



EMPIRICAL ARTICLE

The effectiveness of moral disengagement and social norms as anti-bullying components: A randomized controlled trial

Chloé Tolmatcheff¹ | Benoit Galand¹ | Isabelle Roskam¹ | René Veenstra²

¹Faculty of Psychology and Educational Sciences, University of Louvain, Louvain-la-Neuve, Belgium

²Department of Sociology, University of Groningen, Groningen, The Netherlands

Correspondence

Chloé Tolmatcheff, Psychological Sciences Research Institute (IPSY), 10 Place du Cardinal Mercier, bte L3.05.01, B-1348, Louvain-la-Neuve, Belgium.
Email: chloe.tolmatcheff@gmail.com

Funding information

Fonds De La Recherche Scientifique—FNRS, Grant/Award Number: FC23715

Abstract

This three-armed randomized controlled trial examined how moral disengagement and social norms account for change in bullying behavior and their potential as targets of anti-bullying components within separate interventions among 1200 French-speaking Belgian elementary students (48% boys, 9–12 year-olds, 57 classes, nine schools) during 2018–2019 (no ethnicity data available). Mediation analysis revealed that students' moral disengagement successfully decreased ($\beta = -.46$), which, in turn, reduced both bullying ($\beta = .33$) and outsider behaviors ($\beta = .20$), and increased defending ($\beta = -.10$). Intervening on social norms decreased bullying ($\beta = -.18$), but not through the perceived injunctive class norm as intended. Guidelines to open the “black box” of anti-bullying programs and determine the cost-effectiveness ratio of their components are provided.

Bullying is systematic, unprovoked, peer-to-peer abuse in the context of a power imbalance (Hymel & Swearer, 2015). Short- and long-term negative consequences have been documented not only for victims, but also for bullies and bystanders (e.g., Nishina & Juvonen, 2005). Several anti-bullying programs have been developed to tackle this universal and pervasive phenomenon. Although some programs have shown positive effects on bullying and/or victimization (e.g., KiVa, Kärnä et al., 2011; ViSC, Yanagida et al., 2019), meta-analyses suggest that the results of anti-bullying programs have been mixed and limited overall (a 20% reduction in bullying on average; Gaffney et al., 2019a). Our understanding of how these programs work as well as our knowledge of the psychosocial developmental processes involved in bullying is limited by two main gaps in the evaluations.

First, most researchers do not assess the factors that they aim to manipulate to influence the desired outcome

(see Figure 1a). Instead, they tend to focus on evaluating the behavioral outcomes and overlook the assessment of the hypothesized mediators (Saarento et al., 2015). Assessing mediators enables researchers to test the role of specific processes in bullying and reduces the risk of wrongly concluding that the program or its components have no effect on bullying (when, for instance, the effect is too small to be detected at the distal level or a concurrent effect in the opposite direction is concealing the effect; Rucker et al., 2011).

Second, although anti-bullying programs are often complex and combine several components at different levels of prevention (van Versveld et al., 2021), they are usually assessed as a whole package (Eisner & Malti, 2012). As a result, we do not know the unique contribution of each component, regardless of the global outcomes of the program (Menesini & Salmivalli, 2017). Although assessing the mediators is sometimes used

Abbreviations: BLP-MD, Bullying Literature Project-Moral Disengagement; CFA, confirmatory factorial analysis; CFI, comparative fit index; Fr-MDBS, French-version of the Moral Disengagement in Bullying Scale; LC, latent change; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; TLI, Tucker–Lewis index.

Chloé Tolmatcheff is now at INVEST Flagship Research Center, Department of Psychology and Speech-Language Pathology, University of Turku, Finland.

All material used in the interventions evaluated in this study are openly available in OSF at https://osf.io/w5hd7/?view_only=693d6d86743d4c49980a7768ce1e555b.

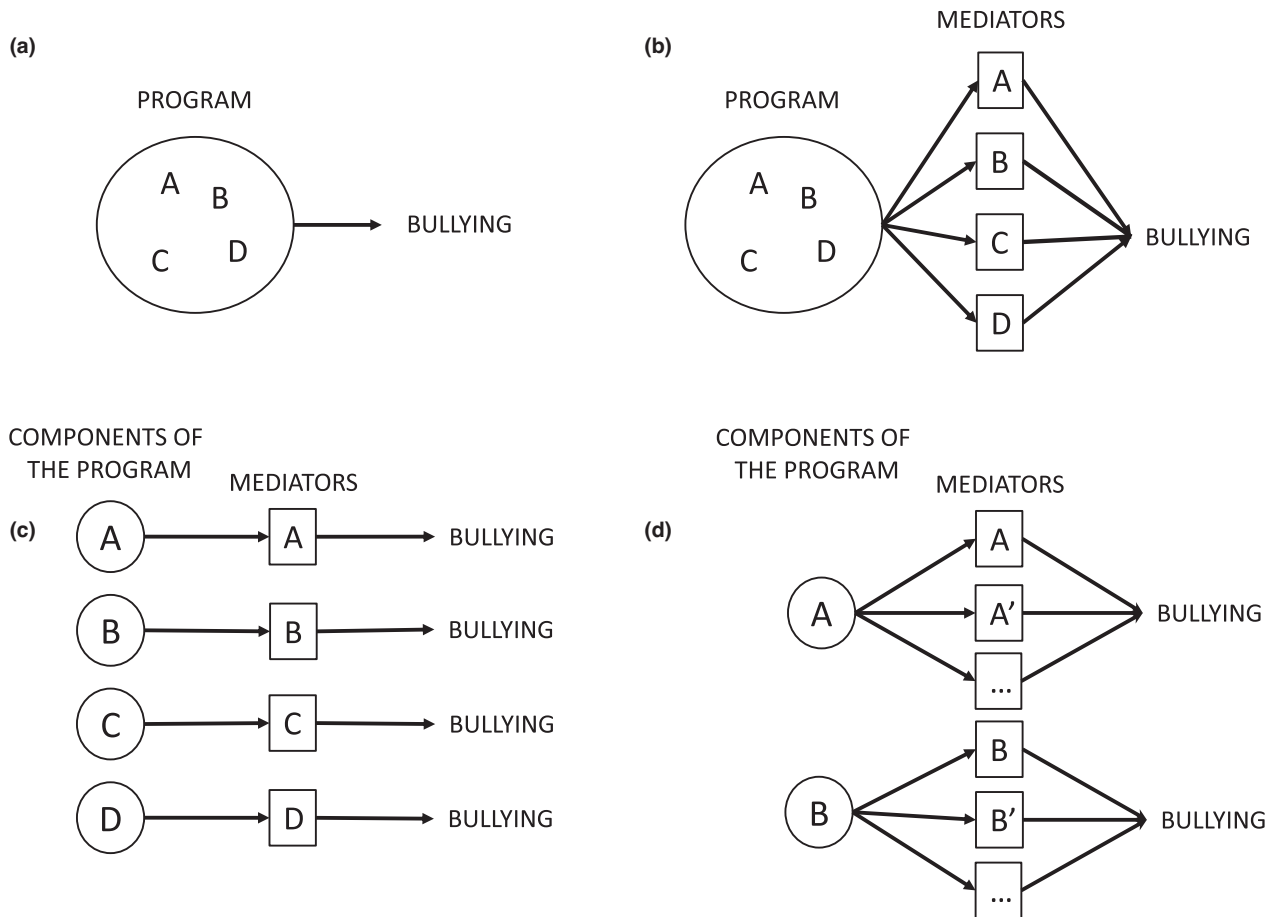


FIGURE 1 Research designs of anti-bullying program evaluation studies. (a) Components of the program assessed as a whole; (b) components of the program assessed as a whole along with their corresponding hypothesized mediators; (c) components of the program and their corresponding hypothesized mediators assessed within separate interventions; (d) components of the program and their corresponding hypothesized mediators assessed within separate interventions along with alternative mediators.

as a way to identify programs' active components (see [Figure 1b](#)), this approach is not ideal as it relies on the assumption that any change in a given mediator can only be caused by the component that has been hypothesized to influence this specific mediator. As change in a given mediator could be unexpectedly related to another component of the program or to the joint action of two or more components, the need for further innovative research to better understand programs' causal mechanisms has been pointed out (Eisner & Malti, 2012).

To address these two methodological issues and open the “black box” of anti-bullying programs, evaluation studies should assess anti-bullying components within separate interventions and test their hypothesized mediating paths (see [Figure 1c](#)). An ideal research design should also allow to test other mediators that may provide alternative explanations for the effect of a given component on bullying (see [Figure 1d](#)). Such a research design is used in the present study.

Apart from these methodological gaps, the limited effectiveness of anti-bullying programs also emphasizes the need to promote research-based innovation in program development and keep studying new directions

to prevent bullying by testing interventions that reflect scientific advances (Eisner & Malti, 2012). To expand our knowledge and understanding of anti-bullying programs, we need firstly to identify and select relevant anti-bullying components to study, based on prior research and theory (Eisner & Malti, 2012). The meaning behind the word “components” differs across studies and is sometimes used to refer to specific characteristics (e.g., whole-school approach; Gaffney et al., 2021) or elements (e.g., digital game; Gaete et al., 2017; student lessons; van Versveld et al., 2021) of the programs. We define the “components” with respect to the specific processes accounting for change in bullying behavior that they target. The examination of the mechanisms of change to understand how intra-individual change occurs (the *why* of development, in contrast to the *how*, referring to the course of change over age) is a major focus of developmental psychology (Schaffer, 2006).

Our study examines two specific processes that have been suggested as maintaining bullying behavior over time: moral disengagement and social norms (e.g., Hymel & Bonanno, 2014; Menesini et al., 2015). Although most students disapprove of bullying and consider it immoral,

many of them bully others and few bystanders defend or support victims (Hymel et al., 2010). Both the moral disengagement and social norms frameworks offer a rationale to explain this discrepancy between students' private attitudes and their actual behaviors, and have been suggested as promising directions for bullying prevention (e.g., Hymel & Bonanno, 2014; Menesini et al., 2015; Thornberg, Pozzoli, et al., 2021) but have rarely, if at all, been targeted as such in programs yet (for exceptions, see Dillon & Lochman, 2019; Perkins et al., 2011; Wang & Goldberg, 2017). The goal of our intervention study was twofold: By addressing the two aforementioned methodological gaps, we aimed (1) to provide insight into the roles of moral disengagement and social norms in maintaining bullying over time, and (2) to determine their potential as anti-bullying prevention program components. Early adolescence corresponds to the beginning of the developmental “peak” of bullying, suggested to occur between ages 9–15 years (Álvarez-García et al., 2015). In addition, early adolescents are more likely to be receptive to anti-bullying interventions, as meta-analyses have indicated that the effectiveness of anti-bullying programs strongly declines in older adolescents (13 years old and over; Yeager et al., 2015). This decline is likely due to developmental trends, such as the striving for autonomy and independence from adult authorities (Yeager et al., 2015), and the progressive increase of peer influence in adolescence (Sandstrom & Bartini, 2010). Early adolescence thus represents a suitable developmental period to examine the processes accounting for change in bullying as well as for prevention. Therefore, we conducted our study in a sample of elementary students between 9 and 12 years old.

Intervening on moral disengagement as an anti-bullying component

Moral disengagement refers to selective disengagement by individuals from their own moral standards (Bandura, 1991). Over the course of their development, children progressively assimilate moral principles that guide their behavior. Following these principles yields self-satisfaction and a sense of self-worth, whereas transgressing them yields internal sanctions such as guilt and shame. In the social cognitive theory of thought and action (Bandura, 1991), eight moral disengagement mechanisms have been identified, which can selectively disengage internal controls that normally lead people to feel guilty when transgressing their moral standards: moral justification, euphemistic labeling, advantageous comparison, responsibility displacement, responsibility diffusion, disregarding or distorting the consequences, dehumanization, and blaming the victim. Students may, for example, argue that the bullying serves a moral and worthy purpose (e.g., “I do it to help my friends”; moral justification), or negate it through sanitized and

innocuous language (e.g., “It’s just a joke”; euphemistic labeling). Moral disengagement is conceptualized as a gradual and reciprocal process, which predicts and maintains immoral behavior such as bullying over time (Bandura, 1991).

Cross-sectional studies have indicated that the use of moral disengagement mechanisms is associated with bullying and bystanders' behaviors (see Killer et al., 2019 for a recent meta-analysis). The key role of bystanders in either enhancing or discouraging bullying has long been acknowledged (Salmivalli et al., 1996). The defender role—referring to bystanders who support the victim or stand up to the bully (Yun, 2020)—is negatively linked with moral disengagement (Killer et al., 2019). The few existing results about the outsider role—referring to bystanders who passively observe bullying or walk away—are more mixed, but recent findings have indicated a positive association with moral disengagement (e.g., Doramajian & Bukowski, 2015; Thornberg, Pozzoli, et al., 2021). Outsiders are likely to rely on moral disengagement mechanisms when refraining from helping, as a way to distance themselves from the victim's distress and avoid self-sanctions (Doramajian & Bukowski, 2015).

Bullying or refraining from supporting the victim is expected to be less likely if students can no longer rely on moral disengagement mechanisms to justify their behavior and therefore escape unpleasant internal sanctions like guilt and shame. To our knowledge, only one study has evaluated an intervention that explicitly targeted moral disengagement to help reduce school bullying, providing mixed but promising results in a sample of 7–9 year olds (Wang & Goldberg, 2017). As the use of moral disengagement mechanisms toward bullying is embedded in students' cognitive and moral development (Bandura, 1991), moral disengagement could even have a more important role in bullying in older students, such as 9–12 year-olds in our sample.

Intervening on social norms as an anti-bullying component

The second anti-bullying component concerns the social norms that shape, constrain, and maintain behaviors of group members over and above individual attitudes (Veenstra & Lodder, 2022). Individuals tend to conform to what they perceive to be the norm in a context, referring to what is typical or desirable in a group or a situation (Veenstra & Lodder, 2022). However, individuals' subjective perceptions of norms are not necessarily accurate (Tankard & Paluck, 2016). Students tend to overestimate, sometimes drastically, the prevalence of risky or problem behaviors and attitudes among their peers, notably toward bullying in both children (8–9 years old; Sandstrom et al., 2013) and early adolescents (ranging from 11 to 14 years old; Dillon & Lochman, 2019; Perkins et al., 2011; Sandstrom & Bartini, 2010).



This misperception can result in *pluralistic ignorance*, which describes a situation in which most members of a group privately reject an opinion or behavior while believing that the others approve of it, and thus behave in accordance with this biased norm (Perkins et al., 2011; Sandstrom et al., 2013). Consequently, students maintain and reinforce the collective misperception over time (Veenstra & Lodder, 2022). Several studies have shown that both assisting or reinforcing the bully and defending the victim are associated with perceived peers' rather than private attitudes toward bullying in both children (8–9 years old; Sandstrom et al., 2013) and early adolescents (ranging from 11 to 14 years old; Pozzoli & Gini, 2010; Rigby & Johnson, 2006; Sandstrom & Bartini, 2010).

A distinction is usually made between *descriptive* norms (referring to the prevalence of a behavior) and *injunctive* (or *prescriptive*) norms (referring to the [dis]approval of a behavior) within a group (Miller & Prentice, 2016). When it comes to pluralistic ignorance and correcting misperception, we also need to distinguish between *actual* and *perceived* norms (Perkins, 2014; Sandstrom & Bartini, 2010; Thornberg, Pozzoli, et al., 2021). Table 1 presents how actual, perceived, descriptive, and injunctive norms are typically operationalized in bullying studies (Thornberg, Pozzoli, et al., 2021; Veenstra & Lodder, 2022).

Social norms theory predicts that providing group members with valid information about their peers' attitudes or behaviors can correct their perception of the norm, which, in turn, induces a change of behavior (Miller & Prentice, 2016). As pluralistic ignorance implies a misperception of others' attitudes toward bullying, interventions should attempt to correct biased perceptions of injunctive norms rather than descriptive norms (Miller & Prentice, 2016; Prentice, 2008). To be effective, norm-based interventions also need to target an appropriate reference group for feedback on the norms. In the case of bullying, the class is a salient normative context for students (Pozzoli et al., 2012). Correcting students' misperception of the injunctive class norm toward bullying should discourage bullies' behaviors and increase bystanders' willingness to intervene. To date, only a few intervention studies have applied this approach to the context of school bullying, providing

mixed but encouraging results in samples of 11–14 years old (Perkins et al., 2011) and 12–14 years old (Dillon & Lochman, 2019).

The present study

This study aimed to provide insights into the roles of moral disengagement and social norms in maintaining bullying behavior over time, and determine their potential for prevention as targets of anti-bullying program components. To do so, we aimed to address the two aforementioned methodological gaps by assessing the effects of intervening on moral disengagement and social norms within separate interventions, while verifying that the effects were mediated through moral disengagement and perceived injunctive class norm toward bullying. We might have an unintended impact on students' perception of the injunctive class norm by delivering a class-wide intervention on moral disengagement, and vice-versa. For that reason, and consistently with our methodological recommendations (see Figure 1d), both moral disengagement and perceived injunctive class norm were also tested as an alternative mediator of the other intervention in our research design.

A three-armed randomized controlled trial with a moral disengagement intervention condition, a social norms intervention condition, and a control group was conducted in a sample of early adolescents (9–12 years old) in French-speaking Belgium. The two interventions exclusively targeted either moral disengagement or the social norms and were inspired by the Bullying Literature Project-Moral Disengagement Version (Wang & Goldberg, 2017) and the Survey of Bullying at Your School (Perkins et al., 2011). Schools from the control group received an intervention of similar duration and intensity on a topic unrelated to bullying (climate and environmental issues). In this way, we attempted to control for changes stemming from nonspecific treatment factors leading to spurious intervention effects. For instance, when teachers spend more time together because of an intervention project, their relationships are likely to improve, which may in turn produce positive effects among their students (Lilienfeld et al., 2014). The study took place over the school year

TABLE 1 Operationalization of actual and perceived descriptive and injunctive norms

	Descriptive norms	Injunctive norms
Actual	Students' individual (self- or peer-reported) scores on a given behavior averaged at the group level (e.g., Peets et al., 2015; Pozzoli et al., 2012; Sentse et al., 2015)	Students' individual attitudes toward a given behavior averaged at the group level (e.g., Dillon & Lochman, 2019; Perkins et al., 2011; Sandstrom & Bartini, 2010; Sandstrom et al., 2013; Sentse et al., 2015)
Perceived	Students' individual perceptions of the prevalence of a given behavior among their peers (e.g., Kubiszewski et al., 2019; Saarento et al., 2015)	Students' individual perceptions of their peers' attitudes toward a given behavior (e.g., Dillon & Lochman, 2019; Perkins et al., 2011; Sandstrom & Bartini, 2010; Sandstrom et al., 2013)

2018–2019. Each group of teachers attended a training day delivered by the first author of this paper between mid-January and the end of February 2019. The five intervention sessions were then delivered by the class teachers to their students for 1 h a week within a month after the training.

This study was mainly confirmatory in nature, although we also conducted some additional exploratory analyses, which are detailed in the Analytical Strategy section. We assumed that challenging students' use of moral disengagement mechanisms would restrain their use of the mechanisms (Hypothesis 1). In parallel, we assumed that providing students with correct information about their classmates' anti-bullying attitudes would correct their biased perception of the injunctive class norm toward bullying (Hypothesis 2). We assumed that each of these two primary outcomes, in turn, would mediate the following secondary behavioral outcomes: a decrease in bullying (Hypotheses 1a and 2a), and a reduction in outsider behaviors along with an increase in defending behaviors (Hypotheses 1b and 2b).

METHOD

Sampling and design

Figure 2 presents a flowchart of the recruitment and retention of intervention and control schools. The [Supporting Information](#) provides detailed information about the power calculation, recruitment, eligibility criteria, randomization process, demographic characteristics, timeline, and data collection. Nine French-speaking Belgian schools (57 classes) of varying size, location, and socioeconomic level agreed to participate in the project and were randomly allocated to one of the three conditions.

Data collection and participants

Data were collected in the fall of 2018 (T1) and the spring of 2019 (T2), with an average of 15 weeks between the two waves. Parental consent had been asked prior to data collection through passive consent forms sent by the schools. Students completed confidential paper questionnaires during class hours under the supervision of one or two research team members, who were available to answer questions. In a 5-min oral introduction, the researchers presented the goal of the survey (“investigating children's well-being at school”) and a few basic rules. Students were reminded that the questionnaire was strictly confidential and not mandatory. The whole procedure was approved by the ethics committee of the psychological sciences research institute of the university of the first author. See

[Supporting Information](#) for the CONSORT Checklist. Table S1 displays the gender, age, and grade of the students in each of the three groups. Specific culturally related constraints in the demographic data collection are also described.

Interventions

[Supporting Information](#) contains a full description of the interventions. The French version of all material provided to the teachers, including a detailed description of each session and how it unfolds, can be found on OSF (link in the Acknowledgements). Teacher training is detailed as [Supporting Information](#).

Intervention 1: Moral disengagement component

The first intervention used in our study is inspired by the Bullying Literature Project-Moral Disengagement Version (hereafter BLP-MD; Wang & Goldberg, 2017), which is a five-session class-wide intervention that uses storybooks displaying bullying situations to prevent bullying among elementary children. A pilot study of the BLP-MD revealed a significant decrease in both moral disengagement and victimization in intervention classes in 7–9 years old elementary students (Wang & Goldberg, 2017). [Supporting Information](#) provides a full description of how we adapted the original BLP-MD to our research goals (e.g., not to manipulate any other processes involved in bullying) and age group, as well as how a typical session unfolded. We created new stories of around 1800 words each, which were adapted to the age of the students in our sample and designed to represent various bullying situations. The stories were intended to illustrate the use of moral disengagement mechanisms by the bullies. Students were invited to identify the moral disengagement mechanisms (presented as “little lies” on which the bullies relied to avoid guilt), which were then invalidated by the teachers by reminding the students that moral disengagement mechanisms are false excuses that do not make the harmful behavior more acceptable. Just like in the original program, every session was closed with an individual writing activity (e.g., writing the end of the story, making a bookmark to illustrate the mechanisms). The English translation of the stories can be found as [Supporting Information](#).

Intervention 2: Social norms component

The second intervention used in our study is based on the “Survey of Bullying at Your School” project, which is a five-session class-wide intervention using students'

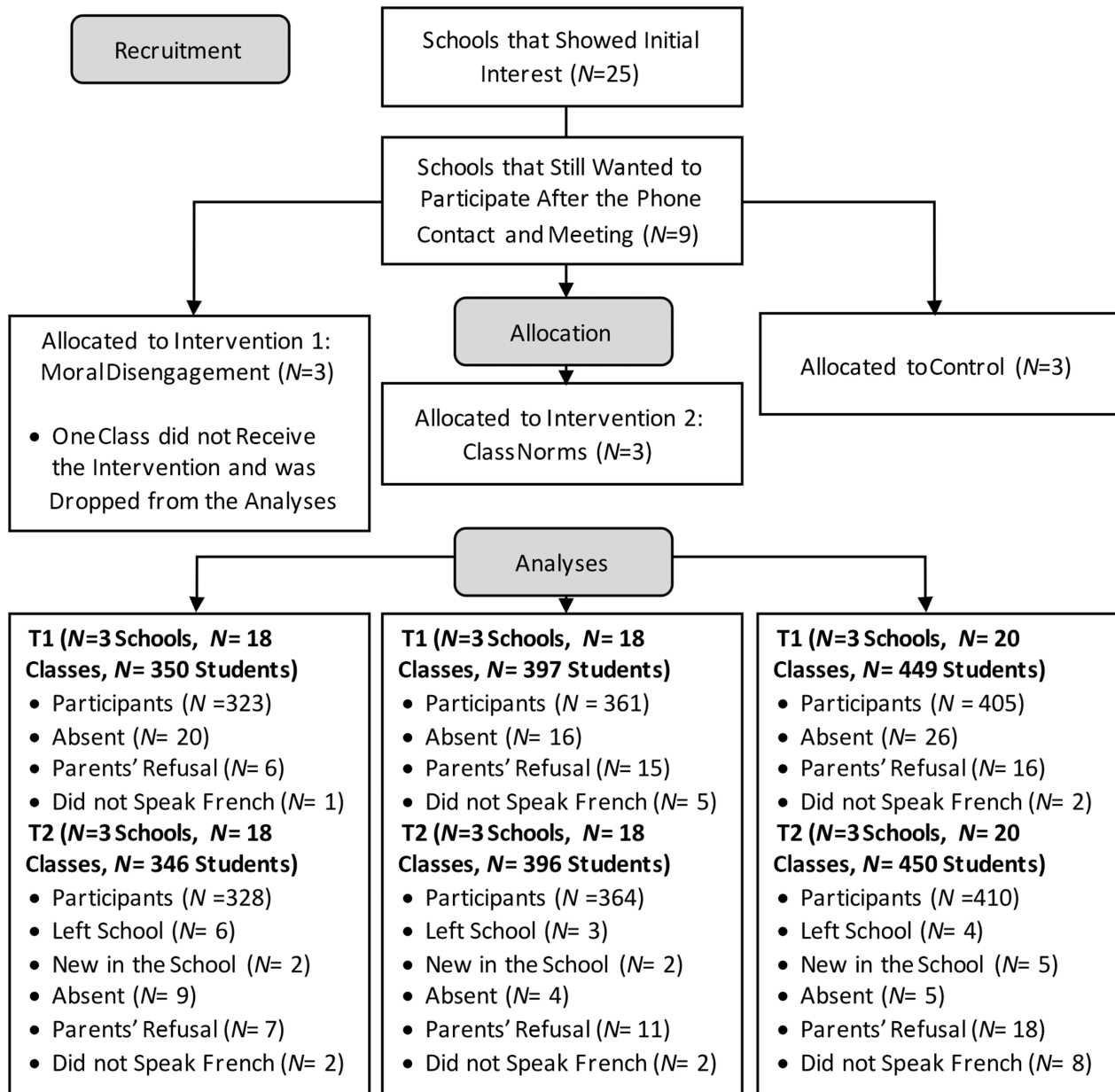


FIGURE 2 Flowchart of recruitment and retention of intervention and control schools. T1, Time 1; T2, Time 2.

online self-reported data to highlight the discrepancy between perceived and actual pro-bullying norms through norm-based messages displayed on posters. A pilot study of the project in 11–14 years old students revealed a significant decrease in both perceived pro-bullying injunctive norms and bullying and victimization (Perkins et al., 2011). [Supporting Information](#) provides a full description of how we adapted the original project to make the students play a more active role in the procedure, in a more entertaining, instant, and offline format, better suited to our participants' age.

During the first session, students were invited to report both their perceptions of the injunctive class norm (“Most of my classmates think that [...] is OK”) and their corresponding personal attitude (“I think that [...] is

OK”) toward some common bullying behaviors (e.g., excluding someone) anonymously. Students were split into work groups and asked to count the number of yes and no answers. The final counts were displayed on two parallel posters for each behavior to visualize and underline the discrepancy between students' perceived attitudinal class norm (e.g., “12 of 20 students thought that most classmates regarded excluding someone as OK”) and their actual class norm (“Only two out of 20 students thought it was OK”). In the following sessions, groups were invited to create by themselves a slogan to describe their class's *actual* norms (e.g., “In our classroom, we include everyone”). Eventually, groups were invited to make an artistic poster with the slogan for display in the school's corridors. In the final session of the intervention, each

class presented the posters to students of another class, and vice-versa.

Control intervention

As mentioned above, schools from the control group received an intervention of similar duration and intensity on a topic unrelated to bullying, that is climate and environmental issues. The themes of the five sessions were water, food chain, biodiversity, insects, and climate change. Each session included moments of reflection (e.g., “Where does water come from?”), exploration activities (e.g., matching successive steps of development from egg to adult stage of different insects such as ladybug, dragonfly, or butterfly), short videos (e.g., “What is the ozone layer?”), and creative activities (e.g., by group, collective paintings of animals living in a specific biotope, such as the forest, desert, or sea). Interested readers can find more details about this intervention through the material and detailed description of each session on OSF (link in the Acknowledgements).

Measures

Moral disengagement

We used the French-version of the Moral Disengagement in Bullying Scale (Fr-MDBS; Tolmatcheff et al., 2022), adapted from Thornberg and Jungert (2014). The mechanisms of dehumanization and blaming the victim are merged into one factor, yielding seven first-order factors (Thornberg & Jungert, 2014). The Fr-MDBS has been validated as a second order-factor structure, which entitles researchers to use either the global score of moral disengagement or its subscales, depending on their interests (Tolmatcheff et al., 2022). Because we aimed to investigate the general reduction in moral disengagement (rather than in specific mechanisms) following the related intervention, we used the global score of moral disengagement. As the questionnaire's length needed to be limited, we removed four out of the 18 items to keep only two items per factor. Students were asked to indicate to what degree they agreed or disagreed with each of the 14 statements (from 0 = *totally disagree* to 4 = *totally agree*), for example, “If people are weird, it is their own fault if they get bullied”. Following recommendations for estimating reliability coefficients with ordinal Likert-type scales, we used the polychoric correlation matrix to compute the ordinal Cronbach's alphas (Gadermann et al., 2012): α was = .84 at T1 and .87 at T2. The results of the confirmatory factorial analysis (CFA) showed a good fit for the higher-order structure: $\chi^2(70, N = 1094) = 120.7, p < .001$, comparative fit index (CFI) = .97, Tucker–Lewis index (TLI) = .97, root mean square error of approximation (RMSEA) = .026, 95% CI:

.018–.033, and standardized root mean square residual (SRMR) = .027. Because of a small negative residual variance for one of the first-order factors (advantageous comparison) at T2, this factor variance was fixed to a small positive value (0.1) to avoid convergence problems.

Perceived injunctive class norm

We translated into French eight items from Salmivalli and Voeten (2004) measuring attitudes toward bullying, for example, “Bullying may be fun sometimes” (from 0 = *totally disagree* to 4 = *totally agree*). Following the original pilot study procedure for measuring the perceived injunctive class norm, we instructed the students to indicate what they believed that most of their classmates would answer, instead of their own opinion (Perkins et al., 2011). Two items were difficult for the students to understand and were deleted to improve the scale's internal consistency. Ordinal Cronbach's alpha for the six remaining items was $\alpha = .78$ at T1 and .80 at T2. As four of the six items were reversed, a method-specific factor was added to the measurement model to take account of this shared specific variance. The results of the CFA showed a good fit: $\chi^2(5, N = 1091) = 9.5, p > .05$, CFI = .99, TLI = .98, RMSEA = .029, 95% CI: .000–.056, and SRMR = .013.

Bullying

We used a 10-item scale intended for primary school students (Tolmatcheff et al., 2020), which assessed direct (3 items: insulting, hitting or kicking or slapping, biting or scratching or pulling hair), indirect (4 items: excluding, gossiping, hiding or damaging personal things, teasing), and cyber (3 items: insulting or intimidating through the Internet [Snapchat, Facebook, Instagram, TikTok, ...]; calling or texting; releasing pictures or videos) bullying behaviors. Students had to indicate how many times (from 0 to 4 = *4 times or more*) they had exhibited these behaviors toward another student over the last 3 months, for example, “I have hidden or damaged another student's things on purpose”; “I have shared a picture or video of a student with others to make fun of him or her.” We used a bifactor model to assess both the general construct of bullying shared by the facets and the specific facets (indirect, direct, and cyber) simultaneously (Chen et al., 2012). The CFA results demonstrated a good fit: $\chi^2(29, N = 1097) = 34.8, p > .05$, CFI = .99, TLI = .99, RMSEA = .014, 95% CI: .000–.028, and SRMR = .025. Ordinal Cronbach's alphas were $\alpha = .85$ (T1) and .83 (T2) for the global construct, $\alpha = .74$ (T1) and .75 (T2) for indirect bullying, $\alpha = .81$ (T1) and .81 (T2) for direct bullying, and $\alpha = .85$ (T1) and .84 (T2) for cyber bullying.



Defender behavior

We adapted three items based on the shortened version of the participant roles scale (Sutton & Smith, 1999), which we formulated as self-reported (instead of peer-reported) statements. For example, the description “Comforts the victim” was turned into “I comfort the victim.” The two other items exploring students’ potential defending behavior when witnessing a bullying episode were as follows: “I ask an adult or peer for help” and “I try to help the victim.” Students had to indicate how often they behaved as described in each item (0 = *never*, 1 = *seldom*, 2 = *sometimes*, 3 = *usually*, 4 = *always*). Ordinal alpha was $\alpha = .67$ at T1 and $.73$ at T2.

Outsider behavior

Students were asked about their outsider behavior when witnessing a bullying episode. We adapted three items based on the shortened version of the participant roles scale (Sutton & Smith, 1999), formulated as self-reported statements, like for the defender behavior items. The three items were as follows: “I pretend not to notice what is happening,” “I don’t look in that direction or I stay away,” and “I don’t do anything” (from 0 = *never* to 4 = *always*). Ordinal alpha was $\alpha = .70$ at T1 and $.79$ at T2.

Analytical strategy

We first ran variance analyses (ANOVA) to compare the three groups at baseline. Prior examination of the Q-Q plots had indicated substantial deviation from normality for bullying, outsider behavior (positively skewed), and perceived injunctive class norm (negatively skewed). Therefore, we performed a square root transformation on these variables beforehand. In addition, we ran non-parametric Kruskal–Wallis tests on the nontransformed variables to check the results’ consistency. Table A1 (Appendix) provides the results of these preliminary analyses.

We performed structural equation modeling with MPlus version 8.4. In this case, the default estimation with complex survey data is maximum likelihood robust, which provides parameter estimates that are robust to non-normality and nonindependence of observations, and is also a leading technique for handling missing data, using full-information maximum likelihood (Enders, 2010). Examination of the intraclass coefficients of all dependent variables (provided as Supporting Information, see Table S3) showed that the between-school variability was negligible. Therefore, only the class level was taken into account in the analyses. In the final model, data were missing on all observed dependent variables in 39 cases (3.2%), which

Mplus automatically excluded from the analyses. The covariance coverage of data was above 92% for all the indicators.

Change between the pre and posttest of both mediating and outcome variables was measured using latent change (LC) models. LC models rely on the breakdown of the latent state variable at Time 2 (State 2) into initial state (State 1) plus change, that is, the latent difference variable (State 2 – State 1; see Figure 3). Latent difference variables represent interindividual differences in true (error-free) intraindividual change over time and can be used like any other latent variable (Geiser, 2013). Method effects of indicators are likely to appear in longitudinal analyses, because indicators share a specific variance with themselves over time, which violates the assumption of uncorrelated residuals. To overcome this issue, we allowed the residuals of the same indicator to correlate over time (Geiser, 2013).

As LC models require at least strong factorial invariance, we tested measurement invariance across time for all dependent variables (provided as Supporting Information, see Table S4). We imposed progressive equality constraints on factor loadings, intercepts, and residual variances and tested the fit of each nested model against the previous (less restrictive) one for all dependent variables (Geiser, 2013). As the Chi-square (χ^2) difference test is highly sensitive to sample size and can thus lead to model rejection even when discrepancies are trivial, specific recommendations have been developed for comparing pairs of nested models in large samples ($N > 300$) in the context of measurement invariance. Weak noninvariance is indicated by a CFI difference greater than .01 supplemented by a difference in the RMSEA greater than .015 or a difference in SRMR greater than .03. The same prevails for strong noninvariance, except that the difference in SRMR should be greater than .01 (Chen, 2007). Although only strong invariance is a prerequisite for a meaningful interpretation of latent mean change over time, we selected the strict invariance model when it held because it is most parsimonious while still representing the data adequately (Geiser, 2013).

Next, we ran a latent path analysis to examine the distinct effects of the interventions on the outcomes of interest through the hypothesized mediators. Because the categorical predictor (X) had three levels (a control group vs. two experimental conditions), we used contrast coding with the control group as the reference category to estimate the direct and indirect effects with a multicategorical X (Hayes, 2018). The final mediation model included the two dummy-coded binary predictors (moral disengagement and social norms components), the two mediating latent difference variables (change in moral disengagement and change in perceived injunctive class norm), and the three outcome latent difference variables: change in bullying, defender behaviors, and outsider behaviors (see Figure 3). To assess the indirect effects’

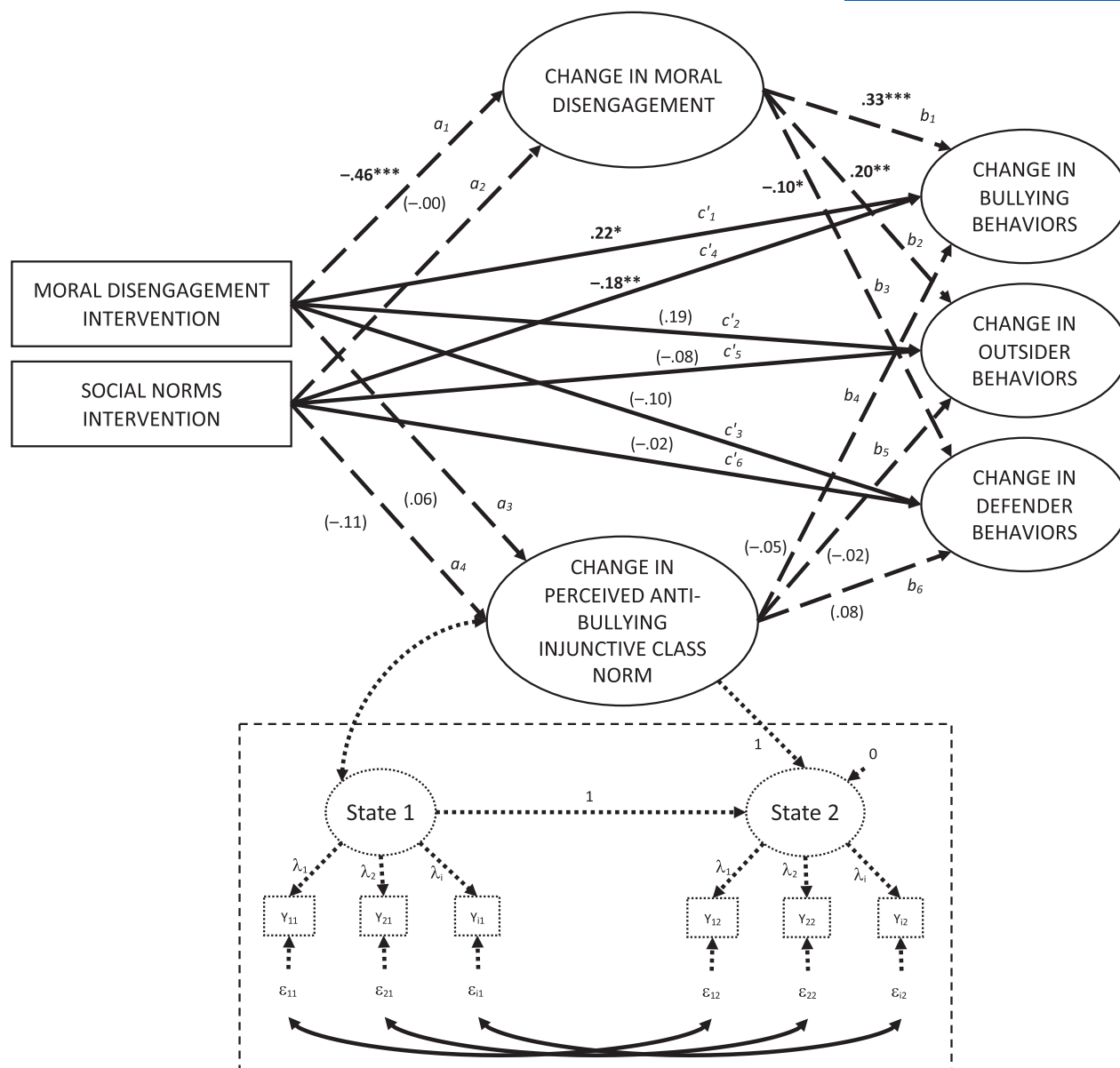


FIGURE 3 Effects of intervening on moral disengagement and social norms on Students' bullying, outsider, and defender behaviors. Each latent variable (indicated by an ellipse) displayed on the graph implies a latent change model with autocorrelated errors. However, only the LC model of perceived anti-bullying injunctive class norm is represented in the graph for the sake of clarity. $*p < .05$. $**p < .01$. $***p < .001$.

significance, we used the asymmetric bootstrap method, which does not assume the normality of the sampling distribution of the indirect effect; 200 bootstraps were required to obtain the 95% confidence intervals. Finally, we conducted additional exploratory analyses to examine the effects of the two interventions on the three specific facets of bullying (indirect, direct, and cyber) as well as the influence of gender and age as covariates.

RESULTS

Table A1 (see Appendix) displays the descriptive statistics (means and standard deviations) of the mediating and outcome variables. **Supporting Information**

provides correlations between all latent variables at T1 (see Table S2). Despite the randomization process, preliminary analyses pointed to significant between-group differences for some variables of interest at baseline. Students in the social norms condition initially reported slightly less bullying behaviors than the other groups. Students in the moral disengagement condition initially reported slightly more outsider behaviors than the other groups. Finally, students in the control group reported slightly more perceived classmates' anti-bullying attitudes than students in the social norms condition. However, this effect was only marginally significant according to the Bonferroni post hoc test ($p = .049$). As indicated by partial eta squared, all three effects can be considered as small.



Strong factorial invariance was reached for all dependent variables, allowing for a meaningful interpretation of latent mean change over time. [Supporting Information](#) provide details of the measurement invariance testing (see Table S4). [Figure 3](#) shows the unstandardized regression coefficients for the full mediation model. Parameter estimates between brackets do not differ significantly from zero. Significant regression coefficients ranged from small to medium effect size. Covariances between the factors are not displayed for the sake of clarity. The model demonstrated good model fit: $\chi^2(2605, N = 1166) = 18,339.6$, $p < .001$, CFI = .91, TLI = .91, RMSEA = .021, 95% CI: .020–.022, and SRMR = .046. The percentage of variance explained by the full model (as indicated by the R^2) was 21.3% for the change in moral disengagement, 13.9% for the change in perceived anti-bullying injunctive class norm, 29.8% for the change in bullying, 25.0% for the change in outsider behaviors, and 17.0% for the change in defender behaviors.

Moral disengagement component

The results indicated that intervening on moral disengagement significantly decreased moral disengagement in students ($a_1 = -.46$). Change in moral disengagement, in turn, was positively linked to change in bullying ($b_1 = .33$) and outsider behaviors ($b_2 = .20$)—thus a decrease in moral disengagement was associated with a decrease in bullying and outsider behaviors. Furthermore, change in moral disengagement was negatively linked to change in defender behaviors ($b_3 = -.10$)—thus, a decrease in moral disengagement was associated with an increase in defender behaviors. The standardized indirect effect on the outcome variables flowing through moral disengagement was significant in all three cases (bullying 95% CI $[-.15, -.05]$, outsider behaviors 95% CI $[-.08, -.02]$, and defender behaviors 95% CI $[.00, .06]$). The results also displayed a significant positive direct effect on the change in bullying behaviors ($c'_1 = .22$; $p < .05$), as shown in [Figure 3](#). However, this effect was not robust, as it was not significant when using the asymmetric bootstrap 95% CI procedure $[-.04, .21]$.

Social norms component

The social norms intervention had a significant direct effect decreasing bullying ($c'_4 = -.18$). However, the perceived injunctive class norm did not mediate this effect. More specifically, the intervention did not yield change in the hypothesized mediator (a_4), which, in turn, did not predict change in any of the outcome variables (b_4 , b_5 , and b_6). Accordingly, none of the indirect effects were significant.

Additional analyses

Additional exploratory analyses yielded similar conclusions. Change in moral disengagement predicted change in indirect ($\beta = .30$), direct ($\beta = .27$), and cyber ($\beta = .17$) bullying. The corresponding standardized indirect effects were significant for indirect (95% CI $[-.22, -.09]$) and direct bullying (95% CI $[-.22, -.08]$), and marginally significant for cyber bullying (95% CI $[-.16, .00]$). Intervening on moral disengagement had a significant direct effect increasing indirect bullying ($\beta = .23$), but not direct or cyber bullying. In addition, the significant direct effect of the social norms intervention related to both indirect ($\beta = -.17$) and direct bullying ($\beta = -.17$), but not to cyber bullying. The percentage of variance explained by the full model (as indicated by the R^2) was 31.7% for the change in indirect bullying, 24.7% for the change in direct bullying, and 42.9% for the change in cyber bullying.

As gender and age can be related to bullying and bystander behaviors, we checked whether the results differed when we controlled for them in the final model. The results were similar to the main analyses. The model controlling for these covariates had a slightly less favorable fit, which is why they were not kept in the final model.

DISCUSSION

Because most intervention studies assess anti-bullying programs' components in combination and omit to test the hypothesized mediators, our understanding of how these programs work and our knowledge of the psychosocial developmental processes involved in bullying remain limited. Additionally, because of the limited effectiveness of anti-bullying programs, we need to test innovative interventions that target under-explored relevant processes involved in bullying. This study aimed to address these methodological gaps to expand our knowledge of the influence of moral disengagement and social norms on bullying, as well as determine their potential as targets of anti-bullying program components.

Our results showed that intervening on moral disengagement decreased students' use of moral disengagement mechanisms, supporting our first hypothesis. It is noteworthy that the moral disengagement component had no effect on the other mediator, the perceived injunctive class norm toward bullying. This supports the idea that the intervention's design successfully isolated moral disengagement and did not incidentally elicit another anti-bullying component. The reduction in moral disengagement, in turn, led to a decrease in bullying and outsider behaviors, and an increase in defending, supporting our other hypotheses. The three corresponding indirect effects were significant, confirming the mediating role of moral disengagement. Additional analyses

revealed that the indirect effects of the intervention were significant for indirect and direct bullying, and marginally significant for cyber bullying. This might be because of the low prevalence of cyber bullying in our sample, which is consistent with prevalence rates generally found among elementary students (e.g., Salmivalli, 2021). These analyses also indicated that the direct effect of the intervention increased indirect (but not direct or cyber) bullying. Although this effect was not robust, it may suggest that some of the students—most likely, some of the bullies—were less receptive to the intervention. These less responsive bullies may have adjusted their behavior by turning to more hidden forms of bullying in reaction to the general change in both moral disengagement and bullying among their classmates.

Regarding the social norms component, our hypotheses were refuted. Although the intervention decreased both indirect and direct bullying, this effect was not mediated through the hypothesized mediator—the perceived injunctive class norm toward bullying—thus, leaving the question of the processes involved open. While providing people with correct information about the actual group norm has been proven effective in rectifying their misperception in various norm-based interventions (Miller & Prentice, 2016), including anti-bullying interventions in early adolescents (11–14 years old; Dillon & Lochman, 2019; Perkins et al., 2011), this was not the case in our study. In addition, no direct effect of the intervention was found for outsider or defender behaviors. It could be that, despite the intervention, standing up to the bullies was too risky for bystanders, who may have feared for their own safety and status (e.g., becoming the next victim). Bullies, on the other hand, may have been more sensitive to the intervention as they are particularly driven by dominance goals (e.g., Sijtsema et al., 2009) and would likely not want to risk their social status, thus being more prone to adjust their behavior to align with the peer norm.

A possible explanation for the absence of change in students' perceptions of the injunctive class norm is that the effect did not flow through the expected mediator. For instance, perhaps the intervention did not lead students to cognitively reconstruct their estimate of the norm, but simply weakened their global perceptions of peer support toward bullying (Miller & Prentice, 2016; Prentice, 2008). It is possible that students grasped the idea that general approval toward a pro-bullying norm within their class was illusory, which might have been enough to initiate a pattern of behavior that became self-sustaining (Prentice, 2008). Similar instances of behavior change occurring without change in norm perception have previously been found in some drinking norm-based interventions (Prentice, 2008).

Alternatively, in the case that bullies have been primarily impacted by the intervention, change in students' bullying behaviors would thus have been mediated by bullies' perceptions of the class injunctive norm toward

bullying. As bullying behaviors are often endorsed by popular peers (e.g., Vaillancourt et al., 2003), this is also in line with previous research suggesting that popular students are powerful norm setters (e.g., Laninga-Wijnen et al., 2017; Peets et al., 2015; Veenstra & Lodder, 2022). Following this explanatory hypothesis, a remaining question is whether this relatively short-term effect on popular students' perceptions and bullying behaviors would have translated into a change in bystanders' perceptions of the injunctive norm—and subsequently in their behaviors—in the long term.

Among the other types of norms that may have mediated the effect of the intervention, the subgroup of friends may also be seen as more influential than the whole class group (e.g., Paluck & Shepherd, 2012; Rigby & Johnson, 2006). Alternatively, teachers also convey social norms among their students, both explicitly and implicitly (Veenstra & Lodder, 2022). It is also possible that the intervention did not impact any type of norm, but rather a totally different mediator. For instance, the intervention could have raised students' awareness of their own anti-bullying attitudes, without necessarily changing students' perceptions of their classmates' attitudes.

Finally, another possible explanation is related to the implementation of the intervention—that is, how the intervention was delivered in practice by the teachers. As norms are believed to be a powerful vehicle of change (e.g., Miller & Prentice, 2016), key messages of norm-based interventions need to be accurately delivered. It is possible that a distorted message might even have a detrimental effect on students' perception of the injunctive class norm. If some of the teachers did not deliver the class norms-based messages as intended, a partial negative effect on the perceived norm could conceal the potential effectiveness of the intervention when implemented correctly.

Strengths and limitations

Strengths of this study are the relatively large sample with high retention rates, the use of LC models, the control for the nested structure of the data, the assessment of specific anti-bullying components in isolation, and the assessment of the hypothesized mediators. With respect to this last point, in particular, this study illustrates the importance of measuring the mediating paths for evaluation of anti-bullying programs (Volk et al., 2017). The significant indirect effects found for the moral disengagement intervention would not have been detected without mediation, because they would have been concealed in the total effects at the distal level (Rucker et al., 2011). By contrast, in the case of the social norms intervention, we would have wrongly assumed that the hypothesized (but not assessed) mediator—the perceived injunctive class norm—mediated change in bullying.



Despite its strengths, this study also has some limitations. First, despite the randomization process, some small yet significant between-groups differences were present at baseline. To minimize these initial differences, researchers can use a so-called stratification randomization procedure, in which allocation takes account of a well-chosen set of initial average levels of, for instance, bullying or anti-bullying attitudes. However, such a process can be tricky when there are relatively few schools per condition like in our study, as it does not always give enough latitude in making a proper allocation.

Second, because all measures were self-reported, results may have been partially inflated because of shared method variance across the mediators and outcome variables. We could have analyzed peer reports of bullying and bystander roles in addition to self-reports. Such a multi-informant design may compensate for their respective biases (see Hymel & Swearer, 2015; Volk et al., 2017).

Third, the absence of a follow-up limits our knowledge of the longer-term effects of the interventions. In the social norms intervention, in particular, a third wave of data collection could have helped clarifying whether the short-term effect on bullying behaviors would subsequently influence bystanders' normative perceptions and behaviors. In the moral disengagement intervention, an additional wave of data collection could have provided insights in the persistence of the effects over time.

Fourth, because this study was not replicated across different sociocultural groups (e.g., in other countries), our results may not generalize to other contexts. Beyond intercultural variations in bullying terminology and definition, cultural and societal differences have been suggested to influence the nature and manifestations of bullying behaviors (e.g., in Eastern vs. Western cultures; Smith et al., 2016), as well as the effectiveness of anti-bullying programs (Gaffney et al., 2019b). However, the BLP-MD's pilot study was originally conducted among American elementary students (Wang & Goldberg, 2017) and yielded similar results to ours. It is thus entirely possible that the cultural context does not threaten our findings' replicability. Nonetheless, both Belgium and the United States belong to the so-called WEIRD—Western, educated, industrialized, rich, and democratic—countries, which represent only a small percentage of the world's population (Henrich et al., 2010).

Finally, this study did not take the implementation of the interventions into account. It is now recognized that implementation of school-based prevention programs in natural contexts varies, especially when delivered by teachers, and that these fluctuations are likely to influence programs' outcomes (Durlak & DuPre, 2008). That most evaluation studies do not take into account implementation (Haataja et al., 2014) may thus also contribute to the “black box” of anti-bullying programs that limits our understanding of how these programs work and the processes involved in bullying. Providing the teachers with a detailed

description of how each session was supposed to unfold may have helped enhance implementation quality, but may not have been sufficient to prevent implementation discrepancies between teachers.

Implications and future directions

Determining the potential of anti-bullying components targeting specific processes involved in bullying is important for two reasons. First, as school bullying represents a serious public health issue, it is critical to tackle it using the most effective and cutting-edge interventions. Second, given the generally limited time and financial resources of schools, it is crucial to identify effective anti-bullying components to design prevention programs with the best possible cost-effectiveness ratio (Menesini & Salmivalli, 2017). This intervention study provides sound evidence supporting the role of moral disengagement not only in bullying, but also in outsider and defender behaviors. Beyond the context of a structured anti-bullying intervention, knowledge of the moral disengagement mechanisms could enable teachers and school staff in general to detect and invalidate them in students when dealing with a bullying situation (Tolmatcheff et al., 2022). In addition, future studies may investigate whether targeting moral disengagement could be effective in tackling cyber bullying in older students.

We hypothesized that some of the bullies adjusted their bullying behavior in reaction to their classmates' change in both moral disengagement and bullying. This finding might also be related to the recent concept of collective moral disengagement, which refers to individuals' perceptions of the extent to which moral disengagement mechanisms are shared within the group (rather than the sum of each individual's moral disengagement in the group; Thornberg, Wänström, et al., 2021). Collective moral disengagement was found to be associated with bullying in recent studies, even after controlling for individual moral disengagement (Thornberg, Wänström, et al., 2021). We suggest that collective moral disengagement may be seen as a perceived norm regarding common or approved moral disengagement mechanisms regarding bullying within the class. The theoretical articulation between individual moral disengagement, collective moral disengagement, and social norms should be further developed.

Even though the causal path of our social norms intervention remains unknown, it had an impact on students' bullying. As such, it is a promising anti-bullying component to further explore. Nevertheless, we still poorly understand the underlying processes of social norms (Veenstra & Laninga-Wijnen, 2022), and research is needed that determines the differential effects of different types of norms (Veenstra & Lodder, 2022).

Future studies should also examine the differential effects of both interventions on the different participant

roles involved in bullying. Differential effects of anti-bullying interventions are also likely to occur for different age groups (Yeager et al., 2015). Both moral judgment and peer influence evolve over the course of development. As children grow, they become increasingly capable of abstract reasoning, and progress from external to increasing self-regulation overall (Bandura, 1991). The use of moral disengagement mechanisms toward bullying is conditioned by students' moral standards, which are adopted, revised, modified, and replaced over the years (Bandura, 1991). Intervening on moral disengagement could thus be increasingly effective with age. This is important for prevention, as meta-analyses have indicated that the effectiveness of anti-bullying programs declines with older adolescents (Yeager et al., 2015). Similarly, intervening on class norms may be more effective in older students, as it has been suggested that the tendency to conform and peer influence increase during adolescence (Sandstrom & Bartini, 2010). Researchers have recently called for studies on how peer norms influence varies developmentally (Veenstra & Lodder, 2022).

By addressing the two methodological gaps, this study aims to provide guidelines for the assessment of anti-bullying components targeting specific processes involved in bullying and to set the course for future evaluation studies. We advocate for a real “return to the roots” for anti-bullying programs to determine the potential of their components and design programs with the best cost-effectiveness ratio. In a second step, relevant components should likely be combined to maximize the effectiveness of interventions. However, the inclusion of each additional component, therefore, needs to be carefully weighted in terms of its cost-effectiveness from a public health perspective (Menesini & Salmivalli, 2017). Finally, future studies should carefully monitor the implementation of anti-bullying interventions by teachers and examine its influence on the outcomes, as this gap likely also contributes to the “black box” of anti-bullying programs.

Overall, our findings provided additional evidence for the involvement of moral disengagement in both bullying and bystander behaviors and indicated that moral disengagement can be easily reduced in early adolescents. However, the intervention itself still needs to be refined to prevent the concurrent increase in indirect bullying in students who did not decrease in moral disengagement. In sum, our findings suggest that a component targeting moral disengagement may be relevant to incorporate in anti-bullying programs. Intervening on social norms is also a promising direction for prevention as the intervention decreased bullying, but further research is needed on the mechanisms through which social norms reduce bullying.

ACKNOWLEDGMENTS

The analyses presented here were not preregistered. The data and code are available from the first author upon

reasonable request. Materials are available at the following URL: https://osf.io/w5hd7/?view_only=693d6d86743d4c49980a7768c4e555b.

FUNDING INFORMATION

This research was funded by a PhD grant (FC23715) from the National Fund for Scientific Research (F.R.S.-F.N.R.S.).

CONFLICT OF INTEREST

None declared.

ORCID

Chloé Tolmatcheff  <https://orcid.org/0000-0002-5833-9176>

Benoit Galand  <https://orcid.org/0000-0003-3387-4305>

Isabelle Roskam  <https://orcid.org/0000-0002-1449-1133>

René Veenstra  <https://orcid.org/0000-0001-6686-6307>

REFERENCES

- Álvarez-García, D., García, T., & Núñez, J. C. (2015). Predictors of school bullying perpetration in adolescence: A systematic review. *Aggression and Violent Behavior*, 23, 126–136. <https://doi.org/10.1016/j.avb.2015.05.007>
- Bandura, A. (1991). Social cognitive theory of moral thought and action. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development* (Vol. 1, pp. 45–103). Erlbaum.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 464–504. <https://doi.org/10.1080/10705510701301834>
- Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, 80, 219–251. <https://doi.org/10.1111/j.1467-6494.2011.00739.x>
- Dillon, C. E., & Lochman, J. E. (2019). Correcting for norm misperception of anti-bullying attitudes. *International Journal of Behavioral Development*. <https://doi.org/10.1177/0165025419860598>
- 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. <https://doi.org/10.13110/merrpalmquar1982.61.1.0144>
- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41, 327–350. <https://doi.org/10.1007/s10464-008-9165-0>
- Eisner, M., & Malti, T. (2012). The future of research on evidence-based developmental violence prevention in Europe—Introduction to the focus section. *International Journal of Conflict and Violence*, 6, 166–175. <https://doi.org/10.4119/ijcv-2909>
- Enders, C. K. (2010). *Applied missing data analysis*. Guilford.
- Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research, and Evaluation*, 17, Article 3. <https://doi.org/10.7275/n560-j767>
- Gaete, J., Valenzuela, D., Rojas-Barahona, C., Valenzuela, E., Araya, R., & Salmivalli, C. (2017). The KiVa antibullying program in



- primary schools in Chile, with and without the digital game component: Study protocol for a randomized controlled trial. *Trials*, 18, 1–9. <https://doi.org/10.1186/s13063-017-1810-1>
- Gaffney, H., Ttofi, M. M., & Farrington, D. P. (2019a). Evaluating the effectiveness of school-bullying prevention programs: An updated meta-analytical review. *Aggression and Violent Behavior*, 45, 111–133. <https://doi.org/10.1016/j.avb.2018.07.001>
- Gaffney, H., Ttofi, M. M., & Farrington, D. P. (2019b). Examining the effectiveness of school-bullying intervention programs globally: A meta-analysis. *International Journal of Bullying Prevention*, 1, 14–31. <https://doi.org/10.1007/s42380-019-0007-4>
- 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>
- Geiser, C. (2013). *Data analysis with Mplus*. Guilford.
- Haataja, A., Voeten, M., Boulton, A. J., Ahtola, A., Poskiparta, E., & Salmivalli, C. (2014). The 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>
- Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis* (2nd ed.). Guilford.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29. <https://doi.org/10.1038/466029a>
- Hymel, S., & Bonanno, R. A. (2014). Moral disengagement processes in bullying. *Theory Into Practice*, 53, 278–285. <https://doi.org/10.1080/00405841.2014.947219>
- Hymel, S., Schonert-Reichl, K. A., Bonanno, R. A., Vaillancourt, T., & Henderson, N. R. (2010). Bullying and morality: Understanding how good kids can behave badly. In S. R. Jimerson, S. M. Swearer, & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 101–118). Routledge/Taylor & Francis Group.
- Hymel, S., & Swearer, S. M. (2015). Four decades of research on school bullying: An introduction. *American Psychologist*, 70, 293–299. <https://doi.org/10.1037/a0038928>
- 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>
- Killer, B., Bussey, K., Hawes, D. J., & Hunt, C. (2019). A meta-analysis of the relationship between moral disengagement and bullying roles in youth. *Aggressive Behavior*, 45, 450–462. <https://doi.org/10.1002/ab.21833>
- Kubiszewski, V., Auzoult, L., Potard, C., & Lheureux, F. (2019). Witnessing school bullying: To react or not to react? An insight into perceived social norms regulating self-predicted defending and passive behaviours. *Educational Psychology*, 39, 1174–1193. <https://doi.org/10.1080/01443410.2018.1530735>
- Laniga-Wijnen, L., Harakeh, Z., Steglich, C., Dijkstra, J. K., Veenstra, R., & Vollebergh, W. (2017). The norms of popular peers moderate friendship dynamics of adolescent aggression. *Child Development*, 88, 1265–1283. <https://doi.org/10.1111/cdev.12650>
- Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2014). Why ineffective psychotherapies appear to work: A taxonomy of causes of spurious therapeutic effectiveness. *Perspectives on Psychological Science*, 9, 355–387. <https://doi.org/10.1177/1745691614535216>
- Menesini, E., Palladino, B. E., & Nocentini, A. (2015). Emotions of moral disengagement, class norms, and bullying in adolescence: A multilevel approach. *Merrill-Palmer Quarterly*, 61, 124–143. <https://doi.org/10.13110/merrillpalmerquar1982.61.1.0124>
- Menesini, E., & Salmivalli, C. (2017). Bullying in schools: The state of knowledge and effective interventions. *Psychology, Health & Medicine*, 22, 240–253. <https://doi.org/10.1080/13548506.2017.1279740>
- Miller, D. T., & Prentice, D. A. (2016). Changing norms to change behavior. *Annual Review of Psychology*, 67, 339–361. <https://doi.org/10.1146/annurev-psych-010814-015013>
- Nishina, A., & Juvonen, J. (2005). Daily reports of witnessing and experiencing peer harassment in middle school. *Child Development*, 76, 435–450. <https://doi.org/10.1111/j.1467-8624.2005.00855.x>
- Paluck, E. L., & Shepherd, H. (2012). The salience of social referents: A field experiment on collective norms and harassment behavior in a school social network. *Journal of Personality and Social Psychology*, 103, 899–915. <https://doi.org/10.1037/a0030015>
- 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>
- Perkins, H. W. (2014). Misperception is reality: The “reign of error” about peer risk behaviour norms among youth and young adults. In M. Xenitidou & B. Edmonds (Eds.), *The complexity of social norms* (pp. 11–36). Springer.
- Perkins, H. W., Craig, D. W., & Perkins, J. M. (2011). Using social norms to reduce bullying: A research intervention among adolescents in five middle schools. *Group Processes & Intergroup Relations*, 14, 703–722. <https://doi.org/10.1177/1368430210398004>
- Pozzoli, T., & Gini, G. (2010). Active defending and passive bystander 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., & Vieno, A. (2012). The role of individual correlates and class norms in defending and passive bystander behavior in bullying: A multilevel analysis. *Child Development*, 83, 1917–1931. <https://doi.org/10.1111/j.1467-8624.2012.01831.x>
- Prentice, D. A. (2008). Mobilizing and weakening peer influence as mechanisms for changing behavior. In M. J. Prinstein & K. A. Dodge (Eds.), *Understanding peer influence in children and adolescents* (pp. 161–180). Guilford Publications.
- Rigby, K., & Johnson, B. (2006). Expressed readiness of Australian schoolchildren to act as bystanders in support of children who are being bullied. *Educational Psychology*, 26, 425–440. <https://doi.org/10.1080/01443410500342047>
- Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, 5, 359–371. <https://doi.org/10.1111/j.1751-9004.2011.00355.x>
- 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>
- Salmivalli, C. (2021, November 4–6). *Special session in collaboration with UNESCO: Cyberbullying. Workshop on Aggression 2021*, Turku, Finland. <https://woa2020.utu.fi/programme/>
- 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. [https://doi.org/10.1002/\(SICI\)1098-2337\(1996\)22:1<1::AID-ABI>3.0.CO;2-T](https://doi.org/10.1002/(SICI)1098-2337(1996)22:1<1::AID-ABI>3.0.CO;2-T)
- 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>
- Sandstrom, M., Makover, H., & Bartini, M. (2013). Social context of bullying: Do misperceptions of group norms influence children's responses to witnessed episodes? *Social Influence*, 8, 196–215. <https://doi.org/10.1080/15534510.2011.651302>
- Sandstrom, M. J., & Bartini, M. (2010). Do perceptions of discrepancy between self and group norms contribute to peer harassment at school? *Basic and Applied Social Psychology*, 32, 217–225. <https://doi.org/10.1080/01973533.2010.495645>
- Schaffer, H. R. (2006). *Key concepts in developmental psychology*. Sage.
- Sentse, M., Veenstra, R., Kiuru, N., & Salmivalli, C. (2015). A longitudinal multilevel study of individual characteristics and classroom norms in explaining bullying behaviors. *Journal of*

- Abnormal Child Psychology*, 43, 943–955. <https://doi.org/10.1007/s10802-014-9949-7>
- Sijtsema, J. J., Veenstra, R., Lindenberg, S., & Salmivalli, C. (2009). Empirical test of bullies' status goals: Assessing direct goals, aggression, and prestige. *Aggressive Behavior*, 35, 57–67. <https://doi.org/10.1002/ab.20282>
- Smith, P. K., Kwak, K., & Toda, Y. (2016). *School bullying in different cultures: Eastern and Western perspectives*. Cambridge University Press.
- Sutton, J., & Smith, P. K. (1999). Bullying as a group process: An adaptation of the participant role approach. *Aggressive Behavior*, 25, 97–111. [https://doi.org/10.1002/\(SICI\)1098-2337\(1999\)25:2<97::AID-AB3>3.0.CO;2-7](https://doi.org/10.1002/(SICI)1098-2337(1999)25:2<97::AID-AB3>3.0.CO;2-7)
- Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, 10, 181–211. <https://doi.org/10.1111/sipr.12022>
- Thornberg, R., & Jungert, T. (2014). School bullying and the mechanisms of moral disengagement. *Aggressive Behavior*, 40, 99–108. <https://doi.org/10.1002/ab.21509>
- Thornberg, R., Pozzoli, T., & Gini, G. (2021). 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*. <https://doi.org/10.1177/08862605211037427>
- Thornberg, R., Wänström, L., Gini, G., Varjas, K., Meyers, J., Elmelid, R., Johansson, A., & Mellander, E. (2021). Collective moral disengagement and its associations with bullying perpetration and victimization in students. *Educational Psychology*, 41, 952–966. <https://doi.org/10.1080/01443410.2020.1843005>
- Tolmatcheff, C., Bebiroglu, N., & Galand, B. (2020). L'aménagement des cours de récréation et l'organisation d'espaces de parole permettent-ils de réduire le harcèlement scolaire? *Revue française de Pédagogie*, 209, 129–147. <https://doi.org/10.4000/rfp.9846>
- Tolmatcheff, C., Galand, B., & Roskam, I. (2022). Validation of the French version of the moral disengagement in bullying scale: Testing Bandura's conceptual model. *Journal of School Psychology*, 91, 81–96. <https://doi.org/10.1016/j.jsp.2022.01.002>
- Vaillancourt, T., Hymel, S., & McDougall, P. (2003). Bullying is power: Implications for school-based intervention strategies. *Journal of Applied School Psychology*, 19, 157–176. https://doi.org/10.1300/J008v19n02_10
- van Versveld, M. D., Fekkes, M., Fekkink, R. G., & Oostdam, R. J. (2021). Effects of implementing multiple components in a school-wide antibullying program: A randomized controlled trial in elementary schools. *Child Development*, 92, 1605–1623. <https://doi.org/10.1111/cdev.13529>
- Veenstra, R., & Laninga-Wijnen, L. (2022). Peer network studies and interventions in adolescence. *Current Opinion in Psychology*, 44, 157–163. <https://doi.org/10.1016/j.copsyc.2021.09.015>
- Veenstra, R., & Lodder, G. M. A. (2022). On the microfoundations of the link between classroom social norms and behavioral development. *International Journal of Behavioral Development*. <https://doi.org/10.1177/01650254221100228>
- Volk, A. A., Veenstra, R., & Espelage, D. L. (2017). So you want to study bullying? Recommendations to enhance the validity, transparency, and compatibility of bullying research. *Aggression and Violent Behavior*, 36, 34–43. <https://doi.org/10.1016/j.avb.2017.07.003>
- Wang, C., & Goldberg, T. S. (2017). Using children's literature to decrease moral disengagement and victimization among elementary school students. *Psychology in the Schools*, 54, 918–931. <https://doi.org/10.1002/pits.22042>
- Yanagida, T., Strohmeier, D., & Spiel, C. (2019). Dynamic change of aggressive behavior and victimization among adolescents: Effectiveness of the ViSC program. *Journal of Clinical Child & Adolescent Psychology*, 48, S90–S104. <https://doi.org/10.1080/15374416.2016.1233498>
- Yeager, D. S., Fong, C. J., Lee, H. Y., & Espelage, D. L. (2015). Declines in efficacy of anti-bullying programs among older adolescents: Theory and a three-level meta-analysis. *Journal of Applied Developmental Psychology*, 37, 36–51. <https://doi.org/10.1016/j.appdev.2014.11.005>
- Yun, H. Y. (2020). New approaches to defender and outsider roles in school bullying. *Child Development*, 91, e814–e832. <https://doi.org/10.1111/cdev.13312>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tolmatcheff, C., Galand, B., Roskam, I., & Veenstra, R. (2022). The effectiveness of moral disengagement and social norms as anti-bullying components: A randomized controlled trial. *Child Development*, 00, 1–16. <https://doi.org/10.1111/cdev.13828>

APPENDIX

TABLE A1 Descriptive statistics and univariate analyses of variance at baseline

Outcome variables	Moral disengagement intervention						Social norm intervention						Control group						Baseline differences			
	T1			T2			T1			T2			T1			T2			F ratio	p	η^2_p	
	N	M	SD	N	M	SD	N	M	SD	N	M	SD	N	M	SD	N	M	SD				
Moral disengagement	321	1.14	0.65	327	0.75	0.66	367	1.08	0.59	380	0.99	0.67	403	1.09	0.65	423	1.02	0.62	0.39	.676	.001	
Perceived attitudinal class norm	321	2.80 ^{ab}	0.80	328	2.84	0.79	366	2.81 ^b	0.88	379	2.74	0.92	404	2.94 ^a	0.77	422	2.87	0.80	3.41*	.033	.006	
Bullying (self-reported)	322	0.62 ^a	0.63	328	0.62	0.61	367	0.42 ^b	0.46	382	0.41	0.45	405	0.52 ^a	0.54	423	0.54	0.57	9.54**	<.001	.017	
Indirect	322	0.84	0.86	329	0.86	0.82	365	0.63	0.68	382	0.59	0.66	404	0.77	0.74	424	0.77	0.79				
Direct	322	0.79	0.92	328	0.82	0.96	365	0.53	0.82	382	0.51	0.78	402	0.64	0.92	423	0.70	0.95				
Cyber	322	0.15	0.49	330	0.24	0.57	365	0.03	0.16	381	0.14	0.44	403	0.09	0.35	424	0.21	0.59				
Defender behavior	322	2.47	0.96	330	2.46	0.96	367	2.46	0.97	382	2.47	1.04	403	2.44	0.92	421	2.49	0.91	0.09	.917	.007	
Outsider behavior	319	0.95 ^a	0.93	329	0.82	0.92	365	0.71 ^b	0.84	380	0.62	0.78	402	0.73 ^b	0.80	420	0.68	0.80	8.37**	<.001	.015	

Note: Means with different superscripts differ at the $p = .05$ level in the Bonferroni post hoc test. Results from nonparametric Kruskal–Wallis tests were consistent with the ANOVA results on the square root transformed variables.

Abbreviations: T1, Time 1; T2, Time 2.

* $p < .05$; ** $p < .001$.