan, China

ARTICLE INFO

Editor: Craig A. Albers
Action Editor: Lyndsay Jenkins

Keywords:
Defending behavior
Anti-bullying intervention
Empathy
Self-efficacy
Outcome expectations
Responsibility to intervene

ABSTRACT

Given that defending victimized peers might help discourage bullying behavior and prevent its harmful consequences, various anti-bullying programs have attempted to increase defending behavior among participating children. However, the cognitions that underlie the effectiveness of interventions in increasing defending remain unknown. Data for this randomized controlled trial (RCT) of the KiVa anti-bullying program were collected in Finnish primary schools at baseline, after 5 months of implementation, and after 9 months of implementation and were used to examine the possible mediating role of seven psychological factors (empathy for the victim, feelings of responsibility to intervene, self-efficacy for defending, negative attitudes towards victims, and outcome expectations that defending would decrease or stop the bullying, be beneficial for one's status, and not increase one's risk of being victimized). Analyses conducted on a sample of 5731 children (baseline $M_{age} = 11$ years; 51% girls) revealed that the positive effects of KiVa on defending behavior after 9 months of implementation could partly be explained by the positive effects of the program on two factors (i.e., feelings of responsibility to intervene and expectations that the defending would make the bullying decrease or stop) after 5 months of implementation. This study provides information regarding the individual-level factors that anti-bullying interventions can target to successfully promote defending of victimized peers in primary schools.

1. Introduction

The initiation and maintenance of school bullying, defined as repeated aggression against a peer in a more vulnerable position, can be inhibited (or facilitated) by the behavior of those who are neither the instigators nor the targets of bullying behavior (i.e., the bystanders; e.g., Nocentini et al., 2013). Defending victimized peers, which can be done by comforting them, seeking help from adults, or directly attempting to stop the bullying (Lambe & Craig, 2020) is thought to help counteract bullying and its consequences.

☆ This research was supported by the INVEST Research Flagship Center (Academy of Finland Flagship Program, decision number: 320162) and by an ERC grant (Grant/Award Number: 884434) awarded to the last author. The last author led the development of the KiVa program; the other authors have no competing interests to declare.

* Corresponding author.
E-mail addresses: clagar@utu.fi (C.F. Garandeau), tmturu@utu.fi (T. Turunen), silsaar@utu.fi (S. Saarento-Zaprudin), christina.salmivalli@utu.fi (C. Salmivalli).

https://doi.org/10.1016/j.jsp.2023.101226
Received 6 September 2022; Received in revised form 10 March 2023; Accepted 27 June 2023
Available online 12 July 2023

Although more longitudinal investigations of the effects of defending on the prevalence of bullying and on the psychosocial adjustment of victimized youth are still needed (e.g., Healy, 2020; Laninga-Wijnen et al., 2023), there is evidence that higher rates of defending in the classroom are associated with lower levels of bullying, both concurrently (Kärnä et al., 2010; Salmivalli, Voeten, & Poskiparta, 2011) and over time (Nocentini et al., 2013), and with lower levels of internalizing problems among victimized children (Kärnä et al., 2010; Yun & Juvonen, 2020). In an observational study, Hawkins et al. (2001) indicated that peer interventions in favor of the victim were also found to put an end to bullying in 57% of bullying occasions. Moreover, victims with at least one defender have concurrently been reported to be better adjusted than non-defended victims (Sainio et al., 2011) and experience higher feelings of belonging 9 months later (Laninga-Wijnen et al., 2023).

The recognition that bystanders play a role in bullying events has provided additional opportunities for anti-bullying interventions. Instead of focusing intervention efforts on changing perpetrators and their victims, modifying the cognitions of bystanders so that they start defending victimized peers appears as a promising strategy to counter bullying (Salmivalli, 2014). The anti-bullying program KiVa, which has been shown to be effective at reducing bullying in several countries (e.g., Garandeau & Salmivalli, 2018; Huitsing et al., 2020; Kärnä et al., 2011; Nocentini & Menesini, 2016), was designed to enhance bystanders' support of victimized peers and deter them from reinforcing the behavior of "ring-leader bullies." After 5 months of implementation, it was found to increase defending in Finnish primary schools ($b = 0.110$, $p = .046$; Kärnä et al., 2011). Other anti-bullying programs that encourage children witnessing bullying to stand up for the victim have been found to increase prosocial peer interventions in situations of bullying (Hedges's $g = 0.20$, 95% confidence interval [CI] = 0.11–0.29, $p < .001$; Polanin et al., 2012).

However, increasing the defending behavior of bystanders is challenging. Fear of becoming the next victim prevents many potential defenders from acting (e.g., Lodge & Frydenberg, 2005; Thornberg et al., 2018) and partly explains why the percentage of students who actively defend victims is estimated to be no higher than 20% (Pouwels et al., 2018; Salmivalli et al., 1996). To develop interventions that are more effective at promoting defending of victimized children, additional knowledge is needed on the mechanisms through which anti-bullying programs foster defending behavior. Although a myriad of studies identified concurrent predictors of defending (see Ma et al., 2019, Lambe et al., 2019, and Zych et al., 2019, for meta-analyses), less is known on the factors that predict increases in defending over time. Moreover, most studies evaluating anti-bullying interventions have examined their effects on bullying and victimization but have less frequently considered their effects on defending and on the cognitions and attitudes that might underlie the effectiveness of the program in increasing defending.

Using data from a randomized control trial (RCT) of the Finnish anti-bullying program KiVa in primary schools, this study aimed to identify how the intervention successfully increased defending behavior among school children by investigating the possible mediating role of seven psychological factors, including (a) affective empathy for the victim, (b) feelings of responsibility to intervene, (c) self-efficacy for defending, (d) negative attitudes towards victims, and three expectations about the outcomes of defending: (e) it will make the bullying stop or decrease, (f) it will make the defender popular, and (g) it will increase the risk of being victimized for the defender. These psychological factors were selected because KiVa was designed to influence them and there is empirical evidence suggesting and/or theoretical reasons to expect that these factors are associated with defending behavior. Other factors, in particular contextual factors, also have been suggested to promote defending. These include, for example, classroom anti-bullying norms and pro-defending norms, the quality of teacher-student relationships, and students' level of popularity and likeability among their peers (e.g., Lambe et al., 2019). However, the mechanisms underlying the effects of these contextual factors on defending are assumed to be psychological factors, such as outcome expectations or self-efficacy for defending. For this reason, the present study focused on the possible mediating role of seven cognitions and attitudes, which should be more directly linked to behavioral change (e.g., Bandura, 1986).

2. Hypothesized mediating factors

2.1. *Affective empathy for the victim*

Affective empathy is defined as an emotional response that stems from another's emotional state or condition and is congruent with the other's emotional state or condition (Eisenberg et al., 1991); it also is one of the most often studied predictors of defending in situations of bullying. Affective empathy is a capacity believed to be necessary for prosocial behavior and has been found to be positively related to interpreting the situation as an emergency and the corresponding act of intervening (Fredrick et al., 2020). According to six systematic reviews and meta-analyses (i.e., Deng et al., 2021; Lambe et al., 2019; Ma et al., 2019; Nickerson et al., 2015; Van Noorden et al., 2015; Zych et al., 2019), those who defend victimized peers have higher levels of affective empathy. Although the average effect size varies across meta-analytic studies (from $r = 0.15$ in Ma et al., 2019, to $r = 0.33$ in Nickerson et al., 2015), this positive association is consistently found across individual studies and meta-analyses.

The KiVa program targets all students and includes several components designed to increase affective empathy for victims of bullying. During KiVa lessons, students are taught how painful the consequences of bullying are for victims and perform exercises in which they reflect on the thoughts and feelings of victimized peers. They also watch filmed interviews of adults who were victimized at school, explaining how it felt and the negative impact the experience has had on their lives. Furthermore, in the KiVa program's online game, which is an online environment where participants are introduced to a virtual school where they can witness bullying incidents with animated characters in the roles of perpetrators, victims, or outsiders, participants are shown – via thought bubbles on the screen – the sad thoughts and emotions of the victim. As KiVa has been found to have a positive effect on affective empathy for victims after 5 months (Saarento et al., 2015) and 9 months of program implementation (Garandeau, Laninga-Wijnen, & Salmivalli, 2022), we expected an indirect effect of KiVa on defending via affective empathy.

2.2. Responsibility to intervene

Early research on the bystander effect among adults has shown that individuals are less likely to help a victim in the presence of other people because they assume that someone else will take responsibility for intervening (Darley & Latané, 1968). In the situational model of bystander behavior (SMB; Latané & Darley, 1970), which has been shown to be relevant for incidents of school bullying (Jenkins et al., 2018; Jenkins & Nickerson, 2017), seeing oneself as responsible for intervening in a troubling situation is one of the key processes underlying the decision to intervene. Numerous studies have shown that feeling responsible for intervening predicts defending in situations of bullying; for instance, a positive association was found between Italian adolescents' sense of responsibility to intervene in favor of the victim and self-reported defending behavior ($r = 0.40$; Pozzoli & Gini, 2010). Moreover, among Canadian (Cappadocia et al., 2012) and Flemish youth (Van Cleemput et al., 2014), the belief that it was not their "place", their "business", or their responsibility to intervene, acted as a barrier to defending victims of bullying. Additionally, in a sample of American adolescents, feelings of responsibility to intervene were found to be strongly correlated ($r = 0.80$) with intention to intervene in a bullying context (Nickerson et al., 2014).

The various components of the KiVa program emphasize that bullying incidents are not only the concern of perpetrators; rather, everyone plays an important role in bullying incidents. In student lessons, all students receive the message that the silence and passivity of witnesses of bullying contribute to the perpetuation of the behavior and that it is everyone's responsibility to do something, even if only reporting the bullying to an adult. We expected that KiVa would increase students' perceptions that it is the responsibility of bystanders to intervene, which in turn would predict the positive effects of the program on defending behavior.

2.3. Negative attitudes towards victims

The perceptions that bystanders have of their victimized classmates should logically contribute to their willingness to defend them. Perceiving victims as deserving of their plight and as weak individuals simply asking for trouble is likely to prevent any action to intervene. Although few studies have investigated the specific role of negative attitudes towards victims on the likelihood to defend, a body of research on moral disengagement, which refers to a set of cognitive mechanisms by which people convince themselves that ethical standards do not apply to them in a particular context (Bandura, 2016), provides insight into this association. Indeed, these mechanisms include blaming the victim (e.g., endorsing statements that people who are weird are to blame for being bullied), moral justification of negative behavior, and minimizing the consequences of the behavior. Numerous studies have shown that children and adolescents with high levels of moral disengagement were less likely to defend victims of bullying (Gini, 2006; Jiang et al., 2022; Mazzone et al., 2016; Pozzoli et al., 2016; Thornberg et al., 2015, 2017, 2022).

The KiVa program teaches students that (a) bullying is never acceptable and (b) everyone must be respected regardless of their differences. KiVa also exposes participants to adults reflecting on their former experiences of peer victimization in school. We expected that exposure to KiVa would decrease negative attitudes towards victims (e.g., perceptions that vulnerable children deserve to be maltreated) which in turn would lead to more defending behavior.

2.4. Self-efficacy for defending

Within social cognitive theory, perceived self-efficacy refers to individuals' beliefs about their own capabilities to execute a particular action successfully; in turn, these beliefs influence how individuals behave (Bandura, 1994). As defending victimized peers is not always an easy behavior to engage in, researchers have theorized that a high level of self-efficacy for defending would be a key predictor of the behavior. There is strong empirical support for the idea that children and adolescents are unlikely to defend victimized peers if they lack defender self-efficacy (DeSmet et al., 2016; Doramajian & Bukowski, 2015; Gini et al., 2008; Peets et al., 2015; Pöyhönen et al., 2010, 2012; Sjögren et al., 2020; Thornberg et al., 2017, 2020; Thornberg & Jungert, 2013). Two longitudinal studies also have shown that higher levels of defending self-efficacy led to higher levels of defending 5 months (Gini et al., 2022) and 7 months later (van der Ploeg et al., 2017) among youth.

Through student lessons, role-play exercises, and an online game, the KiVa program demonstrates various means for participants to defend victims and the positive consequences of such actions. Thus, students get to practice defending behaviors in a safe environment. The program emphasizes that it is possible to defend in ways that are relatively easy to carry out, such as reporting the incident to a teacher. As a result, participants should feel more confident in their abilities to take actions against bullying. Therefore, we hypothesized that the effects of KiVa on defending behavior would be mediated by increases in self-efficacy for defending.

2.5. Outcome expectations for defending

According to social cognitive theory (Bandura, 1994), feelings of self-efficacy and engagement in a particular behavior are determined by the consequences that individuals expect for engaging in the behavior. These outcome expectations are defined as personal beliefs about the likelihood of their behavior leading to a specific outcome. Consequently, when students expect positive outcomes, such as believing that one's defending behavior will make the bullying decrease, stop, or will otherwise benefit one's own status, they should be more likely to defend. Conversely, when they expect negative outcomes, such as believing that defending will make them a target for bullying, they should be less likely to defend. Research on the effects of outcome expectations on defending behavior in situations of bullying is scarce. To our knowledge, the only study to examine these associations was conducted with cross-sectional data collected for the evaluation of KiVa and revealed that children were more likely to defend if they expected that

defending would make the bullying decrease and would improve their own status (Pöyhönen et al., 2012). As students are hesitant to defend victimized peers due to fear of becoming victimized (e.g., Thornberg et al., 2018), the present study considered the expectation of this negative outcome, the expectation that the bullying would decrease, and the expectation that defending would benefit one's status.

The KiVa program promotes defending as an effective way to put an end to bullying and should therefore increase the expectation that it will help decrease or stop the bullying. It also depicts defending behavior in a positive light by describing how other behaviors (e.g., assisting the bullies or remaining passive) often result from succumbing to the pressure of the group. Highlighting the weaknesses of these other behaviors, by contrast, underlines the strength required for defending and may reinforce the belief that defending will heighten one's status. Finally, as KiVa is implemented as a whole-school program and involves most adults in the school, it should make potential defenders feel that they will be protected in their defending attempts and be less afraid of being bullied in retaliation.

2.6. The present study

To develop more effective interventions, information on the mechanisms leading to desired outcomes is urgently needed. Saarento et al. (2015) identified various individual-level mechanisms explaining how the KiVa program leads to reductions in bullying perpetration, namely increases in antibullying attitudes, perceiving an increased number of classmates as defending the victim, and perceiving the teacher as being more disapproving of bullying than before. Although encouraging defending behavior has become one of the major objectives of whole-school anti-bullying programs, tests of the effectiveness of such programs do not always include defending as an outcome and no information is available on the mechanisms through which these interventions might lead students to more frequently defend victimized peers.

We hypothesized that KiVa would lead to increased defending after 9 months of implementation (T3) by influencing several factors after 5 months of implementation (T2), including students' (a) affective empathy for the victim, (b) feelings of responsibility to intervene, (c) self-efficacy for defending, (d) negative attitudes towards victims, (e) expectations that defending would make the bullying decrease, (f) expectations that defending would be beneficial for their own status, and (g) expectations that defending would not increase their risk of being victimized.

3. Method

3.1. Participants

This study used three waves of data collected in 77 Finnish primary schools (39 intervention schools, 38 control schools) for the RCT evaluation of the KiVa anti-bullying program (Kärnä et al., 2011). Among schools providing basic education that had volunteered to participate in the RCT, stratified random sampling was used to include schools from all the provinces of mainland Finland (see Kärnä et al., 2011, for a detailed description).

Among classrooms already participating at the first wave (T1; pretest data), our initial sample included 7357 children participating in at least one of the three waves (including 5255 participating in all three waves). The percentage of parents providing consent for their children to participate was 91%. As the main variable of interest (i.e., defending behavior) was assessed via within-classroom peer nominations, we ensured good reliability of the variable by selecting classrooms with at least 10 students and a minimum 40% participation rate at each wave, resulting in a sample of 5731 participants ($M_{age} = 11$ years, $SD = 1.10$; 51% girls) from 172 intervention and 125 control classrooms. At T1, 33% of the participants were in Grade 3, 33% in Grade 4, and 36% in Grade 5. Ten percent of the participants attended Swedish-speaking schools and 2% were immigrants. In this sample, the percentage of parents providing consent was 96%.

3.2. Procedure

Active consent forms were translated into 15 languages and distributed to the parents of all the students in the target sample. The first wave (T1; pretest data) was collected at the end of the school year (May 2007) when participants (with both parental consent and their own assent for participating) were in Grades 3–5. The second wave (T2) was collected in the winter of the following school year after 4–5 months of program implementation. The third wave (T3) was collected after 9 months of program implementation and occurred approximately 1 year after the first data collection wave. At T2 and T3, participants were in Grades 4–6. In Finnish primary schools, children typically remain in the same classroom from one year to the next during these grades. The overall implementation level of the KiVa curriculum (i.e., student lessons) during the RCT was classified as relatively good. For instance, primary school teachers implemented an average of 8.7 out of 10 lessons and an average of 68% of lesson content was delivered, with a mean duration of a lesson being 79 min (Haataja et al., 2014; Salmivalli, Haataja, & Poskiparta, 2011). Almost 80% of the teachers were female and their average teaching experience was 15 years (Haataja et al., 2015).

Data were collected through online questionnaires that were completed during regular school hours in the school computer lab under the supervision of teachers. The questionnaires were either in Finnish or in Swedish, depending on the language of the school. The teachers were given detailed instructions about the procedure 2 weeks prior to data collection. In addition, the teachers were provided with an option of receiving support through phone or email prior to and during the data collection in case they had any questions. The order of the questionnaires presented to students, as well as the order of the items within questionnaires, were randomized. In accordance with the Declaration of Helsinki, all parents gave written informed consent and participating children

provided their assent. When the KiVa research project began, neither institutional nor national guidelines required an ethics approval for non-invasive questionnaire studies. Nevertheless, this study was conducted in accordance with the recommendations of the Ethics Board of the University of Turku, Finland.

3.3. Measures

3.3.1. Defending behavior

At the three data collection waves, participants were asked to nominate the classmates who fit various descriptions of bystander behaviors. Defending was assessed with three items from the Participant Role Questionnaire (Salmivalli & Voeten, 2004): (a) "He/She comforts the victim or encourages him/her to tell the teacher about the bullying"; (b) "He/She tells the others to stop bullying"; and (c) "He/She tries to make the others stop bullying". Participants were able to nominate an unlimited number of classmates. For each item, the total number of nominations received by each student was summed and divided by the number of possible nominators. Composite scores were created by averaging across the three items resulting in proportion scores for defending that ranged from 0 to 1 (T1 $\alpha = 0.92$; T2 $\alpha = 0.92$; T3 $\alpha = 0.94$).

3.3.2. Affective empathy for the victim

Four self-report items designed for the evaluation of the KiVa program were used to measure students' affective empathy for victimized peers: (a) "When the bullied student starts to cry, I also feel bad"; (b) "When someone is bullied, I start to get angry on his/her behalf"; (c) "When the bullied student feels sad, I want to comfort him/her"; and (d) "When the bullied student is sad, I also feel sad". Responses were provided on a 4-point scale ranging from 0 (*Never*) to 3 (*Always*). The total score was the average score from the four items. The Cronbach's alphas were high at T1 ($\alpha = 0.81$) and T2 ($\alpha = 0.84$).

3.3.3. Responsibility to intervene

Children's feelings of responsibility to intervene in bullying situations were measured by four items specifically created for the evaluation of the KiVa program: (a) "If someone bullies another, others don't have to care about it" (reverse-coded); (b) "It is everyone's collective task to prevent bullying in class"; (c) "If someone bullies another, that issue is just between them" (reverse-coded); and (d) "Everyone should for his/her part take care that nobody gets bullied". Responses were provided on a 5-point scale ranging from 0 (*Completely disagree*) to 4 (*Completely agree*). The scores across the four items were averaged to create a composite score at T1 and T2, with higher scores on the scale reflecting stronger feelings of responsibility to intervene (T1 $\alpha = 0.58$; T2 $\alpha = 0.67$).

3.3.4. Negative attitudes towards victims

Three items that were adapted from the Provictim Scale (Rigby & Slee, 1991) measured participants' negative attitudes towards victimized children. Participants were asked to rate the extent to which they agreed with the following items: (a) "Kids who are weak are just asking for trouble", (b) "Soft kids make me sick", and (c) "Nobody likes a wimp". Responses were given on a scale ranging from 0 (*Completely disagree*) to 4 (*Completely agree*). Higher scores on these items indicate more negative attitudes towards victims. The scores for the three items were averaged at T1 ($\alpha = 0.64$) and T2 ($\alpha = 0.69$).

3.3.5. Self-efficacy for defending

Participants' beliefs in their capacity to defend victims of bullying were assessed with three items previously used by Pöyhönen et al. (2010). Participants were asked to rate how easy or difficult it would be for them to behave in the following ways: (a) "Trying to get others to stop bullying", (b) "Comforting the bullied person or encouraging him/her to report the bullying to the teacher", and (c) "Asking others to stop bullying or saying that bullying is stupid". Answers were given on a 4-point scale ranging from 0 (*Very easy*) to 3 (*Very difficult*). The three items were reverse coded so that higher scores would reflect higher self-efficacy; items then were averaged to form a composite score (T1 $\alpha = 0.69$; T2 $\alpha = 0.74$).

3.3.6. Outcome expectations for defending behavior

Three types of outcome expectations for defending behavior were assessed using items created specifically for the evaluation of the KiVa program. First, children were asked to imagine a situation when one of their classmates is bullied and to rate the likelihood of various consequences they would expect for three types of defending behavior they might engage in, including (a) if they tried to stop the bully, (b) if they comforted the bullied person or told them to report the bullying to a teacher, and (c) if they asked others to stop the bullying or said that bullying is stupid. Each item measuring an outcome expectation for defending was repeated for each of the three types of defending. Outcome expectations that defending would make the bullying stop or decrease were assessed by two items: (a) "It would end or decrease the bullying" or (b) "It would increase the bullying" (reverse-coded). Responses were provided on a 4-point scale ranging from 0 (*Not likely at all*) to 3 (*Very likely*). The total score for outcome expectations that the bullying would stop was computed by averaging the scores of the six items (two for each type of defending; T1 $\alpha = 0.75$; T2 $\alpha = 0.76$). Outcome expectations that defending would benefit the status of the defender was assessed by the item "It would make the others think highly of you (you would be valued)". Responses were provided on a 4-point scale ranging from 0 (*Not likely at all*) to 3 (*Very likely*). The total score for outcome expectations that defending would benefit the defender's status was computed by averaging the scores of the three items (one for each type of defending; T1 $\alpha = 0.80$; T2 $\alpha = 0.83$). Outcome expectations that defending would increase risks of being victimized for the defender was assessed by the item "It would make you unpopular and you would be bullied". Responses were provided on a 4-point scale ranging from 0 (*Not likely at all*) to 3 (*Very likely*). The total score for outcome expectations that defending would increase risks of

being victimized was computed by averaging the scores of the three items (one for each type of defending; T1 α = 0.79; T2 α = 0.82). This variable was reverse coded so that higher scores indicated that the participant found it not likely that their defending would make them unpopular and victimized.

3.4. Analysis plan

To determine whether the implementation of the KiVa program led to increases in defending at T3 via the effects of KiVa on seven hypothesized mediators at T2, a series of structural equation models were estimated in *Mplus 8.6* (Muthén & Muthén, 1998–2017). As testing all seven hypothesized, correlated mediating factors in the same model might result in a loss of power to detect indirect effects (Hayes, 2017), we decided to only include in our final mediation model factors that were shown to be either significantly affected by the KiVa program at T2 (controlling for T1 levels) or that significantly predicted defending at T3 (controlling for T2 defending). Therefore, we first conducted models testing these effects. All models were run using maximum likelihood estimation with robust standard errors (MLR) and accounting for T1 within-classroom dependence. Several indices were used to evaluate model fit based on recommended cut-off criteria (Hu & Bentler, 1999), including the Comparative Fit Index (CFI; > 0.95), the Tucker-Lewis Index (TLI; > 0.90), the Root Mean Square Error of Approximation (RMSEA; < 0.06), and the Standardized Root Mean Square Residual (SRMR; < 0.08). Finally, the significance of the indirect effects of KiVa on defending via the hypothesized mediators was tested with bias-corrected bootstrap confidence intervals (95%) for indirect and direct effects, using 10,000 bootstrap draws. Indirect and direct effects were considered significant when the value of zero was outside of the range of confidence intervals (Preacher & Hayes, 2004).

First, we tested whether the implementation of KiVa had a positive effect on defending behavior at T2 and at T3, controlling for age, gender, and defending at T1 (Model 1). Second, we examined the effects of KiVa on all hypothesized mediators at T2, controlling for age, gender, and hypothesized mediators at T1 (Model 2). All T1 variables were allowed to correlate. Third, we examined the effects of the seven hypothesized mediating variables at T2 on defending at T3, controlling for T2 defending, KiVa implementation, age, and gender (Model 3). Finally, the hypothesized mediators, which were either significantly predicted by KiVa or significantly predicted defending at T3, were included in a model designed to examine their mediating role in the association between KiVa implementation and T3 defending (Model 4; see Fig. 1). The key hypotheses of the study, which predicted that the KiVa program would have a positive effect on defending via its effects on these seven psychological factors, were tested in this fourth model.

This final model (Model 4) was an autoregressive panel model in which any variable regressed on itself at a previous time point represents change in that variable. Our model included (a) regression paths from KiVa to all variables at T2 and on defending at T3; (b) autoregressive paths between T1 and T2 measurements of the mediators and defending, and between T2 and T3 measurements of defending; (c) within-time correlations (T1) and residual correlations (T2) among the mediators; (d) cross-lagged associations between all mediators and defending at T1 and T2; and (e) regression paths from the mediators to defending at T3. To account for T1 variables that may have differed between KiVa and control schools, our model included correlations between the following variables: KiVa

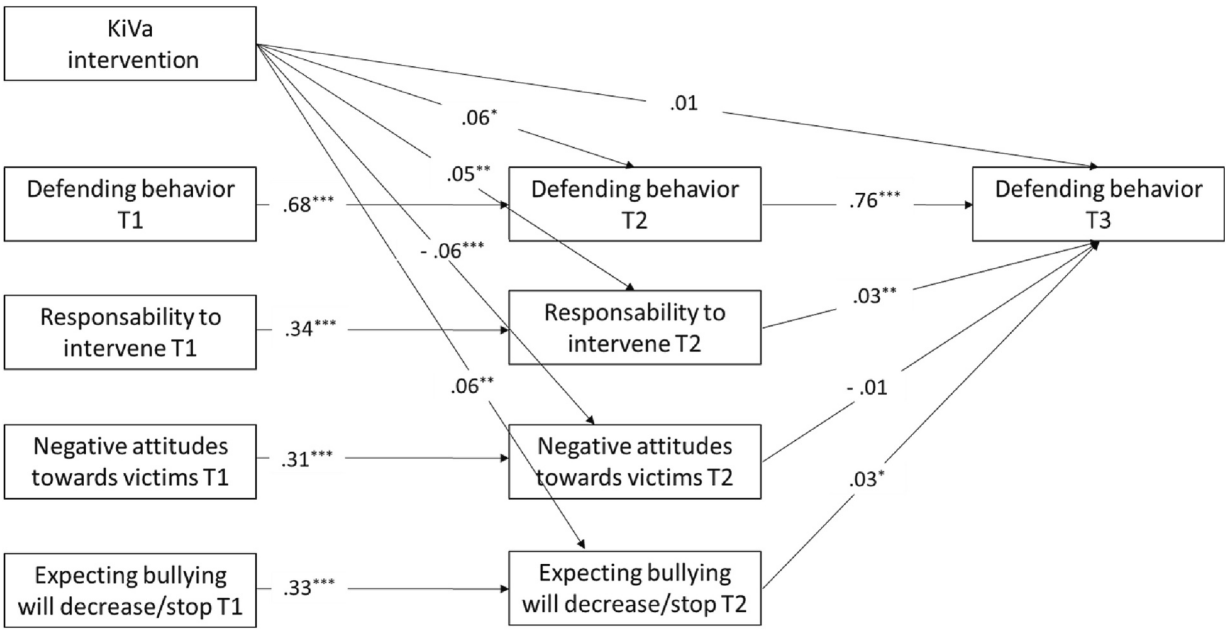

Fig. 1. Standardized coefficients for the final model testing for the indirect effects of KiVa on T3 defending. Within-timepoint correlations (T1), residual correlations (T2), and cross-lagged paths between T1 and T2 variables were estimated. Effects of age and gender on all T1, T2 and T3 variables were controlled for.
Note. *$p < .05$, **$p < .01$, ***$p < .001$.

Table 1
Correlations among defending at the three time points and seven potential mediators at T1 and T2.

	1.	2	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
1. Defending T1	–																
2. Defending T2	0.75	–															
3. Defending T3	0.69	0.81	–														
4. Empathy T1	0.33	0.31	0.29	–													
5. Empathy T2	0.30	0.34	0.31	0.53	–												
6. Responsibility T1	0.25	0.22	0.21	0.42	0.33	–											
7. Responsibility T2	0.25	0.28	0.28	0.35	0.45	0.45	–										
8. Neg. att. vict T1	−0.23	−0.23	−0.20	−0.32	−0.27	−0.45	−0.34	–									
9. Neg. att. vict T2	−0.26	−0.28	−0.26	−0.28	−0.35	−0.33	−0.54	0.43	–								
10. Self-efficacy T1	0.13	0.15	0.16	0.17	0.21	0.13	0.15	−0.11	−0.13	–							
11. Self-efficacy T2	0.16	0.18	0.16	0.18	0.22	0.16	0.24	−0.14	−0.21	0.38	–						
12. Exp. decrease T1	0.14	0.14	0.14	0.17	0.14	0.22	0.19	−0.20	−0.17	0.15	0.19	–					
13. Exp. decrease T2	0.16	0.19	0.19	0.16	0.24	0.17	0.28	−0.18	−0.25	0.20	0.27	0.37	–				
14. Exp. status T1	0.10	0.12	0.12	0.19	0.14	0.14	0.14	−0.09	−0.08	0.10	0.13	0.33	0.22	–			
15. Exp. status T2	0.10	0.13	0.14	0.13	0.23	0.11	0.19	−0.09	−0.09	0.17	0.18	0.21	0.40	0.32	–		
16. Exp. fear T1	0.10	0.09	0.08	$.02_b$	$.04_a$	0.11	0.12	−0.19	−0.13	0.13	0.17	0.47	0.28	0.12	0.16	–	
17. Exp. fear T2	0.13	0.13	0.13	0.05	0.06	0.10	0.14	−0.14	−0.22	0.18	0.28	0.25	0.49	0.13	0.21	0.37	–
$M_{intervention}$	0.20	0.22	0.19	1.85	1.84	3.04	3.05	1.08	1.09	1.89	1.80	1.93	1.86	1.59	1.57	1.86	1.79
$SD_{intervention}$	0.14	0.15	0.15	0.73	0.73	0.76	0.77	0.94	0.93	0.72	0.73	0.59	0.58	0.77	0.75	0.80	0.81
$M_{control}$	0.20	0.20	0.17	1.83	1.79	2.98	2.95	1.13	1.22	1.83	1.80	1.91	1.78	1.60	1.54	1.81	1.72
$SD_{control}$	0.14	0.14	0.14	0.72	0.76	0.78	0.81	0.92	0.96	0.72	0.74	0.60	0.58	0.77	0.76	0.80	0.81

Note. All correlations were significant at $p < .001$, except correlations with subscript a ($p < .01$) and b ($p > .05$). Neg. att. Vict = Negative attitudes towards victims; Exp decrease = Outcome expectation that defending will make the bullying decrease or stop; Exp status = Outcome expectations that defending would benefit the defender's status; Exp fear = Outcome expectations that defending would not increase risks of being victimized.

intervention (vs control), T1 defending, and all hypothesized mediators at T1. In addition, the effects of demographic variables (i.e., age and gender of the students) were controlled for by correlating them with all T1 variables and regressing the mediator and outcome variables at T2 and T3 on them.

The percentage of missing data was low for most of the key variables. For defending, it was <0.1% across the three time points. For all other variables, it ranged from 4.5% to 6.8% across T1 and T2, except for responsibility to intervene at T1, which had 20.9% missingness due to a technical issue that led to missing data from some schools. Therefore, we assumed the data to be missing at random (MAR) and used full information maximum likelihood estimation (FIML) in order to use all available data to estimate the model without imputing it.

4. Results

4.1. Descriptive statistics

Table 1 provides the means, standard deviations, and correlations for all study variables, computed separately for KiVa intervention schools and for control schools. All hypothesized mediators at T1 and T2 were significantly and positively correlated with defending behavior at each time point. Independent-sample *t*-tests indicated that intervention schools at T1 were higher on mean levels of responsibility to intervene, $t = 2.47$, $p = .014$, self-efficacy, $t = 3.31$, $p < .001$, and outcome expectations that defending would not increase risks of being victimized, $t = 2.12$, $p = .034$. At T2, mean level of negative attitudes towards victims was lower in intervention schools and mean levels of other variables (responsibility to intervene, affective empathy, outcome expectations that defending would decrease bullying and would not increase risks of being victimized) were higher in intervention schools (all *p*-values < .02), except for outcome expectations that defending would benefit the defender's status, $t = 1.33$, $p = .183$, and self-efficacy for defending, $t = 0.85$, $p = .932$.

4.2. Effects of KiVa on defending and hypothesized mediators

In Model 1, we tested whether the KiVa program had a significant effect on defending after 5 months (T2) and 9 months (T3) of program implementation, controlling for baseline levels of defending behavior, age, and gender. Results of this saturated model are shown in Table 2. Consistent with our expectations, KiVa had a positive effect on defending behavior at T2 ($p = .008$) and at T3 ($p = .009$). When both T1 and T2 defending were controlled for at the same time, the effect of KiVa on T3 defending was no longer significant, $\beta = 0.009$, SE = 0.006, $p = .154$. However, the indirect effect of KiVa on T3 defending via T2 defending was significant, est. = 0.040, $SE = 0.015$, $p = .008$.

The effects of KiVa on all hypothesized mediators at T2 were tested in Model 2 (see Table 3). The model fit was adequate, with $\chi^2(44) = 695.071$, $p < .001$, RMSEA = 0.051, CFI = 0.937, TLI = 0.871, SRMR = 0.057. To improve model fit, we added four paths suggested by modification indices, including (a) effects of negative attitudes towards victims and affective empathy for the victim at T1 on responsibility to intervene at T2 and (b) effect of responsibility to intervene at T1 on negative attitudes towards victims and affective empathy for the victim at T2. The fit of the final model fit was good, $\chi^2(40) = 490.766$, $p < .001$, RMSEA = 0.044, CFI = 0.957, TLI = 0.901, SRMR = 0.048. The results indicated that KiVa had a positive effect on T2 responsibility to intervene ($p = .001$), negative attitudes towards victims ($p < .001$), and outcome expectations that the bullying would decrease or stop ($p = .001$).

4.3. Effects of hypothesized mediators on T3 defending

Model 3 examined the effects of each hypothesized mediator at T2 on T3 levels of defending behavior. The results are presented in Table 4. In this saturated model, defending at T3 was significantly lower among boys and older children and positively predicted by T2 defending. Among the hypothesized mediators, only T2 responsibility to intervene had a statistically significant positive effect on defending at T3 ($p = .014$).

4.4. Indirect effects of KiVa on T3 defending

Based on the results of Model 2 and Model 3, we included the following three hypothesized mediators in the final model (Model 4) testing for the indirect effects of KiVa on T3 defending: (a) responsibility to intervene, (b) outcome expectations that the bullying

Table 2
Model 1: Unstandardized estimates for the effects of KiVa on defending at T2 and T3 accounting for within-classroom dependence and covariance between the two outcomes ($N = 5731$).

		Defending T2			Defending T3	
	Est (*SE*)	95% CI	*p*	Est (*SE*)	95% CI	*p*
Age	−0.013 (0.004)	[−0.021, −0.006]	< 0.001	−0.018 (0.004)	[−0.026, −0.010]	< 0.001
Boy	−0.035 (0.005)	[−0.045, −0.026]	< 0.001	−0.034 (0.005)	[−0.044, −0.023]	< 0.001
KiVa intervention	0.018 (0.007)	[0.005, 0.031]	0.008	0.021 (0.008)	[0.005, 0.037]	0.009
Defending T1	0.705 (0.027)	[0.653, 0.758]	< 0.001	0.652 (0.030)	[0.592, 0.712]	< 0.001

Table 3
Model 2: Unstandardized estimates for the effects of KiVa on T2 potential mediators controlling for age, gender, outcome at T1, within-classroom dependence, and covariation among all hypothesized mediators at T1 and T2 (N = 5731).

	Predictors											
	KiVa			Age			Boy			Outcome at T1		
T2 Outcomes	Est (*SE*)	95% CI	*p*	Est (*SE*)	95% CI	*p*	Est (*SE*)	95% CI	*p*	Est (*SE*)	95% CI	*p*
Empathy	0.04 (0.02)	[−0.01, 0.08]	0.112	−0.06 (0.01)	[−0.09, −0.04]	< 0.001	−0.22 (0.02)	[−0.26, −0.19]	< 0.001	0.41 (0.02)	[0.38, 0.44]	< 0.001
Responsibility	0.08 (0.03)	[0.03, 0.13]	0.001	−0.05 (0.01)	[−0.08, −0.02]	< 0.001	−0.16 (0.02)	[−0.20, −0.12]	< 0.001	0.32 (0.02)	[0.29, 0.36]	< 0.001
Neg. att. Vict.	−0.11 (0.03)	[−0.17, −0.06]	< 0.001	0.04 (0.02)	[0.01, 0.07]	0.010	0.32 (0.03)	[0.27, 0.37]	< 0.001	0.30 (0.02)	[0.27, 0.33]	< 0.001
Self-efficacy	−0.02 (0.02)	[−0.07, 0.02]	0.319	0.00 (0.01)	[−0.02, 0.03]	0.884	−0.12 (0.02)	[−0.16, −0.09]	< 0.001	0.33 (0.02)	[0.30, 0.36]	< 0.001
Exp. decrease	0.07 (0.02)	[0.03, 0.11]	0.001	−0.04 (0.01)	[−0.06, −0.02]	< 0.001	−0.10 (0.02)	[−0.13, −0.07]	< 0.001	0.26 (0.01)	[0.24, 0.29]	< 0.001
Exp. status	0.03 (0.02)	[−0.02, 0.08]	0.206	−0.02 (0.01)	[−0.05, 0.00]	0.107	−0.07 (0.02)	[−0.11, −0.03]	0.001	0.26 (0.01)	[0.23, 0.28]	< 0.001
Exp. fear	0.05 (0.03)	[−0.01, 0.10]	0.087	−0.02 (0.01)	[−0.04, 0.01]	0.260	−0.10 (0.02)	[−0.14, −0.06]	< 0.001	0.29 (0.01)	[0.27, 0.32]	< 0.001

Note. Neg. att. Vict = Negative attitudes towards victims; Exp decrease = Outcome expectation that defending will make the bullying decrease or stop; Exp status = Outcome expectations that defending would benefit the defender's status; Exp fear = Outcome expectations that defending would not increase risks of being victimized.

Table 4
Model 3: Unstandardized estimates for the effects of T2 defending and T2 potential mediators on T3 defending controlling for age, gender, KiVa intervention, within-class dependence, and covariation among all mediators (N = 5731).

		Defending at T3	
Predictors at T2	Est *(SE)*	95% CI	*p*
Age	−0.008 (0.003)	[−0.015, −0.002]	0.007
Boy	−0.016 (0.004)	[−0.024, −0.007]	< 0.001
KiVa intervention	0.006 (0.006)	[−0.007, 0.018]	0.353
Defending at T2	0.774 (0.022)	[0.731, 0.817]	< 0.001
Affective empathy	0.001 (0.002)	[−0.002, 0.005]	0.426
Responsibility to intervene	0.005 (0.002)	[0.001, 0.009]	0.014
Negative attitudes towards victims	−0.002 (0.002)	[−0.005, 0.001]	0.191
Self-efficacy for defending	0.000 (0.002)	[−0.004, 0.003]	0.919
Outcome expectations decrease	0.004 (0.003)	[−0.001, 0.010]	0.128
Outcome expectations status	0.003 (0.002)	[−0.001, 0.007]	0.127
Outcome expectations fear	0.001 (0.002)	[−0.003, 0.005]	0.588

Note. Outcome expectations decrease = Outcome expectations that defending will make the bullying decrease or stop; Outcome expectations status = Outcome expectations that defending would benefit the defender's status; Outcome expectations fear = Outcome expectations that defending would not increase risks of being victimized.

would decrease or stop, and (c) negative attitudes towards victims. Standardized regression coefficients for this model are displayed in Fig. 1. Model fit (without bootstrapping) was good, $\chi^2(6) = 42.755$, $p < .001$, CFI = 0.1.00, TLI = 0.97, RMSEA = 0.03, SRMR = 0.01. Results with bias-corrected bootstrap confidence intervals (95%) indicated that there were significant indirect effects of KiVa on T3 defending via T2 responsibility to intervene (95% CI = [0.00011, 0.00097]) and via T2 outcome expectations that bullying would decrease or stop (95% CI = [0.00010, 0.00096]). However, the indirect effect of KiVa on T3 defending via T2 negative attitudes towards victims was not statistically significant (95% CI = [−0.00009, 0.00067]). The results remained similar when the model was run without the direct effect of KiVa on T3 defending.

5. Discussion

Along with decreasing bullying behavior directly, encouraging all children to stand up for victimized peers has become a key objective of many intervention programs. However, the effects of school-based bullying prevention and intervention programs on bystander behaviors and on the psychological factors that may play a role in these behaviors are not often investigated. One of the main obstacles to improving the effectiveness of anti-bullying interventions may be our lack of knowledge regarding the causal mechanisms that are associated with changes in the desired outcomes (the so-called "black box", see Harachi et al., 1999). Saarento et al. (2015) previously had uncovered that the reductions in peer-reported bullying behavior obtained by the KiVa anti-bullying program were mostly due to changes in students' antibullying attitudes and their collective perceptions of teacher attitudes towards bullying. The present study also used the three waves of data collected for the RCT of the KiVa program to examine, for the first time, possible mechanisms explaining the effects of an anti-bullying program on defending behavior. Seven potential mediating factors assessed after 5 months of program implementation were investigated and our analyses revealed that two of them played a significant mediating role in the effects of the KiVa program on defending behavior after 9 months of program implementation: (a) feelings of responsibility to intervene and (b) outcome expectations that defending would make the bullying stop or decrease, both of which are discussed more below.

5.1. The mediating role of feelings of responsibility to intervene

A central objective of the KiVa program is to make children aware that the behavior they choose to adopt in bullying incidents may influence how the situation evolves. The lessons in the program teach them that siding with the aggressors or remaining passive serves to reinforce the bullying and they have a duty to act when witnessing a peer being victimized. As expected, KiVa had a positive effect on children's feelings that everyone has a responsibility to intervene when witnessing bullying among their peers. After 4–5 months of implementation, the mean levels of responsibility to intervene had increased slightly in intervention schools and decreased slightly in control schools.

In turn, those who more strongly endorsed the view that all students have a responsibility to intervene in situations of bullying were more likely to increase in defending behavior. This finding is aligned with the bystander model that lists "feeling responsible" as one of the requirements for intervening in situations of crisis (Latané & Darley, 1970; Nickerson et al., 2014). This supports the idea that diffusion of responsibility, which refers to situations when individuals fail to provide help to someone in need because they assume that others present share their responsibility to help (Latané & Darley, 1970), may be an important obstacle to intervention when witnessing someone in danger. The diffusion of responsibility phenomenon has indeed been demonstrated in experimental studies with children (Plötner et al., 2015). In the early research on bystander interventions, personal responsibility to intervene decreased as the number of bystanders increased (Latané & Nida, 1981). Other studies have suggested that the larger the group, the less likely people are to help because their responsibilities are shared among all members of the group (e.g., Barron & Yechiam, 2002; Wiesenthal et al.,

1983). Diffusion of responsibility is believed to be due to a decrease in self-awareness that individuals experience when they are in the presence of many others (Wegner & Schaefer, 1978). Specific to school bullying, it remains unclear whether the number of bystanders plays a role. More insight into the factors that promote or deter children's acceptance of responsibility to intervene in favor of victimized peers is needed.

5.2. The mediating role of outcome expectations that bullying would stop or decrease

Our results indicate that the KiVa program had a positive effect on the perception that children have of the consequences of their defending behavior with respect to its effectiveness at making the bullying stop or decrease. Although, on average, children's expectations that defending would decrease bullying declined in both intervention and control schools, children's expectations declined significantly more in control schools. As predicted by Bandura's (1994) social cognitive theory, these beliefs in the effectiveness of the defending behavior increased the likelihood that children would defend. These findings suggest that demonstrating to children that defending will have a positive impact by making the bullying stop is an effective way to encourage them to intervene.

An important question for future research will be to determine whether defending is indeed an effective solution to put an end to bullying. Currently, there is a lack of longitudinal investigations examining the effects of being defended on future victimization (Healy, 2020). The only longitudinal study to date that has addressed this question did not find any evidence that victimized children with at least one defender at the beginning of the school year experienced less victimization by the end of the year than undefended victims (Laninga-Wijnen et al., 2023). It is possible that the success of defending attempts depends on other factors – personal or contextual – that remain to be identified. Moreover, the consequences of defending on the defenders themselves need to be better understood as it might be safe for some students and riskier for other students (Malamut et al., 2021, 2023).

5.3. Strengths and limitations

In addition to the large sample size, a key strength of the present study was the use of a three-wave longitudinal design to examine factors mediating the effects of an anti-bullying intervention on defending behavior. Much of the prior research on the individual characteristics associated with defending has tended to be cross-sectional. Furthermore, the present study has simultaneously investigated a wide range of relevant constructs that are often separately examined.

The study also has several limitations. Out of the seven hypothesized mediators, only two were shown to have a significant indirect effect. This might partly be due to the measurement of the main variables of the study. Several scales, including responsibility to intervene, outcome expectations for defending, and affective empathy for the victims, were created specifically for the RCT evaluation of the KiVa program. At the time of this evaluation, no such scale existed to capture these constructs among children. As the program was designed to influence these constructs, it was important to create items to measure them. However, this implies that the scales were not already validated in prior studies, which could explain why the reliability coefficients of some of the scales were not very high. Moreover, the items used in the self-efficacy for defending measure, which included an aggressive element ("Saying that bullying is stupid") did not exactly match the items used to assess defending behavior. Many of the items used to assess the various hypothesized mediators require the participants – explicitly or not – to imagine a hypothetical bullying event when answering the question. Participating children may have responded to these items differently depending on whether they had previously witnessed situations of bullying or whether they needed to mostly rely on their imagination.

Regarding the responsibility to intervene measure, it is important to note that the measure did not specifically capture personal feelings of responsibility as would be the case with items such as "I feel responsible". Instead, the four items assessed participants' beliefs that it is the responsibility of all witnesses to defend victims of bullying. Furthermore, these items did not measure participants' evaluation of the situation as an emergency or a problem in need of intervention, although this evaluation must precede the assumption of responsibility to intervene in the bystander model of Latané and Darley (1970).

Defending behavior was the only study variable that was assessed with peer nominations. Although this method has the advantage of relying on multiple informants and being less likely than self-reports to be affected by socially desirable responding, it assumes that defending behavior is visible for the rest of the peer group. This could lead to an underestimation of specific types of defending behavior that bring support to the victim but remain hidden from peers. Moreover, at the time of data collection, defending behavior was generally conceptualized as a unidimensional construct. In the past decade, research has shown that defending could encompass a variety of behaviors, including comforting victims, confronting bullies by asking them to stop, reporting the incident to a teacher, or even aggressing the bullies (e.g., Lambe & Craig, 2020; Wang et al., 2023). These different types of defending have different correlates (e.g., Garandeau, Vermande, et al., 2022; Lambe & Craig, 2020; Reijntjes et al., 2016) and it is likely that anti-bullying programs do not increase the prevalence of all types of defending equally and the psychological mechanisms that explain changes in one type of defending might differ from the ones that explain changes in another type. It is noteworthy that one of the defending items combined comforting the victim and encouraging the victim to tell the teacher about the bullying, although these can be considered as distinct behaviors that do not necessarily co-occur.

The fact that the data used in the present study were collected 15 years ago should also be considered when interpreting the findings. The growing prevalence of cyberbullying in the past two decades implies that situations of bullying are no longer restricted to the school but can happen anytime online. This might affect how bystanders intervene in such situations, and psychological mechanisms through which school-based programs successfully increase defending behavior. Therefore, future research should seek to replicate the present findings with recent data.

Although some of the factors of interest may take time to change, the time lag between the different data collection waves was

relatively short. For example, the KiVa program has been shown to have a positive effect on affective empathy after 9 months of program implementation (Garandeau, Laninga-Wijnen, & Salmivalli, 2022); however, statistically significant effects after 4–5 months were not detected in our models. Therefore, even if our analyses did not find evidence for an indirect effect of empathy, such a mechanism is still possible but might take longer to develop.

The present study focused on a sample of primary school children – where the KiVa program had been shown to positively affect defending - and did not consider adolescents. Although research shows that anti-bullying programs are generally less effective in reducing bullying perpetration among adolescents in comparison to younger children (see Salmivalli et al., 2021), a prior meta-analysis indicated that program effects on bystander interventions are larger in high school than in middle school (Polanin et al., 2012). Thus, it will be important for future research to investigate whether the factors that predict increases in defending behavior and that mediate the effectiveness of intervention on defending are the same in childhood and in adolescence. For example, the expectation that defending will benefit one's status might play a stronger role among adolescents, compared to children, as popularity among peers becomes increasingly important at that age (LaFontana & Cillessen, 2010).

5.4. Implications for practice and research

The present findings show that, at least in primary school, it is possible for school personnel to increase the likelihood that students witnessing bullying would be more likely to intervene on behalf of the victimized student by increasing their beliefs that it is their responsibility to do so and their intervention will help put an end to the bullying. Therefore, these cognitions may be the ones that teachers aiming to increase defending among bystanders of bullying should primarily target. The specific program components that effectively affected these statistically significant mediators were not identified. Knowledge of these components would be important to better inform anti-bullying practices beyond the KiVa program (e.g., Cunningham et al., 2020). New exercises specifically aimed at increasing students' feelings of responsibility to intervene, as well as their expectations that their defending will be effective, could be developed. As mentioned above, it will be crucial for future research to determine how effective defending behavior is in putting an end to bullying, whether some types of defending are more effective than others, and whether this effectiveness depends on the context and on personal characteristics of the victimized child and of the defender.

More research is also needed to identify the psychological mechanisms that can promote defending behavior in older age groups (e.g., adolescents) and to investigate whether the cognitions that underlie the effects of anti-bullying programs on defending behavior differ across types of defending. This information would be helpful for teachers. For example, it might be easier to increase self-efficacy for defending for safer intervention strategies, such as privately comforting a victim, than it is for directly confronting a bullying perpetrator.

Finally, the psychological mechanisms underlying changes in defending behavior are not necessarily the same as the psychological mechanisms underlying changes in bullying behavior (e.g., Saarento et al., 2015). This might be because the children who decrease in bullying behavior after exposure to the program are not the same children as those who increase in defending. As educators striving to increase defending among their students generally also aim at decreasing bullying behavior, our findings suggest that it is important for educators to try and influence various types of cognitions needed to obtain beneficial effects among many children.

6. Conclusion

Understanding why anti-bullying programs achieve desired outcomes is essential for improving anti-bullying interventions. Recent research has investigated the effectiveness of specific program components in reducing bullying and victimization (Gaffney et al., 2021; Hensums et al., 2022). Yet, studies investigating the mechanisms underlying the effectiveness of anti-bullying programs are still scarce. This study was the first to investigate potential mediating factors of the effectiveness of an anti-bullying program on defending behavior, thus contributing to opening the "black box" of interventions. Among the seven factors examined, two were shown to play a significant mediating role: (a) feelings of responsibility to intervene and (b) outcome expectations that the bullying would decrease or stop. Researchers should continue investigating other potential mediators, including the ones that were not found to have a statistically significant mediating effect in the present study as these other factors may play a role in the longer term or with a different age group. More research is also needed to determine if the cognitions that the program promotes and that lead to increases in defending are the same for all students. As suggested by Cunningham et al. (2020), both personal characteristics of the children such as their dispositional reactance and contextual factors such as school norms may influence the extent to which anti-bullying interventions affect students' willingness to intervene. We encourage future research to identify (a) the specific program components leading to changes in feelings of responsibility to intervene and expectation that defending would help decrease bullying and (b) possible individual and contextual characteristics that might moderate the indirect effects of anti-bullying programs on defending behavior.

References

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice Hall.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), *Vol. 4. Encyclopedia of human behavior* (pp. 71–81). Academic Press.

Bandura, A. (2016). *Moral disengagement: How people do harm and live with themselves*. Worth Publishers.

Barron, G., & Yechiam, E. (2002). Private e-mail requests and the diffusion of responsibility. *Computers in Human Behavior, 18*, 507–520. https://doi.org/10.1016/S0747-5632(02)00007-9.

Cappadocia, M. C., Pepler, D., Cummings, J. G., & Craig, W. (2012). Individual motivations and characteristics associated with bystander intervention during bullying episodes among children and youth. *Canadian Journal of School Psychology, 27*, 201–216. https://doi.org/10.1177/0829573512450567.

Cunningham, C. E., Rimas, H., Vaillancourt, T., Stewart, B., Deal, K., Cunningham, L., … Thabane, L. (2020). What antibullying program designs motivate student intervention in grades 5 to 8? *Journal of Clinical Child and Adolescent Psychology, 49*, 603–617. https://doi.org/10.1080/15374416.2019.1567344.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*, 377–383. https://doi.org/10.1037/h0025589.

Deng, X., Yang, J., & Wu, Y. (2021). Adolescent empathy influences bystander defending in school bullying: A three-level meta-analysis. *Frontiers in Psychology, 12*, Article 690898. https://doi.org/10.3389/fpsyg.2021.690898.

DeSmet, A., Bastiaensens, S., Van Cleemput, K., Poels, K., Vandebosch, H., Cardon, G., & De Bourdeaudhuij, I. (2016). Deciding whether to look after them, to like it, or leave it: A multidimensional analysis of predictors of positive and negative bystander behavior in cyberbullying among adolescents. *Computers in Human Behavior, 57*, 398–415. https://doi.org/10.1016/j.chb.2015.12.051.

Doramajian, C., & Bukowski, W. M. (2015). A longitudinal study of the associations between moral disengagement and active defending versus passive bystanding during bullying situations. *Merrill-Palmer Quarterly, 61*, 144–172. http://doi.org/10.13110/merrpalmquar1982.61.1.0144.

Eisenberg, N., Shea, C. L., Carlo, G., & Knight, G. P. (1991). Empathy-related responding and cognition: A "chicken and the egg" dilemma. In W. M. Kurtines, & J. L. Gewirtz (Eds.), *Vol. 1. Theory; Vol. 2. Research; Vol. 3. Application. Handbook of moral behavior and development* (pp. 63–88). Lawrence Erlbaum Associates, Inc.

Fredrick, S. S., Jenkins, L. N., & Ray, K. (2020). Dimensions of empathy and bystander intervention in bullying in elementary school. *Journal of School Psychology, 79*, 31–42. https://doi.org/10.1016/j.jsp.2020.03.001.

Gaffney, H., Ttofi, M. M., & Farrington, D. P. (2021). What works in anti-bullying programs? Analysis of effective intervention components. *Journal of School Psychology, 85*, 37–56. https://doi.org/10.1016/j.jsp.2020.12.002.

Garandeau, C. F., Laninga-Wijnen, L., & Salmivalli, C. (2022). Effects of the KiVa anti-bullying program on affective and cognitive empathy in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 55*, 515–529. https://doi.org/10.1080/15374416.2020.1846541.

Garandeau, C. F., & Salmivalli, C. (2018). Le programme anti-harcèlement KiVa. *Enfance, 3*, 491–501. https://doi.org/10.3917/enf2.183.0491.

Garandeau, C. F., Vermande, M. M., Reijntjes, A. H. A., & Aarts, E. (2022). Classroom bullying norms and peer status: Effects on victim-oriented and bully-oriented defending. *International Journal of Behavioral Development, 46*, 401–410. https://doi.org/10.1177/0165025419894722.

Gini, G. (2006). Social cognition and moral cognition in bullying: What's wrong? *Aggressive Behavior, 32*, 528–539. https://doi.org/10.1002/ab.20153.

Gini, G., Albiero, P., Benelli, B., & Altoè, G. (2008). Determinants of adolescents' active defending and passive bystanding behavior in bullying. *Journal of Adolescence, 31*, 93–105. https://doi.org/10.1016/j.adolescence.2007.05.002.

Gini, G., Pozzoli, T., Angelini, F., Thornberg, R., & Demaray, M. K. (2022). Longitudinal associations of social-cognitive and moral correlates with defending in bullying. *Journal of School Psychology, 91*, 146–159. https://doi.org/10.1016/j.jsp.2022.01.005.

Haataja, A., Ahtola, A., Poskiparta, E., & Salmivalli, C. (2015). A process view on implementing an antibullying curriculum: How teachers differ and what explains the variation. *School Psychology Quarterly, 30*, 564–576. https://doi.org/10.1037/spq0000121.

Haataja, A., Voeten, M., Boulton, A., Ahtola, A., Poskiparta, E., & Salmivalli, C. (2014). KiVa antibullying curriculum and outcome: Does fidelity matter? *Journal of School Psychology, 52*, 479–493. https://doi.org/10.1016/j.jsp.2014.07.001.

Harachi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. B. (1999). Opening the black box: Using process evaluation measures to assess implementation and theory building. *American Journal of Community Psychology, 27*, 711–731. https://doi.org/10.1023/A:1022194005511.

Hawkins, D. L., Pepler, D. J., & Craig, W. M. (2001). Naturalistic observations of peer interventions in bullying. *Social Development, 10*, 512–527. https://doi.org/10.1111/1467-9507.00178.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press.

Healy, K. L. (2020). Hypotheses for possible iatrogenic impacts of school bullying prevention programs. *Child Development Perspectives, 14*, 221–228. https://doi.org/10.1111/cdep.12385.

Hensums, M., de Mooij, B., Kuijper, S. C., BIRC: the anti-Bullying Interventions Research Consortium, Fekkes, M., & Overbeek, G. (2022). What works for whom in school-based anti-bullying interventions? An individual participant data meta-analysis. *Prevention Science*. https://doi.org/10.1007/s11121-022-01387-z.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. https://doi.org/10.1080/10705519909540118.

Huitsing, G., Lodder, G., Browne, W. J., Oldenburg, B., Van der Ploeg, R., & Veenstra, R. (2020). A large-scale replication of the effectiveness of the KiVa antibullying program: A randomized controlled trial in the Netherlands. *Prevention Science, 21*, 627–638. https://doi.org/10.1007/s11121-020-01116-4.

Jenkins, L., Fredrick, S. S., & Nickerson, A. (2018). The assessment of bystander intervention in bullying: Examining measurement invariance across gender and grade. *Journal of School Psychology, 69*, 73–83. https://doi.org/10.1016/j.jsp.2018.05.008.

Jenkins, L. N., & Nickerson, A. B. (2017). Bullying participant roles and gender as predictors of bystander intervention. *Aggressive Behavior, 43*, 281–290. https://doi.org/10.1002/ab.21688.

Jiang, S., Liu, R.-D., Ding, Y., Jiang, R., Fu, X., & Hong, W. (2022). Why the victims of bullying are more likely to avoid involvement when witnessing bullying situations: The role of bullying sensitivity and moral disengagement. *Journal of Interpersonal Violence, 37*. NP3062-NP3083 https://doi.org/10.1177/0886260520948142. NP3062-NP3083.

Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Kaljonen, A., & Salmivalli, C. (2011). A large-scale evaluation of the KiVa antibullying program: Grades 4-6. *Child Development, 82*, 311–330. https://doi.org/10.1111/j.1467-8624.2010.01557.x.

Kärnä, A., Voeten, M., Poskiparta, E., & Salmivalli, C. (2010). Vulnerable children in varying classroom contexts: bystanders' behaviors moderate the effects of risk factors on victimization. *Merrill-Palmer Quarterly, 56*, 261–282. https://digitalcommons.wayne.edu/mpq/vol56/iss3/4.

LaFontana, K. M., & Cillessen, A. H. N. (2010). Developmental changes in the priority of perceived status in childhood and adolescence. *Social Development, 19*, 130–147. https://doi.org/10.1111/j.1467-9507.2008.00522.x.

Lambe, L. J., & Craig, W. M. (2020). Peer defending as a multidimensional behavior: Development and validation of the defending behaviors scale. *Journal of School Psychology, 78*, 38–53. https://doi.org/10.1016/j.jsp.2019.12.001.

Lambe, L. J., Della Cioppa, V., Hong, I. K., & Craig, W. M. (2019). Standing up to bullying: A social ecological review of peer defending in offline and online contexts. *Aggression and Violent Behavior, 45*, 51–74. https://doi.org/10.1016/j.avb.2018.05.007.

Laninga-Wijnen, L., van den Berg, Y. H. M., Garandeau, C. F., Mulder, S., & Orobio De Castro, B. (2023). Does being defended relate to decreases in victimization and improved psychosocial adjustment among victims? *Journal of Educational Psychology, 115*, 363–377. https://doi.org/10.1037/edu0000712.

Latané, B., & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help? Appleton-Century-Croft*.

Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin, 89*, 308–324. https://doi.org/10.1037/0033-2909.89.2.308.

Lodge, J., & Frydenberg, E. (2005). The role of bystanders in school bullying: Positive steps toward promoting peaceful schools. *Theory Into Practice, 44*, 329–336. https://doi.org/10.1207/s15430421tip4404_6.

Ma, T., Meter, D., Chen, W., & Lee, Y. (2019). Defending behavior of peer victimization in school and cyber context during childhood and adolescence: A meta-analytic review of individual and peer-relational characteristics. *Psychological Bulletin, 145*, 891–928. https://doi.org/10.1037/bul0000205.

Malamut, S., Trach, J., Garandeau, C. F., & Salmivalli, C. (2021). Examining the potential mental health costs of defending victims of bullying: A longitudinal analysis. *Research on Child and Adolescent Psychopathology, 49*, 1197–1210. https://doi.org/10.1007/s10802-021-00822-z.

Malamut, S. T., Trach, J., Garandeau, C. F., & Salmivalli, C. (2023). Does defending victimized peers put youth at risk of being victimized? *Child Development, 94*, 380–394. https://doi.org/10.1111/cdev.13866.

Mazzone, A., Camodeca, M., & Salmivalli, C. (2016). Interactive effects of guilt and moral disengagement on bullying, defending and outsider behavior. *Journal of Moral Education, 45*, 419–432. https://doi.org/10.1080/03057240.2016.1216399.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Nickerson, A. B., Aloe, A. M., Livingston, J. A., & Feeley, T. H. (2014). Measurement of the bystander intervention model for bullying and sexual harassment. *Journal of Adolescence, 37*, 391–400. https://doi.org/10.1016/j.adolescence.2014.03.003.

Nickerson, A. B., Aloe, A. M., & Werth, J. M. (2015). The relation of empathy and defending in bullying: A meta-analytic investigation. *School Psychology Review, 44*, 372–390. http://doi.org/10.17105/spr-15-0035.1.
Nocentini, A., & Menesini, E. (2016). KiVa anti-bullying program in Italy: Evidence of effectiveness in a randomized control trial. *Prevention Science, 17*, 1012–1023. https://doi.org/10.1007/s11121-016-0690-z.
Nocentini, A., Menesini, E., & Salmivalli, C. (2013). Level and change of bullying behavior during high school: A multilevel growth curve analysis. *Journal of Adolescence, 36*, 495–505. https://doi.org/10.1016/j.adolescence.2013.02.004.
Peets, K., Pöyhönen, V., Juvonen, J., & Salmivalli, C. (2015). Classroom norms of bullying alter the degree to which children defend in response to their affective empathy and power. *Developmental Psychology, 51*, 913–920. https://doi.org/10.1037/a0039287.
van der Ploeg, R., Kretschmer, T., Salmivalli, C., & Veenstra, R. (2017). Defending victims: What does it take to intervene in bullying and how is it rewarded by peers? *Journal of School Psychology, 65*, 1–10. https://doi.org/10.1016/j.jsp.2017.06.002.
Plötner, M., Over, H., Carpenter, M., & Tomasello, M. (2015). Young children show the bystander effect in helping situations. *Psychological Science, 26*, 499–506. https://doi.org/10.1177/0956797615569579.
Polanin, J., Espelage, D., & Pigott, T. (2012). A meta-analysis of school-based bullying prevention programs' effects on bystander intervention behavior. *School Psychology Review, 41*, 47–65. https://doi.org/10.1080/02796015.2012.12087375.
Pouwels, J. L., van Noorden, T. H. J., Lansu, T. A. M., & Cillessen, A. H. N. (2018). The participant roles of bullying in different grades: Prevalence and social status profiles. *Social Development, 27*, 732–747. https://doi.org/10.1111/sode.12294.
Pöyhönen, V., Juvonen, J., & Salmivalli, C. (2010). What does it take to stand up for the victim of bullying? *Merrill-Palmer Quarterly, 56*, 143–163. https://doi.org/10.1353/mpq.0.0046.
Pöyhönen, V., Juvonen, J., & Salmivalli, C. (2012). Standing up for the victim, siding with the bully or standing by? Bystander responses in bullying situations. *Social Development, 21*, 722–741. https://doi.org/10.1111/j.1467-9507.2012.00662.x.
Pozzoli, T., & Gini, G. (2010). Active defending and passive bystanding behavior in bullying: The role of personal characteristics and perceived peer pressure. *Journal of Abnormal Child Psychology, 38*, 815–827. https://doi.org/10.1007/s10802-010-9399-9.
Pozzoli, T., Gini, G., & Thornberg, R. (2016). Bullying and defending behavior: The role of explicit and implicit moral cognition. *Journal of School Psychology, 59*, 67–81. https://doi.org/10.1016/j.jsp.2016.09.005.
Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*, 717–731. https://doi.org/10.3758/BF03206553.
Reijntjes, A., Vermande, M., Olthof, T., Goossens, F. A., Aleva, L., & van der Meulen, M. (2016). Defending victimized peers: Opposing the bully, supporting the victim, or both? *Aggressive Behavior, 42*, 585–597. https://doi.org/10.1002/ab.21653.
Rigby, K., & Slee, P. T. (1991). Bullying among Australian school children: Reported behavior and attitudes toward victims. *Journal of Social Psychology, 131*, 615–627. https://doi.org/10.1080/00224545.1991.9924646.
Saarento, S., Boulton, A. J., & Salmivalli, C. (2015). Reducing bullying and victimization: Student- and classroom-level mechanisms of change. *Journal of Abnormal Child Psychology, 43*, 61–76. https://doi.org/10.1007/s10802-013-9841-x.
Sainio, M., Veenstra, R., Huitsing, G., & Salmivalli, C. (2011). Victims and their defenders: A dyadic approach. *International Journal of Behavioral Development, 35*, 144–151. https://doi.org/10.1177/0165025410378068.
Salmivalli, C. (2014). Participant roles in bullying: How can peer bystanders be utilized in interventions? *Theory Into Practice, 53*, 286–292. https://doi.org/10.1080/00405841.2014.947222.
Salmivalli, C., Haataja, A., & Poskiparta, E. (2011, August). Implementation fidelity of the KiVa antibullying program during randomized controlled trial and broad dissemination. In *Presentation at the 15th European conference of Developmental Psychology, Bergen, Norway*.
Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the group. *Aggressive Behavior, 22*, 1–15. http://doi.org/10.1002/(SICI)1098-2337(1996)22:1<1::AID-AB1>3.0.CO;2-T.
Salmivalli, C., Laninga-Wijnen, L., Malamut, S. T., & Garandeau, C. F. (2021). Bullying prevention in adolescence: Solutions and new challenges from the past decade. *Journal of Research on Adolescence, 31*, 1023–1046. https://doi.org/10.1111/jora.12688.
Salmivalli, C., & Voeten, M. (2004). Connections between attitudes, group norms, and behaviour in bullying situations. *International Journal of Behavioral Development, 28*, 246–258. https://doi.org/10.1080/01650250344000488.
Salmivalli, C., Voeten, M., & Poskiparta, E. (2011). Bystanders matter: Associations between reinforcing, defending, and the frequency of bullying behavior in classrooms. *Journal of Clinical Child and Adolescent Psychology, 40*(5), 668–676. https://doi.org/10.1080/15374416.2011.597090.
Sjögren, B., Thornberg, R., Wänström, L., & Gini, G. (2020). Associations between individual and collective efficacy beliefs and students' bystander behavior. *Psychology in the Schools, 57*, 1710–1723. https://doi.org/10.1002/pits.22412.
Thornberg, R., & Jungert, T. (2013). Bystander behavior in bullying situations: Basic moral sensitivity, moral disengagement and defender self-efficacy. *Journal of Adolescence, 36*, 475–483. https://doi.org/10.1016/j.adolescence.2013.02.003.
Thornberg, R., Landgren, L., & Wiman, E. (2018). It depends: A qualitative study on how adolescent students explain bystander intervention and non-intervention in bullying situations. *School Psychology International, 39*, 400–415. https://doi.org/10.1177/0143034318779225.
Thornberg, R., Pozzoli, T., & Gini, G. (2022). Defending or remaining passive as a bystander of school bullying in Sweden: The role of moral disengagement and antibullying class norms. *Journal of Interpersonal Violence, 37*, 19–20. https://doi.org/10.1177/08862605211037427.
Thornberg, R., Pozzoli, T., Gini, G., & Jungert, T. (2015). Unique and interactive effects of moral emotions and moral disengagement on bullying and defending among school children. *The Elementary School Journal, 116*, 322–337. https://doi.org/10.1086/683985.
Thornberg, R., Wänström, L., Elmelid, R., Johansson, A., & Mellander, E. (2020). Standing up for the victim or supporting the bully? Bystander responses and their associations with moral disengagement, defender self-efficacy, and collective efficacy. *Social Psychology of Education, 23*, 563–581. https://doi.org/10.1007/s11218-020-09549-z.
Thornberg, R., Wänström, L., Hong, J. S., & Espelage, D. L. (2017). Classroom relationship qualities and social-cognitive correlates of defending and passive bystanding in school bullying in Sweden: A multilevel analysis. *Journal of School Psychology, 63*, 49–62. https://doi.org/10.1016/j.jsp.2017.03.002.
Van Cleemput, K., Vandebosch, H., & Pabian, S. (2014). Personal characteristics and contextual factors that determine "helping," "joining in," and "doing nothing" when witnessing cyberbullying. *Aggressive Behavior, 40*, 383–396. https://doi.org/10.1002/ab.21534.
Van Noorden, T. H., Haselager, G. J., Cillessen, A. H., & Bukowski, W. M. (2015). Empathy and involvement in bullying in children and adolescents: A systematic review. *Journal of Youth and Adolescence, 44*, 637–657. https://doi.org/10.1007/s10964-014-0135-6.
Wang, Z., Laninga-Wijnen, L., Garandeau, C. F., & Liu, J. (2023). Development and validation of the Adolescent Defending Behaviors Questionnaire among Chinese early adolescents. *Assessment*, Article 10731911221149082. https://doi.org/10.1177/10731911221149082.
Wegner, D. M., & Schaefer, D. (1978). The concentration of responsibility: An objective self-awareness analysis of group size effects in helping situations. *Journal of Personality and Social Psychology, 36*, 147–155. https://doi.org/10.1037/0022-3514.36.2.147.
Wiesenthal, D. L., Austrom, D., & Silverman, I. (1983). Diffusion of responsibility in charitable donations. *Basic and Applied Social Psychology, 4*, 17–27. https://doi.org/10.1207/s15324834basp0401_2.
Yun, H.-Y., & Juvonen, J. (2020). Navigating the healthy context paradox: Identifying classroom characteristics that improve the psychological adjustment of bullying victims. *Journal of Youth and Adolescence, 49*, 2203–2213. https://doi.org/10.1007/s10964-020-01300-3.
Zych, I., Ttofi, M. M., & Farrington, D. P. (2019). Empathy and callous-unemotional traits in different bullying roles: A systematic review and meta-analysis. *Trauma, Violence & Abuse, 20*, 3–21. https://doi.org/10.1177/1524838016683456.