

Opening the black box of anti-bullying programs

Opening the black box of anti-bullying programs

An investigation of the effects, mediating roles, and implementation of moral disengagement and class norms as anti-bullying program components

Chloé Tolmatcheff

Promoteur(s)

Benoît GALAND (UCLouvain)
Isabelle ROSKAM (UCLouvain)

Thèse présentée en vue de
l'obtention du grade de Docteur
en sciences psychologiques et
de l'éducation

Comité d'accompagnement et jury

Emmanuelle ZECH (UCLouvain, Présidente de jury)
Benoît GALAND (UCLouvain, Secrétaire du jury)
Isabelle ROSKAM (UCLouvain, Secrétaire du jury)
Céline BAGÈS (Université de Lille, France)
René VEENSTRA (University of Groningen, The Netherlands)
Hilde COLPIN (KULeuven, Belgium)

Table of contents

Table of contents	5
Résumé	7
Abstract	7
Remerciements (acknowledgments)	9
General introduction	15
Validation of the French Version of the Moral Disengagement in Bullying Scale Testing Bandura's Conceptual Model	35
The Effectiveness of Moral Disengagement and Class Norms as Anti-Bullying Components A Randomized Controlled Trial	63
Anti-Bullying Interventions Delivered by Teachers Examination of the Influence of the Implementation Fidelity and Quality in a Randomized Controlled Trial	95
General discussion	129
References	148
Appendices	Erreur ! Signet non défini.

Résumé

La plupart des études d'intervention ne distinguent pas les différents composants des programmes anti-harcèlement, omettent de tester les médiateurs et ne tiennent pas compte de la mise en œuvre dans l'interprétation des résultats. Par conséquent, notre compréhension du fonctionnement des programmes et de leurs composants efficaces demeure limitée. En outre, l'efficacité restreinte des programmes existants révélée par les méta-analyses soulignent la nécessité de développer des interventions innovantes et reflétant les avancées scientifiques. Cette thèse de doctorat vise à réaliser un premier pas dans l'ouverture de la « boîte noire » des programmes anti-harcèlement en remédiant aux trois lacunes méthodologiques susmentionnées, ainsi qu'à déterminer l'efficacité de deux composants anti-harcèlement suggérés comme des pistes prometteuses pour la prévention mais jusqu'ici peu, voire pas exploités comme tels dans les programmes : le désengagement moral et les normes du groupe-classe.

Abstract

As most evaluation studies assess anti-bullying program components in combination, omit the testing of hypothesized mediating paths, and do not take implementation into account, our understanding of how these programs work and our knowledge of their effective components remain limited. In addition, the limited effectiveness of anti-bullying programs pointed out by meta-analyses emphasizes the need to develop innovative interventions that reflect scientific advances. This thesis aims to make a first step in opening the 'black box' of anti-bullying programs by addressing the three aforementioned methodological gaps, as well as to test the effectiveness of two anti-bullying components that have been suggested as promising directions for prevention but have rarely, if at all, been targeted as such in programs: moral disengagement and class norms.

Opening the black box of anti-bullying programs

Remerciements (acknowledgments)

Pfiuuu...

Ça y est, nous y voilà : la fin de thèse. Ce fut un chemin bien solitaire et pourtant, j'ai tant de personnes à remercier !

En premier lieu, bien évidemment, je souhaite remercier les Professeurs Benoît Galand et Isabelle Roskam, mes promoteurs de thèse. Benoît, un immense merci d'avoir accordé cette rencontre, il y a six ans, à une jeune psychologue sortie de l'U.L.B. un an auparavant, qui voulait faire un doctorat sur le harcèlement scolaire et ne savait pas vraiment par où commencer. Un plus grand merci encore de l'avoir recontactée, un mois plus tard, pour lui proposer un premier contrat d'assistante de recherche de quelques mois. Au fil des années qui ont suivi, tu n'as eu cesse de m'offrir des opportunités de me développer professionnellement. Six ans plus tard, on peut dire qu'elles auront porté leur fruits. Tu as été un mentor formidable et tu m'auras certainement transmis ta considération et ton attention sincères pour les besoins des acteurs de premier plan, ceux et celles qui font face, bien avant nous, chercheurs, aux situations de harcèlement scolaire en première ligne. Je suis heureuse et fière d'être ta collaboratrice, pour encore longtemps j'espère.

Isa, je mesure ma chance de t'avoir eu comme co-promotrice ! Ta rigueur scientifique et ton exigence, mais aussi la richesse de tes feedbacks, toujours si pertinents, m'ont permis de donner le meilleur de moi-même dans l'écriture de mes articles. Tu étais là pour redonner la priorité à ma thèse quand je m'éparpillais dans toutes sortes d'autres projets passionnants, me permettant au final d'en venir à bout dans le temps imparti, une victoire plus qu'appréciable. Merci du fond du cœur pour tes encouragements et félicitations tout au long du processus d'écriture ces deux dernières années ; ce fut sans aucun doute le meilleur des carburants.

Je voudrais également remercier les différents membres de mon jury, qui ont tous et toutes contribué non seulement à ma thèse, mais également à mon parcours de doctorante : les Professeurs Emmanuelle Zech, Céline Bagès, René Veenstra et Hilde Colpin. Un grand merci, Manu, pour tes apports très enrichissants, déjà même avant le début de ma thèse ! Tes ressources en matière de formation aux compétences d'aide et d'entretien m'auront été précieuses bien au-delà du cadre dans lequel je les avais sollicitées, notamment lorsqu'il s'est agi de gérer tant bien que mal des interactions parfois fort compliquées entre les différents acteurs impliqués dans ce projet. Un grand merci également, Céline, notamment pour m'avoir permis de te remplacer, le temps d'un quadrimestre, et de donner ton cours de Psychologie de la violence scolaire à l'Université de Lille, une première expérience de ce genre pour moi.

I want to thank René, in particular, for inviting me to stay in his research lab for four months. Although the pandemic made things a bit different than usual, this was a great opportunity to discover another work environment and meet others currently researching bullying. I was amazed by the quality of their work, and I have learned a lot as a result of

being there. Through your expertise and advice, you have made a considerable contribution to the quality of my thesis papers. I was also happy to learn that you are not only an excellent researcher but also a kind and benevolent person, who is a pleasure to work with. I am looking forward to going back to Groningen at the first opportunity available in order to continue our collaboration. I also want to thank Hilde for agreeing to be part of my doctoral jury and for her insightful comments and suggestions which helped me improve my thesis. I really appreciated our conversations, and I am looking forward to further collaboration with you in the future.

Je voudrais également sincèrement remercier les neuf directions¹, cinquante-sept enseignant·e·s² et mille-deux-cents et quelques élèves des écoles primaires qui ont accepté de prendre part à ce projet de recherche. Cette thèse n'aurait pas pu exister sans eux. Par ailleurs, je voudrais spécifiquement remercier Stéphane, Quentin, Marc, Sandra Melis, Violaine Dautrebande et Alain Nicolas ainsi que l'Association des Écoles Lasalliennes, qui m'ont introduite auprès de ces écoles et de leurs directions en premier lieu. Je vous remercie sincèrement pour votre confiance.

Ceci m'amène également à remercier particulièrement mes collègues Florence, Christine et Nathalie, qui m'ont accompagnée et épaulée pour former les équipes éducatives de plus de cinquante personnes lors des journées pédagogiques de remerciement qui ont suivi le projet. Merci également à Michaël, pour sa disponibilité et son aide face aux syntaxes MPlus, ainsi qu'à Thomas, pour ses conseils et nos discussions statistiques entre deux portes. Je remercie du fond du cœur l'ensemble de mes collègues, anciens et nouveaux, auprès de qui j'ai eu la joie d'évoluer pendant ces six dernières années : ma chère Virginie, Noémie, Morgane, Huy, Mickaël, Samir, Margot F., Claire, Élise, Margot D., Antoine, Joséphine, Maurice, Christelle, Simon, Amandine, Florence V., Charlie et ceux ou celles que j'oublie peut-être au moment d'écrire ces remerciements. Des midis d'équipe aux manifestations pour le climat, en passant par la VR et le compost, je peux difficilement imaginer plus chouettes collègues. Si la pandémie a malheureusement indiscutablement bouleversé notre vie d'équipe, je garde néanmoins en mémoire tous les formidables moments passés ensemble et me réjouis de ceux à venir.

Je tiens à remercier Emmeline, qui, bien plus qu'une stagiaire, s'est révélée être une précieuse collaboratrice sur laquelle j'ai véritablement pu m'appuyer. Merci également aux mémorantes qui ont contribué à la récolte et/ou à l'encodage des données, Célie, Jennifer,

¹ Nathalie Watrin et Garance Dombret, Valentine Donnay, Jean Bruynbroeck et Valérie Cacheux, Carine Lenaerts, Christine Toumpsin, Céline Martin, Fabienne Gaziaux, Sabine Brasseur, et Fabienne Tasco.

² Valentine Riga, Jessica Leemans, Christine Verschueren, Laurent Vanaudenhove, Anne T'sas, Benoît Hilbert, Gwenaél Denies, Valérie Rodrigues, Alizée Streibel, Antoine Vancryenveld, Chan Ngu, Valine Lees, Julie Urdinov, Champa Somsouthi, Frédéric Missoul, Régine Janssis, Martine Falla, Jean-Paul Piotto, Aurélie Coheur, Émilie Soutmans, Fanny Duquène, Carine Loox, Christophe Mathot, Anne Tasiaux, Olivier Defresne, Céline Haeghens, Mathieu Hebette, Anne-Françoise Caillou, Justin Colaux, Christel Delval, Nathalie Rouard, Hélène Daiwaille, Olivia Rabau, Odile Bette (†), Manon Faut, Aline Verly, Laeticia De Wolf, Morgane Toby, Benoît Weber, Delphine Derèse, Gaëlle Faway, Samantha Di Vito, Christelle Gehu, Emilie Wellens, Mathieu Lambillon, Marion Cnockaert, Najima Abakach, Line Daune, Valentin De Rouge, Vladimir Brouhns, Yvan Ortiz, Robert Nys, Jean-Pascal Bodart, Katty Rocha, Isabelle Goffin, Aubrey Debaty, Joëlle Pironet, Sophie Melot.

Joy et Laura. Je voudrais aussi remercier mes amies Evelyne Jadot et Anna Motyl pour les centaines d'heures passées ensemble en ligne à terminer d'écrire nos thèses respectives. Ces derniers mois de co-working ont réellement été salvateurs.

Pour finir, je voudrais remercier ma famille pour leur soutien sans faille tout au long de ces années. Merci avant tout à ma maman, ma plus grande fan, et à mon oncle Frédéric, pour n'avoir jamais douté de mes capacités. J'aurais bien sûr voulu que mon papa soit encore là pour vivre ce moment, mais je sais aussi qu'il avait la certitude tranquille que je réussirais. Merci à Antoine, pour son affection et son soutien à toute épreuve. Merci, enfin, à Thibault, pour tout. Pour m'avoir soutenue et poussée à dépasser mes limites, pour avoir été à mes côtés face à toutes les épreuves qui ont parsemé cette thèse, pour tous nos échanges et tes précieux conseils, et, plus généralement, pour partager ma vie et celle de Matcha. Cette thèse n'aurait pas été la même sans toi.

Chapter 1

General introduction

General introduction

1. Introduction

School bullying has received increasing attention from researchers, policymakers, education professionals, and the general public over these past decades. Despite being studied in depth for over two decades, bullying behavior remains difficult to define (Volk et al., 2017). The definition proposed by the pioneer in this field, Dan Olweus, in 1993 highlights the intentionality, the repetition, and the imbalance of power that characterize bullying. Recently, some authors have started questioning these three features, preferring to define bullying as a goal-directed rather than intentional behavior and emphasizing the perceived (subjective) harm rather than how often bullying occurs (Volk et al., 2017). Most of bullying research, nevertheless, is based on Olweus' seminal definition (Volk et al., 2017).

Olweus made a preliminary distinction between direct bullying, which includes relatively open forms of aggression such as hitting, insulting or threatening, and indirect bullying, which includes less visible forms of bullying such as deliberate exclusion from a group or social manipulation (1994). New communication technologies brought their own set of online forms of bullying (although far less frequently than traditional forms of bullying; Salmivalli & Peets, 2018), referred to as cyberbullying (Brighi et al., 2012). Although studies often consider an overarching dimension of bullying, these three subtypes of bullying have been found to be differentially associated with various factors such as empathy (e.g., Jolliffe & Farrington, 2006), gender (e.g., Salmivalli & Kaukiainen, 2004), consequences for psychosocial health (e.g., Baldry, 2004; van der Wal et al., 2003) or teacher intervention likelihood (e.g., Yoon & Kerber, 2003).

Beyond this theoretical definition, bullying is intimately related to the social dynamics of the peer group, as reflected in its original designation as 'mobbing' (in reference to the term 'mob' – an aggressive and violent crowd) in early writings on the topic. In the 90s, the social nature of bullying was particularly brought to the foreground (Salmivalli, 2010). Bullying was notably suggested to be a means for bullies to achieve status goals and a dominant position in the peer group hierarchy. As bullying mostly takes place in the presence of bystanders, their key role in either enhancing or discouraging bullies' behavior was acknowledged (Salmivalli, 2010). Different participant roles outside the bully-victim dyad were identified. Bystanders were initially classified into assistants (who join in the bullying but do not initiate it), reinforcers (who display encouraging behaviors, like laughing or cheering), outsiders (who are not directly involved in bullying and passively observe it or walk away), and defenders (who provide support to the victim or stand up to the bully) (Salmivalli et al., 1996). More recently, this categorization has met with some criticism because the roles of bully, assistant, and reinforcer have been found to be too highly correlated to be considered as distinct roles (e.g., Crapanzano et al., 2011; Troop-Gordon et al., 2019).

Considerable variation has been found among studies measuring the prevalence of bullying. While bullying is known to vary to some extent according to sample characteristics such as students' age, gender, or country (e.g., Currie et al., 2012), some of the variation is likely to be due to how bullying is operationalized and measured (Salmivalli & Peets, 2018). At the international level, the global prevalence of bullying and victimization (i.e., being a recipient of bullying) has been estimated to reach 35% each in a meta-analysis of 80 studies (Modecki et al., 2014).

Research has brought out compelling evidence of the adverse consequences of bullying. Victims face many adjustment difficulties and are at increased risk of depression and other mental health problems in adulthood (e.g., Lereya et al., 2015). To a lesser degree, witnessing bullying can also negatively impact bystanders (e.g., Nishina & Juvonen, 2005; Rivers et al., 2009). Bullies, in turn, are at greater risk of later maladjustment and externalizing problems (e.g., Lösel & Bender, 2011; Kretschmer et al., 2017).

2. The 'black box' of anti-bullying program evaluations

Given the high prevalence of bullying in schools and its detrimental consequences for the children involved, an increasing number of anti-bullying programs have been developed over the past three decades (Low et al., 2014). Some of these programs have shown positive effects on bullying and/or victimization (e.g., KiVa, Kärnä et al., 2011; ViSC, Yanagida, et al., 2016). However, meta-analyses and systematic reviews suggest that the results of anti-bullying programs have been mixed overall (Bradshaw, 2015; Gaffney et al., 2019). Although the most favorable meta-analyses suggest a 20% reduction in bullying on average, effect sizes vary considerably across studies (Farrington & Ttofi, 2009; Gaffney et al., 2019), and many programs have not yielded positive outcomes (Menesini & Salmivalli, 2017). While anti-bullying programs can be effective, they do not consistently translate into positive outcomes (Bradshaw, 2015; Cross & Barnes, 2014).

Below we develop three main reasons that might explain the limited and/or mixed results of anti-bullying programs. We argue that evaluations of anti-bullying programs comprise a 'black box', which limits our knowledge of how these programs work and may conceal their potential effectiveness. This 'black box' results from three gaps in the evaluations.

2.1 The hypothesized mediating paths are not tested (gap #1)

First, most evaluation studies do not measure the factors targeted by the interventions to ultimately impact bullying (or other related behaviors) (see Figure 1a). They generally assess the expected behavioral outcomes only, at the expense of the hypothesized mediators (Saarento et al., 2015). This impairs our knowledge and understanding of how programs work. As well as the effects of anti-bullying programs, we need to investigate whether and which of these effects flow through the expected mediators (Menesini & Salmivalli, 2017; Saarento et al., 2015). Assessing mediators not only enables researchers to test the hypotheses and theoretical frameworks behind anti-bullying programs accurately, but also reduces the

risk of wrongly concluding that the program or its components have no effect on bullying. An effect on bullying can easily be missed in intervention studies, for instance because the effect is too small to be detected at the distal level (i.e., on the behavioral outcome) or because a concurrent effect in the *opposite* direction (i.e., increasing bullying) is concealing the effect (Rucker et al., 2011). For example, an intervention enhancing cognitive empathy (the ability to understand another person's emotions) might in fact facilitate effective bullying by increasing the bullies' manipulative skills (Jolliffe & Farrington, 2011). The lack of consideration of the hypothesized mediators in the evaluation of programs might thus be a first reason for the limited and/or mixed results of anti-bullying programs.

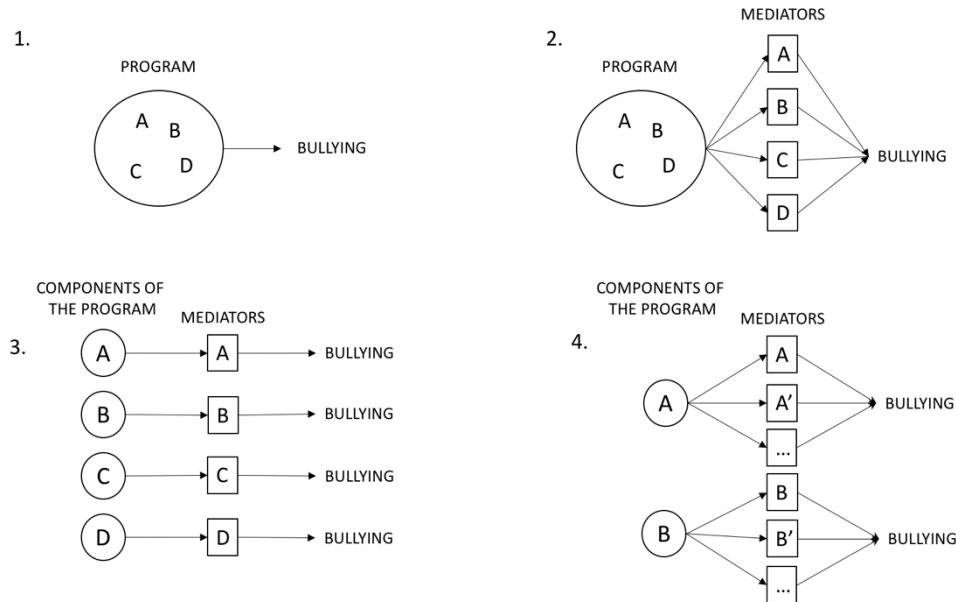
2.2 The different anti-bullying program components are not assessed within separate interventions (gap #2)

Anti-bullying programs are usually assessed as a whole. However, they generally contain many components targeting various factors associated with bullying. As a result, we are unaware of the unique contribution of each component. Regardless of the global outcomes of the program, we do not know what are the effective ingredients. We need to determine the unique effect of each component separately to improve our understanding of anti-bullying programs (Menesini & Salmivalli, 2017).

Assessing the hypothesized mediators of a program, in response to the first gap we pointed out, is sometimes also presented as a means to address this second issue (see Figure 1b). Let us consider a fictional example of a program that contains lessons on bullying to encourage students' anti-bullying attitudes and watching testimonial videos to enhance students' empathy toward victims. If we evaluate this program as a whole, assessing the two hypothesized mediating paths to bullying might be suggested as a way to determine the effectiveness of each component. This approach, however, is not ideal as it relies on the assumption that any change in a given mediator (e.g., students' empathy) can only be caused by the component that has been hypothesized to influence this specific mediator (in this example, watching testimonial videos). Yet change in a given mediator could be unexpectedly related to another component of the program (e.g., the lessons on bullying) or to the joint action of two or more components (e.g., watching testimonial videos in combination with theoretical lessons on bullying).

Although assessing the mediators may give researchers some insights about how anti-bullying programs work, there has been a call for further innovative research rather than evaluating whole packages of interventions. In order to better understand programs' causal mechanisms, researchers have been encouraged to assess promising anti-bullying components within separate interventions, after identifying them based on prior research and theory (Eisner & Malti, 2012). In summary, to address these first two issues, evaluation studies should assess anti-bullying components within separate interventions and test their hypothesized mediating paths (see Figure 1c). Moreover, an ideal research design should also make it possible to test other relevant mediators that might provide alternative explanations

for the effect of a given component on bullying (see Figure 1d). Such a research design is used in this dissertation.



(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.

Figure 1

2.3 Implementation is not taken into account (gap #3)

In most evaluation studies, the outcomes represent the overall global effect of a given program (Kallestad & Olweus, 2003). On closer examination, the programs' effects are seen to fluctuate strongly between classrooms, and, to a lesser degree, between schools. These variations are likely to be related to natural fluctuations in implementation, that is how the programs are put into practice (Kärnä et al., 2011). This is especially true when programs are delivered by teachers, which is often the case in bullying prevention (Goncy et al., 2015). The third reason for the limited and/or mixed results of anti-bullying programs might thus be that teachers do not necessarily implement the programs as expected by the developers (Salmivalli et al., 2013). Despite the widely acknowledged influence of implementation on school-based prevention programs' outcomes (Durlak & Dupre, 2008), school bullying evaluation studies have rarely assessed implementation, or only at a very general level (Salmivalli et al., 2013). An increasing interest in the study of implementation has emerged, and school bullying researchers have called for rigorous assessment of implementation in anti-bullying programs (Goncy et al., 2015).

3. The selection of relevant anti-bullying components

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). Apart from the three aforementioned methodological gaps, the limited effectiveness of anti-bullying programs (i.e., a 20% reduction in bullying at best; Gaffney et al., 2019) also emphasizes the need to promote research-based innovation in program development (Eisner & Malti, 2012). Although testing the replicability of existing interventions is important, researchers also need to keep studying new directions to prevent bullying by testing interventions that target under-explored relevant predictors of bullying. To improve effective prevention, we need to develop interventions that reflect scientific advances (Eisner & Malti, 2012). In that respect, this research aimed to investigate some anti-bullying components that have been suggested as promising directions for prevention but have rarely if at all been targeted as such in programs yet.

There is a discrepancy between students' private attitudes toward bullying and their actual behaviors. Whereas most of them consider bullying immoral and disapprove of it (e.g., Gini et al., 2011), many bully others and few of them would stand up for the victims (Hymel et al., 2010; Salmivalli, 2001). Both the moral disengagement and class norms frameworks provide a rationale to explain this paradox and have been suggested as promising directions for bullying prevention (e.g., Almeida et al., 2009; Menesini et al., 2015).

Moral disengagement has almost never been targeted in anti-bullying programs so far (for an exception, see Wang & Goldberg, 2017). Likewise, few studies have directly manipulated the class norms through a norm-based intervention in the context of bullying (for exceptions, see Dillon & Lochman, 2019; Perkins et al., 2011). Several programs, however, refer more or less explicitly to the improvement of social norms as a way to prevent bullying or enhance defending in bystanders. For instance, the goal of KiVa (Kärnä et al., 2011) is to influence the peer group in such a way that more students will express their disapproval of bullying and stand up for the victims (p.62) – the rationale being, in part, that perceptions of what is normative (i.e., common or accepted) in the classroom may influence the public expression of private attitudes toward bullying (Saarento et al., 2015). Nevertheless, in these programs, the norms are not directly manipulated as is the case in norm-based interventions (see below). Rather, program developers assume that norm change stems from improvement in other variables, such as students' awareness of bullying or empathy toward the victims.

The growing interest in morality and social norms among bullying researchers is also reflected in the current research trends in international conferences. As an illustration, the last edition of the World Anti-Bullying Forum³ opened with a keynote session entitled "Bullying through the lens of morality and peer dynamics", which articulated the influence of moral disengagement and social processes in bullying both theoretically and empirically (Salmivalli & Hymel, 2019). This thesis thus aims to take a first step in opening the black

³ The WABF is an international conference gathering researchers, policymakers, and practitioners in the fight against bullying (<https://worldantibullyingforum.com>).

box of anti-bullying programs, and to test two novel and promising anti-bullying components: intervening on moral disengagement and class norms.

4. Intervening on moral disengagement

The apparent discrepancy between students' private attitudes toward bullying and their actual behaviors can be analyzed through the lenses of Bandura's model of moral disengagement (Hymel & Bonnano, 2014). This framework is based on the idea that individuals can selectively disengage from the moral principles that normally guide their behavior. Children integrate these moral principles over the course of their development. Following these principles leads to self-satisfaction and a sense of self-worth, whereas transgressing them leads to self-sanctions such as guilt and shame. However, internal controls can be selectively disengaged through some cognitive processes, which are used to reinterpret the situation. Relying on these so-called moral disengagement mechanisms, one can behave in an immoral way without undergoing self-sanctions.

The eight moral disengagement mechanisms identified by Bandura are not restricted to aggressive and violent behavior. These mechanisms are used in everyday life and belong to normal functioning (Bandura, 1990). We all routinely face conflicting situations in which personal benefits are obtained or secured by transgressing our moral standards. As soon as external inducements overpower the strength of our internal controls, we are likely to adopt self-serving behaviors that have detrimental effects for others (Bandura, 1990). However, behaving in accordance with our internalized moral standards contributes to the achievement of self-determination and a sense of continuity and purpose in our lives. Using moral disengagement mechanisms to deactivate self-sanctions is therefore necessary to preserve our self-esteem (Bandura, 1991).

The theoretical framework of moral disengagement is not without similarities with the cognitive dissonance theory of Festinger (Hymel et al., 2010), which states that inconsistent cognitions yield psychological discomfort. This discomfort, in turn, prompts individuals to reduce the dissonance, notably by redefining the cognitions themselves – downplaying dissonant cognitions and reinforcing or creating consonant cognitions (Festinger, 1957). Cognitive dissonance, however, does not relate to moral issues or involve internalized self-sanctions, unlike moral disengagement. Cognitive dissonance refers to any discrepancy between two cognitions, which encompasses events, behavior, values, beliefs, feelings, information and knowledge about the environment, oneself, or one's behavior (e.g., smoking while knowing that it is bad for one's health) (Festinger, 1957).

Moral disengagement is also different from an emotion-regulation strategy, such as cognitive reappraisal (i.e., reframing an emotional situation) (Gross, 1998). Although moral disengagement also implies some cognitive restructuring, it is used to transgress one's moral principles while escaping self-condemnation. Cognitive reappraisal involves reframing one's thoughts about a stimulus to change its emotional impact, and can be considered a type of emotion-focused coping (Troy et al., 2013, p. 2505). Moreover, whereas cognitive reappraisal is generally seen as adaptive, and is associated with greater psychological health

(Troy et al., 2013), it has been suggested that moral disengagement leads to reduced well-being and poorer general health through less prosocial behavior (Guo et al., 2021).

4.1 The moral disengagement mechanisms

The eight moral disengagement mechanisms are categorized into four domains of moral disengagement in Bandura's conceptual model (Bandura, 1990). The cognitive restructuring domain includes the mechanisms of moral justification, euphemistic labeling, and advantageous comparison, which redefine the moral nature of the behavior itself. Through *moral justification*, one can make a reprehensible conduct seem righteous by presenting it as serving moral purposes. In other words, a moral end is held to justify immoral means. Such justifications have been used in many instances throughout history, justifying the perpetration of inhumane and destructive conduct in the name of religious, ideological, or nationalistic imperatives (Bandura, 1990). In the specific context of bullying, students can, for instance, claim that they bullied a classmate in order to help or protect their friends.

Language shapes the way we think and perceive reality. *Euphemistic labeling* is thus an easy and convenient mechanism to downplay the seriousness of detrimental behaviors or even make them appear honorable. The power of language is, of course, well known and widely used in various domains such as marketing, politics, or the military (e.g., “collateral damage”) (Bandura, 1990). Students who bully may, for example, pretend that “it’s just a joke” or that “they are just kidding” (Hymel et al., 2010).

Immoral behaviors can also be downplayed by contrasting them with worse actions through *advantageous comparison*. Because of the flagrant severity of what it is compared to, the behavior concerned appears of little consequence, or even trivial (Bandura et al., 1996). Bullies may, for example, argue that making fun of classmates is not as serious as beating them or compare their actions to the worse actions of others (Hymel et al., 2010).

The second domain of minimizing one's agentic role operates through the mechanisms of responsibility displacement and responsibility diffusion. Moral disengagement is easier when personal responsibility is made ambiguous. *Displacement of responsibility* occurs when people consider their behavior as determined by external circumstances, such as the dictates of a legitimate authority. Nazi camp commanders, for instance, declined any personal responsibility for their atrocious actions, arguing that they had simply obeyed orders (Bandura, 1990). Students may claim, for example, that bullying would not happen if school staff prevented it from happening (Hymel et al., 2010), or that other students initiated the bullying.

Through *diffusion of responsibility*, people disseminate their personal responsibility to the collectivity. Where everyone is responsible, no one is really responsible (Bandura, 1990, p.37). This may occur through the fragmentation of a process into isolated individual contributions, like in a bomb production line, for instance. In school bullying, it is not unusual to hear justifications such as “everybody does it anyway”.

In the third domain of distorting the consequences, one can *disregard or distort the consequences*. When individuals commit reprehensible actions to gain or secure personal benefits, they minimize or avoid facing the detrimental consequences. As long as the harmful effects of their actions are ignored or distorted, self-sanctions are not likely to be activated. This is why aggression is easier when aggressors do not see the harm they inflict, as with the use of mass destruction weapons (Bandura, 1990). Bullies can deny the harmfulness of their actions for the victims (“being teased does not hurt”) or even pretend that their behavior has positive consequences, for instance by pretending that the victim is pleased to get attention.

The last domain of victim attribution consists of dehumanization and blaming the victim. Empathy is activated when we perceive similarity with someone. It is easier to hurt people if we do not see them as human beings with feelings, hopes, and concerns. Our moral principles do not apply to subhuman objects, which explains why *dehumanization* is an essential ingredient in the perpetration of inhumanities such as slavery or genocides (Bandura, 1990, p. 38). In the context of bullying, the repeated use of nicknames and insults progressively dehumanizes the victim in both the bullies’ and bystanders’ eyes.

In the last mechanism, *blaming the victim*, victims are considered at least partly responsible for bringing suffering on themselves. Importantly, bystanders of immoral conduct are likely to infer from the situation that the victims probably deserved their plight. This mechanism is typical of sexual aggression toward women (e.g., rape victims are alleged to have provoked rape by their appearance or behavior) (Bandura, 1990). School bullies may argue that their actions were motivated by retaliation, in response to what they consider as provocation from the victim. Another example is to claim that the victims are responsible because they knew they were going to be bullied if they came near the bullies.

4.2 Empirical results and rationale for intervening on moral disengagement

Moral disengagement has been found to be associated with both bullying and bystanders’ behaviors (see Killer et al., 2019 for a recent meta-analysis), as well as cyberbullying (e.g., Hoareau et al., 2019; Wachs, 2012), in cross-sectional studies. It is positively related to bullying and negatively related to defender behaviors. Fewer results are available regarding the outsider role, but most recent findings have indicated a positive association with moral disengagement (e.g., Doramajian & Bukowski, 2015; Sjögren et al., 2020; Thornberg et al., 2017). As defenders’ behavior is in line with their moral standards, there is no need for them to rely on moral disengagement mechanisms. By contrast, outsiders are likely to rely on moral disengagement mechanisms to distance themselves from the victim’s distress and avoid self-sanctions (Doramajian & Bukowski, 2015).

Moral disengagement has been suggested as a promising avenue for prevention (e.g., Almeida et al., 2009; Hymel & Bonanno, 2014; Wang & Goldberg, 2017). If students can no longer use moral disengagement mechanisms, they will be unable to deactivate internal controls when transgressing their moral principles, whether as bullies or as bystanders. As a result, they will undergo the unpleasant self-sanctions that they were trying to escape, which

will in turn discourage their immoral behavior. To our knowledge, only one study has evaluated an intervention that explicitly targeted moral disengagement in students to help reduce school bullying, with mixed but promising results (Wang & Goldberg, 2017). After five weeks of intervention in third graders, significant time-by-treatment interactions were found for both moral disengagement and victimization (but not for prosocial behavior and socio-emotional skills). Partial eta squared (η^2) indicated moderate effect sizes (Wang & Goldberg, 2017). In addition, previous evaluations of juvenile delinquents and students on a prevention program targeting “self-serving cognitive distortions”, a concept close to moral disengagement (Ribeaud & Eisner, 2010), have shown positive effects on both recipients’ cognitions and behaviors (Gibbs, et al., 1996; van der Velden et al., 2010). For instance, a significant improvement in eight graders’ attitude toward antisocial behavior and self-serving cognitive distortions was found in the experimental group after 16 weeks of intervention (van der Velden et al., 2010). Both effect sizes were small.

4.3 Operationalization of moral disengagement

As moral disengagement has been suggested both as a relevant theoretical framework to understand bullying and a promising avenue for prevention (e.g., Almeida et al., 2009; Hymel & Bonanno, 2014; Obermann, 2011; Wang & Goldberg, 2017), a valid and reliable measurement tool is required to operationalize this concept. For this research, in particular, having a valid measure of moral disengagement was critical, as moral disengagement was the targeted mediator of our intervention and therefore at the heart of our research design. Bandura originally conceptualized moral disengagement as a multidimensional construct encompassing eight mechanisms, split into four domains. Three scales have been developed to measure moral disengagement, but none of them have fully matched the theoretical model so far. The first two failed to identify distinct dimensions and were validated as unidimensional structures (Bandura et al., 1996; Caprara et al., 1995; Hymel et al. 2005). The third had a seven-factor structure that fitted the theoretical subdivision into mechanisms, except for two of them, which formed a common factor (Thornberg & Jungert, 2014). This is to date the validated scale closest to Bandura’s conceptual subdivision into mechanisms. However, because the domains and mechanisms of moral disengagement are conceptually distinct but supposed to operate together (Thornberg & Jungert, 2014), an even better suited factorial structure should include the first-order factors corresponding to the mechanisms, and a second-order factor corresponding to the overarching dimension of moral disengagement. This discrepancy between the conceptualization of moral disengagement and the factorial structure of the scales calls into question either the theoretical model itself or its operationalization, and needs to be clarified.

5. Intervening on class norms

The second explanatory framework for the apparent paradox between students’ private attitudes toward bullying and their actual behavior concerns social norms, referring to what is typical or desirable in a group or situation (Prentice & Miller, 1996). Norms shape, constrain, and maintain behaviors of group members over and above individual attitudes

(Pozzoli et al., 2012; Veenstra et al., 2018). However, what individuals perceive to be the norm in a context is 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 regarding bullying (Perkins et al., 2011; Sandstrom & Bartini, 2010). As a result, most of them privately disapprove of bullying and pro-bullying attitudes but believe that others approve of it. This biased norm, in turn, may enhance bullying and restrain defending behavior, as students mistakenly feel support or social pressure to do so (Perkins et al., 2011; Sandstrom et al., 2013).

The main explanation for this overestimation bias is that negative cues capture our attention more than positive cues in general (Lansu & Troop-Gordon, 2017). Why negative valence elicits more brain activity than positive valence might be related to evolutionary strategies, as threat-related stimuli are more important than rewarding ones to survive and therefore need to be detected more quickly (e.g., Öhman et al., 2001). Moreover, negative behaviors such as bullying are often endorsed by popular peers (e.g., Vaillancourt et al., 2003), and high status attracts attention, even more in children and adolescents than in adults (Koski et al., 2015). From a resource control theory perspective, it may be relevant to pay more attention to high-status peers because they control physical and social resources (Hawley, 1999). It may also be that they represent a potential threat because of their greater propensity for relational and overt aggression (Sandstrom & Cillessen, 2006). In sum, this salience of antisocial behavior, possibly emphasized by high-status peers' behavior, may lead individuals to perceive a norm that is not accurate.

Pluralistic ignorance arises when most members of a group privately reject a norm but publicly behave in accordance with it because they believe that most others approve of it (Perkins et al., 2011; Sandstrom et al., 2013). With regard to bullying, students fail to recognize that the need they feel to align with the perceived norm may also be driving their classmates' behaviors (Dillon & Lochman, 2019). This phenomenon has some similarities with the fundamental attribution error, referring to our natural inclination to explain others' behavior by dispositional traits rather than situational factors (Andrews, 2001). Because conforming to the norm is usually rewarded with social approval, inclusion, and status, whereas deviation from it puts students at risk of social disapproval, exclusion, rejection, and victimization, students tend to conform to this biased norm despite the resulting conflict with their own values (Dillon & Lochman, 2019; Veenstra & Lodder, 2022). As they thus actively ignore their dissent and copy the negative behaviors that they assume to be the norm, students maintain and reinforce the collective misperception over time (Veenstra & Lodder, 2022).

In sum, whether students are driven by a desire to gain social status, belong to the group, or avoid social sanctions, they are likely to hide their disapproval or even show support for bullying and refrain from defending. And this is even more true when antisocial behaviors and attitudes are endorsed by high-status peers, who are known to be important norm-setters in the class (Dijkstra et al., 2008; Veenstra & Lodder, 2022).

5.1 Empirical results and rationale for intervening on class norms

A discrepancy between perceived and actual group norms has been found in a variety of health-related and social risky or problem behavior or attitudes (e.g., alcohol and tobacco consumption, drug use, unsafe sex, violence against women, sexual aggression) (Perkins et al., 2011; Sandstrom & Bartini, 2010). A similar discrepancy regarding peer support for aggression was found to persist from 6th to 8th grade (Henry et al., 2013). With regard to bullying, several studies have revealed that students consistently perceive their peers to hold more pro-bullying norms than they themselves do, both in elementary and middle schools (Dillon & Lochman, 2019; Perkins et al., 2011; Sandstrom & Bartini, 2010; Sandstrom et al., 2013). Moreover, several studies have shown that both assisting or reinforcing the bully and defending the victim were associated with peers' perceived attitudes rather than private attitudes toward bullying (Pozzoli & Gini, 2010; Rigby & Johnson, 2006; Sandstrom & Bartini, 2010; Sandstrom et al., 2013).

Interventions targeting the norms aim to rectify people's misperceptions of their peers' attitudes or behaviors (e.g., Dillon & Lochman, 2019). Social norms theory predicts that providing group members with correct information can change their perception of the norm, making a behavior or attitude more or less desirable. This change of perception, in turn, is likely to induce a change of behavior (Miller & Prentice, 2016). Norm-based interventions can vary in the way they convey information about the correct norm (Miller & Prentice, 2016). In social norms marketing (SNM), correct information about the high prevalence of some desirable behavior or attitude is disseminated to the whole group (e.g., Burchell et al., 2013). Personalized normative feedback (PNF) provides each person with information about the gap between their individual perception of the norm and the actual norm (e.g., Dillon & Lochman, 2019). The focus group discussion (FGD) involves facilitators leading interactive group discussions about the misperception between the perceived and actual norm (e.g., Reilly & Wood, 2008). Regardless of their format, all norm-based interventions are based on the simple observation that people do not usually have access to accurate information about what their peers think, feel, or do (Miller & Prentice, 2016). Norm-based interventions of all formats have been successfully used in diverse contexts (e.g., alcohol use, gambling, sunscreen use), leading to a clear reduction in norm misperception and intended or actual changes of behavior in the expected direction (Dillon & Lochman, 2019). To date, however, only a few studies have applied this approach in the context of school bullying (Dillon & Lochman, 2019; Perkins et al., 2011). In essence, the majority of students who privately disapprove of bullying realize that they are not alone in the opposition, and that most of their peers do not expect them to support bullying, thus reducing the feeling of social pressure to do so (Perkins et al., 2011). Moreover, this sets a power balance less in favor of the bullies, which is likely to make students more inclined to defend the victims or confront the bullies (Garandeau et al., 2019; Veenstra & Lodder, 2022). For the minority of students who actually hold pro-bullying attitudes, social norms theory assumes that change will operate through cognitive dissonance (Berkowitz, 2004). By revealing to these students the discrepancy between their own and their classmates' attitudes, the intervention challenges their feeling of

fitting in with their peer group, triggering some adjustment in the direction of the true norm to reduce the dissonance (Dillon & Lochman, 2019). In sum, correcting students' misperception of their classmates' attitude toward bullying should discourage bullies' behavior and increase bystanders' willingness to intervene.

5.2 Operationalization of class norms

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). While descriptive norms are measured through the mean level of a given behavior (e.g., bullying), injunctive norms are typically operationalized as the mean level of private attitudes among classmates (Veenstra & Lodder, 2022).

When it comes to pluralistic ignorance and correcting misperception, however, we also need to distinguish between *actual* and *perceived* norms (Perkins, 2014; Sandstrom & Bartini, 2010). A typical way of doing so for injunctive norms is to ask students (1) to what extent *they* (dis)agree, and (2) to what extent they think that *their classmates* (dis)agree with each of a set of statements about bullying (or a related variable) attitudes (e.g., Dillon & Lochman, 2019; Prentice & Miller, 1993; Shelton & Richeson, 2005). Such a procedure reveals (1) students' private attitudes and (2) their perception of classmates' attitudes. When averaged at the class level, these two scores represent (1) the mean level of private attitudes within the class (injunctive class norm) and (2) the mean level of the perceived attitudinal class norm (perceived injunctive class norm).

The perceived descriptive norm corresponds to the perceived prevalence of a behavior (Miller & Prentice, 2016), but has rarely been used in the context of bullying (for exceptions, see Kubiszewski et al., 2019; Saarento et al., 2015), with most class norm studies assessing actual rather than perceived descriptive norms, that is, the average prevalence of a given behavior within the class rather than students' perception of its prevalence among their classmates. Table 1 recapitulates the differences between actual, perceived, descriptive, and injunctive norms.

Table 1. Operationalization of actual and perceived descriptive and injunctive class norms

	Descriptive norms	Injunctive norms
Actual	Students' individual reports of a given behavior averaged at the class level.	Students' individual attitudes toward a given behavior averaged at the class level.
Perceived	Students' individual perceptions of the prevalence of a given behavior among their classmates.	Students' individual perceptions of their classmates' attitudes toward a given behavior.

In the case of pluralistic ignorance, however, it is more appropriate for interventions to target the biased perceived attitudinal (injunctive) norm (see bottom right cell of Table 1) rather than the biased perceived behavioral (descriptive) norm (Miller & Prentice, 2016;

Prentice, 2008). This is because pluralistic ignorance arises when most members of a group privately disapprove of a behavior while being convinced that most other members have positive attitudes toward it and therefore publicly behave in accordance with this misperceived attitudinal norm (Perkins et al., 2011; Sandstrom et al., 2013). As peers' private attitudes cannot be easily accessed by students and are almost never uniform, the perception that everybody supports a given behavior is the best target for intervention (Prentice, 2008). Finally, to be effective, norm-based interventions need not only to convey credible information (Miller & Prentice, 2016) but also 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).

6. The relevant implementation features in anti-bullying program evaluation

Identifying and selecting relevant anti-bullying components is necessary. However, the proper implementation of an intervention (or essential parts of it) is a precondition for any effects to occur (Kallestad & Olweus, 2003; Salmivalli et al., 2013). Evaluation of implementation is thus crucial to interpreting the effects of anti-bullying programs accurately (Low et al., 2014). The concept of implementation encompasses different research areas, such as the many parameters and factors that support or impede subsequent implementation, the issue of program sustainability over time, and the delivery process of the program content to the recipients (Cross & Barnes, 2014; Durlak & Dupre, 2008). This latter aspect is critical to interpreting the effects of anti-bullying interventions accurately (Low et al., 2014). In this thesis, we thus refer to the delivery process of the intervention content from the teachers to their students as "implementation". Several literature reviews have highlighted that the greater the implementation of school-based prevention programs, the more likely the expected outcomes arise, in domains such as drug abuse (Dusenbury et al., 2003), mental health (Weare & Nind, 2011), and aggressive behavior (Wilson & Lipsey, 2007). In the specific context of bullying prevention, however, studies examining the association between implementation and outcomes have been scarce (Haataja et al., 2014). This lack of knowledge limits our ability to support effective practice (Hirschtien et al., 2007). More studies investigating this issue are therefore needed (Haataja et al., 2014).

Several relevant implementation features have been suggested by different theoretical frameworks from the prevention literature (e.g., Berkel et al., 2011; Domitrovich et al., 2012; Durlak & Dupre, 2008; Fixsen et al., 2005). Which are decisive in influencing outcomes of anti-bullying interventions, however, has not been determined yet. Besides, there is a wide variability in the way implementation features have been defined, operationalized, and assessed. Despite these variations, the different concepts assessed throughout the studies most often fall into one of the following four implementation features: *fidelity* (the extent to which the delivered intervention corresponds to the prescribed content and procedure), *dosage* (how much of the prescribed program has been delivered), *quality* (how well the program has been delivered), and *responsiveness* (recipients' attitudes toward the program, interest, and participation) (Durlak & Dupre, 2008).

Fidelity has been assessed in most studies. Dose-outcome associations are relevant to examine, provided that there is enough variability in dosage – which is not always the case in small-scale evaluation trials (Haataja et al., 2014). Because of the key role of the teachers in influencing students' behavior and peer relationships through both explicit and implicit messages (e.g., Bierman, 2011; Farmer et al., 2011), quality has been suggested as particularly important (Berkel et al., 2011; Durlak, 2010; Goncy et al., 2015). Besides, quality has close links with the last feature – student responsiveness, which is sometimes directly included within the measure of quality as reflecting the teachers' ability to get students interested and involved in the lessons (e.g., Hirschstein et al., 2007). Most implementation research – including in the domain of bullying prevention – has long focused on a single feature of implementation (Durlak & Dupre, 2008; Low et al., 2014). Because different features of implementation are likely to interact and to jointly influence students' outcomes, there has been a call for empirical studies investigating combined dimensions to explain variations in anti-bullying interventions' effectiveness (Durlak & Dupre, 2008; Hirschstein et al., 2007; Low et al. 2014; Schultes et al., 2014). Notably, recommendations from the implementation literature suggest assessing both fidelity and quality (e.g., Berkel et al., 2011; Goncy et al., 2015).

Fidelity is usually operationalized as the proportion or sum of activities or core components delivered as prescribed by the program (Haataja et al., 2014). Thus specific items need to be created for assessing a given program's or lesson's fidelity (Leadbeater et al., 2018). Quality includes not only technical (e.g., instructional clarity) but also relational dimensions (e.g., classroom management, emotional tone) (Goncy et al., 2015, Haataja et al., 2014), and sometimes student responsiveness (Hirschstein et al., 2007). Among the different assessment methods, direct observation is considered by many to be the gold standard (Biggs et al., 2018), as it is held to be more reliable (Durlak & Dupre, 2008; Hirschstein et al., 2007). However, observer ratings are highly resource- and time-consuming, and can be resisted by teachers (Dusenbury et al., 2004). Self-reports (e.g., Haataja et al., 2014), electronic logs (e.g., Sharma et al., 2020), and interviews (e.g., Matischek-Jauk et al., 2018) have also been commonly used to assess anti-bullying program implementation. However, teacher self-reports can be incomplete or subject to social desirability biases (Dusenbury et al., 2004).

A long-lasting debate concerns the role of adaptation by the providers. Although adaptation was initially seen as a threat to fidelity, both quantitative and qualitative studies have suggested that some adaptation leads to better implementation and program outcomes (Durlak & Dupre, 2008). In this perspective, adjustments made by teachers to best fit the local context and their students' needs can be seen as contributing to the quality (Cross & Barnes, 2014; Goncy et al., 2015). Most researchers nowadays suggest that a balanced assessment of fidelity requires the identification of the core components that should not be modified for the program to remain effective (Durlak & Dupre, 2008).

Because researchers have called for a systematic assessment of implementation in anti-bullying programs (Goncy et al., 2015), more evaluation studies report an index of implementation nowadays. However, this index is usually reported to demonstrate that the intervention under examination was correctly implemented, or to explain mixed or non-

significant outcomes (e.g., Axford et al., 2020; Clarkson et al., 2019; Gage et al., 2019; Nickerson et al., 2019). The association between implementation and outcomes has only been examined in a minority of studies so far (Haataja et al., 2014).

7. Research questions and design

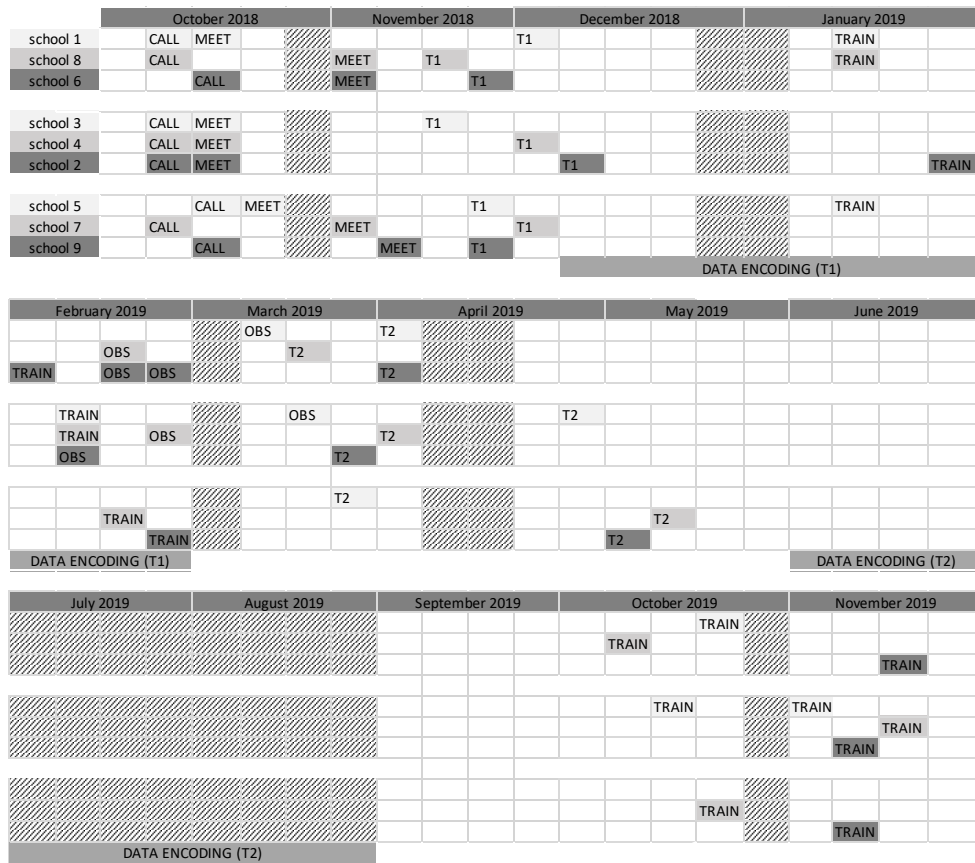
Because most evaluation studies omit to test the hypothesized mediators, assess anti-bullying programs' components as a whole, and do not take their implementation into account, our understanding of how these programs work remains limited. In addition, there is a need to study new directions in the prevention of bullying by testing promising yet under-explored intervention components. In the research for this dissertation, we aimed to determine the effectiveness of intervening on moral disengagement and class norms as anti-bullying components, as well as to investigate the influence of implementation on the effects of these two components. We specifically used a research design that assessed the components within separate interventions and tested their hypothesized mediating paths as well as an alternative relevant mediator (see Figure 1d).

By delivering a class-wide intervention on moral disengagement, we might have an unintended impact on students' perception of the class norms. For instance, during the reading of stories displaying bullying situations, students might detect subtle reactions of their classmates reflecting disapproval of the bully characters' actions. They might infer that their classmates disapproved of bullying more than they thought, thus modifying their perception of the attitudinal class norm toward bullying. Conversely, the class norm intervention might incidentally reduce students' use of moral disengagement mechanisms. By reinforcing the unacceptability of bullying through the class norm intervention, we might deter students from deactivating self-sanctions through moral disengagement mechanisms and/or weaken the power of these mechanisms as guilt-suppressors. Although we designed our interventions so as to target either moral disengagement or the class norm, we could not rule out the possibility that they might incidentally elicit the other variable as a mediator of the intervention. For that reason, and consistently with our methodological recommendations (see Figure 1d), both moral disengagement and perceived injunctive class norm were also tested as a possible mediator of the other intervention in our research design.

Our understanding of how anti-bullying programs work is intimately related to our knowledge of the psychosocial processes involved in bullying. Information about the former provides valuable insights about the latter, and vice-versa. Intervention studies offer an ideal research design for both research goals (Eisner & Malti, 2012). Our research design, in particular, provided direct information about the involvement of moral disengagement and class norms in bullying, as each of the two interventions was designed to target and manipulate only one of these mediators. Our 'opening of the black box of anti-bullying programs' is thus a means of expanding our knowledge and understanding of not only anti-bullying programs, but also bullying itself.

A three-armed randomized controlled trial (RCT) with a moral disengagement intervention condition, a class norm intervention condition, and an active control group was

conducted on a sample of about 1,200 elementary school students (Grades 4-6). The two interventions exclusively targeted either the moral disengagement component or the class norms component. They 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) respectively. We designed a whole new intervention of similar duration and intensity on a topic unrelated to bullying (climate and environmental issues) for the control group. The use of an active control group allowed us to control for changes stemming from non-specific treatment factors such as improvement of relationships between teachers due to the project collaboration, which can lead to spurious intervention effects (Lilienfeld et al., 2014).



Note. CALL = first phone contact with the school principal; MEET = first meeting with Grade 4-6 teachers; T1 = pre-test data collection; T2 = post-test data collection; TRAIN = training day; OBS = implementation assessment through direct observation. Cells with cross-hatching correspond to school vacation periods.

Figure 2. Calendar of recruitment, data collection, training, implementation assessment, and data encoding

After the recruitment drive was launched in September 2018, we contacted the schools that expressed an interest in the project by phone. Nine school principals confirmed their interest after we gave them a broad outline of the project. Grade 4-6 teachers in these schools were met between mid-October and mid-November during lunch times. We presented the research project (including the randomization process details) and asked them to indicate whether they were willing to join in a short confidential form. All teachers agreed to participate in the project ($N = 57$). Data were collected in the fall of 2018 and the spring of 2019, with an average of 15 weeks between the two waves. Each group of teachers attended a training day delivered between mid-January and the end of February 2019. The five intervention sessions were then delivered by the class teachers to their students for one hour a week. Implementation data (fidelity and quality) were collected by two independent trained raters who attended in parallel one specific lesson in each intervention classroom. The next school year, each school received a day of anti-bullying training for the whole teaching team for free between October and November 2019. Figure 2 shows an overview of the recruitment process, data collection waves, trainings, implementation assessment through direct observation, and encoding of the paper questionnaires.

8. Description of the chapters

Following this general introduction, the second chapter of this thesis corresponds to a first paper entitled “Validation of the French Version of the Moral Disengagement in Bullying Scale: Testing Bandura’s Conceptual Model”. None of the existing scales of moral disengagement have fully matched Bandura’s model so far, which calls into question the accuracy of either the theoretical model itself or its operationalization. As moral disengagement is the target of one of our interventions and is hypothesized to mediate change in student outcomes, using a valid measurement tool is even more crucial in this dissertation. In this paper, we attempted to narrow the gap between the moral disengagement conceptual model and its measurement tools by testing five factorial structures based on both theory and previous empirical work. We aimed to validate a second-order factor structure, which is congruent with the conceptual model. This first paper is thus a precondition for us to validly examine either the mechanisms or the global dimension of moral disengagement, using the French Version of the Moral Disengagement in Bullying scale (Fr-MDBS).

In the third chapter, we present a paper entitled “The Effectiveness of Moral Disengagement and Class Norms as Anti-Bullying Components: A Randomized Controlled Trial”. In this paper, we tested the effects of intervening on moral disengagement and class norms within separate interventions to examine the unique effects of these two anti-bullying components. We used mediation analysis to verify that the effects flowed through the expected mediators, that is, students’ moral disengagement on the one hand, and students’ perception of their classmates’ anti-bullying attitudes (the injunctive class norm) on the other hand. By doing so, we addressed the first two methodological gaps in evaluation studies that we previously highlighted. The paper examined the effects of the two anti-bullying components on bullying, defender, and outsider behaviors. In addition, sensitivity analyses were conducted to examine the effects on direct, indirect, and cyber bullying specifically.

The findings of this paper are critical to determine whether the use of moral disengagement and class norm components should be supported in anti-bullying programs. Details of power calculation, the recruitment process and eligibility criteria, the demographic characteristics of our sample, data collection, teacher training, and the content of the two interventions are provided in this chapter.

The fourth chapter presents a paper that addressed the last methodological gap that we previously emphasized, and examined how the implementation influenced the two aforementioned interventions' effects. We first presented a literature review of 22 evaluation studies of anti-bullying programs that examined the association between implementation and outcomes. We detailed which implementation features were assessed and how they were operationalized, as well as how their impact on the program outcomes was examined throughout the studies. Then, we examined how the combination of fidelity and quality of the implementation interacted with the effects of our interventions on bullying within a moderated mediation model. The results of this paper provided insights into the influence of fidelity and quality on the outcomes of these two anti-bullying components.

The fifth and last chapter is the general discussion of this thesis. All the material used in the interventions is provided in the Appendix.

Chapter 2

Validation of the French Version of the Moral Disengagement in Bullying Scale

Testing Bandura's Conceptual Model

Although Bandura conceptualized moral disengagement as a global dimension encompassing a set of mechanisms, the validated factorial structures of the existing scales have not fully matched the conceptual model so far. The present study attempts to narrow the gap between the moral disengagement conceptual model and its measurement tools by testing five factorial structures based both on theory and previous empirical work. A sample of 1,107 pupils from grades 4 to 6 completed the French version of the Moral Disengagement in Bullying Scale (Fr-MDBS). Confirmatory Factorial Analyses revealed that a second-order factor structure, congruent with the conceptual model, was better than a single-factor and as good as a first-order multi-factorial structure. Results also supported both construct validity (gender comparison) and predictive validity of the scale. Moral disengagement mechanisms were associated with self- and peer-reported bullying and victimization, defender, and outsider behaviors. Euphemistic labeling, diffusion of responsibility, and distorting consequences were also related to bullying while moral justification and victim attribution were related to both bullying and victimization. Euphemistic labeling, diffusion of responsibility, distorting consequences, and victim attribution were negatively associated with defender behavior and positively related to outsider behavior. Advantageous comparison was not related to any of the examined variables. The Fr-MDBS seems to be a useful assessment tool to validly examine either the mechanisms or the global dimension of moral disengagement..

Reference

Tolmatcheff, C., Galand, B., & Roskam, I. (2021). Validation of the French version of the moral disengagement in bullying scale: Testing bandura's conceptual model. *Journal of School Psychology* (accepted).

Validation of the French Version of the Moral Disengagement in Bullying Scale

Testing Bandura's Conceptual Model

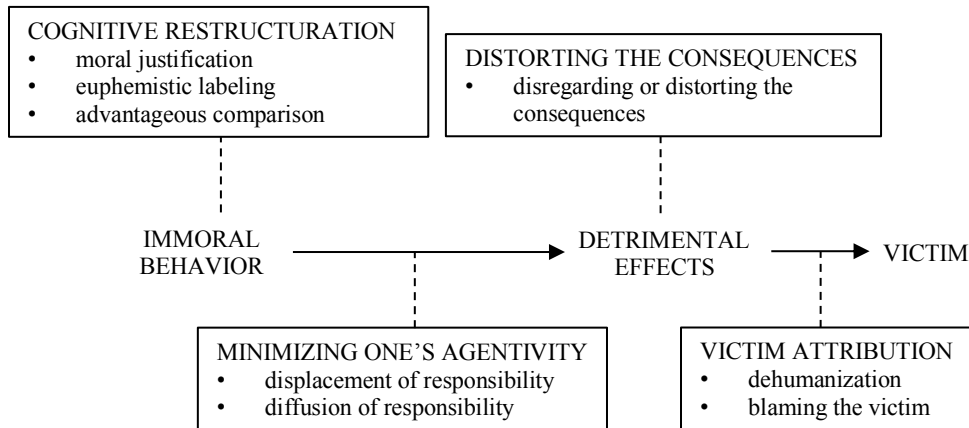
1. Introduction

How can we explain that otherwise kind and thoughtful children and adolescents sometimes behave in despicable and hurtful ways toward their peers and yet do not feel guilty about it? How can we explain that most students who witness these aggressions do not intervene to support the victims? These questions are particularly relevant to the context of school bullying, a situation of systematic, unprovoked, peer-to-peer abuse in a setting of power imbalance (Gini et al., 2011). Bullying is a prevalent and pervasive problem in schools worldwide, with potential negative consequences that may persist into adulthood, such as depression, anxiety, or suicidality (e.g., Wolke & Lereya, 2015). Whereas most school-age children and adolescents report to disapprove bullying and to be aware of its immoral nature (e.g., Gini et al., 2011; Salmivalli, 2001), a substantial portion of them do bully others, and reactions of defense or support toward victims from bystanders are rare (Hymel et al., 2010; Salmivalli, 2001). This discrepancy between students' attitudes and their actual behavior can interestingly be analyzed through the lenses of Bandura's model of moral disengagement (Hymel & Bonnano, 2014).

1.1 Mechanisms of Moral Disengagement

In his social cognitive theory of thought and action, Bandura (1986) offers an explanatory framework based on the central role of selective disengagement from one's moral principles. When growing up, children integrate moral standards that guide their conduct. Behaving in accordance with these moral principles leads to self-satisfaction and a sense of self-worth whereas transgressing them leads to self-sanctions like guilt and shame. However, some cognitive processes can selectively disengage internal controls, allowing one to act in an immoral way without undergoing self-condemnation. Bandura identified eight of these moral disengagement mechanisms, split into four domains of moral disengagement. The full conceptual model is organized as follows (see Figure 1). The cognitive restructuring domain includes the mechanisms of *moral justification* (the immoral behavior serves a moral and worthy purpose), *euphemistic labeling* (the immoral behavior is negated through a sanitized and innocuous language), and *advantageous comparison* (the immoral behavior is made benign by comparing it with an even more harmful one). The second domain of minimizing one's agentive role operates through the mechanisms of *responsibility displacement* (responsibility belongs to someone/something else), and *responsibility diffusion* (responsibility is diffused among several people and no one bears it). In the third domain of distorting the consequences, one can minimize, deny, or even pretend that the

consequences of the detrimental behavior are positive (*disregarding or distorting the consequences*). The last domain of victim attribution consists of *dehumanization* (victims are divested of their human being quality), and *blaming the victim* (victims are blamed for the situation).



Note. This figure was adapted from “Mechanisms of moral disengagement in the exercise of moral agency,” by A. Bandura, C. Barbaranelli, G. V. Caprara, and C. Pastorelli, 1996 *Journal of Personality and Social Psychology*, 71(2), p. 365. (<https://doi.org/10.1037/0022-3514.71.2.364>) Copyright 1996 by the American Psychological Association.

Figure 1. Bandura’s Framework of Moral Disengagement

The use of these mechanisms of moral disengagement is not limited to exceptional contexts in which large-scale atrocities take place. Instead, these mechanisms are encountered in everyday life and belong to normal functioning (Bandura, 1986). Bandura’s conceptual model of moral disengagement has been thought to be especially relevant in the context of school bullying, as cross-sectional studies have highlighted associations between moral disengagement on the one hand, and bullying perpetration, victimization, as well as bystanders’ roles, on the other.

1.2 Moral Disengagement and Bullying

With regard to the perpetration of bullying (i.e., engaging in bullying behavior towards other students), a recent meta-analysis examining 49 studies found a mean correlation of $r = .31$ with moral disengagement (Killer et al., 2019). In addition, the relation between moral disengagement and bullying perpetration holds even when the influence of other factors like aggression efficacy (Barchia & Bussey, 2011) or parental monitoring and mother-child warmth/support (Pelton et al., 2004) are controlled.

Results on the association between moral disengagement and victimization (i.e., being a recipient of bullying) are inconsistent. While some studies report a negative relation (Pornari & Wood, 2010), others point out a positive relation (Allison & Bussey, 2017), or no relation

(Pozzoli et al., 2016). On the one hand, victims might have an enhanced sensitivity to moral responsibility due to their experience of being victimized (Perren et al., 2012). On the other hand, as victims often blame themselves for their plight (Graham & Juvonen, 1998; Tolmatcheff et al., 2019), they might display higher levels of moral disengagement, notably at the victim level (i.e., *dehumanization* and *blaming the victim*) (Perren et al., 2012). A meta-analytic review of 21 studies found a significant, yet weak, positive association between moral disengagement and victimization, with a mean effect size of $r = .08$ (Killer et al., 2019).

It should be noted that whether bullying perpetration and victimization are self- or peer-reported has been shown to influence their link with moral disengagement (Gini et al., 2014; Killer et al., 2019). For instance, the association with moral disengagement was equal to $r = .34$ for self-reported bullying but equal to $r = .17$ for peer-reported bullying in a recent meta-analysis (Killer et al., 2019). As moral disengagement is usually self-reported, this might be the result of shared method variance (Killer et al., 2019). There has been a longstanding debate to determine which of self- and peer-reported data are the most suitable method to assess bullying perpetration and victimization. Recent recommendations support the use of both forms of reports to compensate for their respective biases as well as to capture a more accurate picture of bullying (e.g., Hymel & Swearer, 2015; Volk et al., 2017).

Bystanders' key role in school bullying by either enhancing or dissuading it has long been recognized (Jenkins & Troop-Gordon, 2020; Salmivalli et al., 1996). A negative association was found between moral disengagement and the defender role (i.e., bystanders providing support to the victim or even standing up to the bully), with a mean correlation of $r = -.17$ computed from 23 studies (Killer et al., 2019). As defenders' behavior is in line with their moral standards, there is no use for them to rely on moral disengagement mechanisms (Doramajian & Bukowski, 2015). For outsiders (i.e., bystanders not directly involved in bullying who passively observe it or walk away), the few existing results are mixed. Some studies reported a negative (e.g., Gini, 2006; Thornberg & Jungert, 2013) or null association (e.g., Gini et al., 2015). Most of recent findings, however, indicated a positive association (e.g., Doramajian & Bukowski, 2015; Sjögren et al., 2020; Thornberg et al., 2017), which is consistent with the theoretical expectations. As they refrain from helping, outsiders are likely to rely on moral disengagement mechanisms to distance themselves from the victim's distress and avoid self-sanctions (Doramajian & Bukowski, 2015). The assistant (i.e., bystander who joins in the bullying but does not initiate it) and reinforcer (i.e., bystander who displays encouraging behaviors, like laughing or cheering) roles have been traditionally considered in the literature as well (Salmivalli et al., 1996; Troop-Gordon et al., 2019). However, this categorization has met criticism because of the strong covariances between the bully, assistant, and reinforcer roles that question the conceptual distinction between them (e.g., Crapanzano, et al., 2011; Troop-Gordon et al., 2019). For this reason, we focused on defenders and outsiders in the present study.

1.3 Measurement of Moral Disengagement

Based on this set of results, moral disengagement has been suggested both as a relevant theoretical framework to understand bullying, and a promising avenue for prevention (e.g.,

Almeida et al., 2009; Hymel & Bonanno, 2014; Obermann, 2011; Wang & Goldberg, 2017). As such, a valid and reliable measurement tool is required to operationalize the concept of moral disengagement as well as to conduct intervention studies targeting moral disengagement. Although three scales inspired by Bandura's framework have been developed to measure moral disengagement, their validated factorial structures have not fully matched the theoretical model so far. The most commonly used scale was developed by Bandura and colleagues (Bandura et al., 1996; Caprara et al., 1995) in three versions, namely for elementary children (14 items), adolescents (24 items) and adults (32 items). Another 13-item scale was proposed by Hymel et al. (2005) to assess moral disengagement in the specific context of school bullying. Whereas Bandura originally presented moral disengagement as a construct encompassing four domains and eight mechanisms, these two scales failed to identify distinct dimensions and were validated as unidimensional structures, thus calling into question either the theoretical model itself or its operationalization. Subsequent studies using either scale repeatedly displayed a single-factor solution (e.g., Almeida et al., 2009; Bandura et al., 2001; Gini, 2006; Obermann, 2011; Paciello et al., 2008; Pelton et al., 2004). To our knowledge, only one study using the shortened version of Bandura's scale identified the four broader domains -but not the eight mechanisms- of moral disengagement in the context of school bullying (Pozzoli et al., 2012). In that respect, Thornberg and Jungert (2014) made a valuable step forward in developing and validating a third scale, entitled Moral Disengagement Bullying Scale (MDBS). They initially created 55 items designed to capture the eight mechanisms of moral disengagement, and tested it in a sample of 372 Swedish children aged 10 to 14 years old. Through a systematic and iterative process of repeated Explorative Factor Analysis (EFA), they found a seven-factor solution based on 18 items that fit the theoretical subdivision into mechanisms, except for *dehumanization* and *blaming the victim*, which formed a common factor. Although Thornberg and Jungert's scale thus encompasses only seven of the eight theoretical mechanisms, it is to date the validated scale closest to Bandura's conceptual subdivision into mechanisms.

At this point, it is not clear whether moral disengagement domains and mechanisms are all highly inter-related (unidimensional construct) or partially independent (multidimensional construct). Moreover, since the domains and mechanisms of moral disengagement are conceptually distinct but supposed to operate together (Thornberg & Jungert, 2014), an even better suited factorial structure should include first-order factors corresponding either to the four domains or to the eight mechanisms, and a second-order factor corresponding to moral disengagement. In a recent study using the MDBS (Bjärehed et al., 2020), such a factorial structure encompassing the seven first-order factors and a higher order factor presented a better fit than a single-factor structure. However, whether this higher order factorial structure also prevails over the original seven first-order factor structure still needs to be clarified. In addition, bullying is recognized as a serious problem not only in high schools but also in primary schools (Hong & Espelage, 2012), with a "peak" suggested to occur between ages 9 and 15 (Álvarez-García et al., 2015; Carney & Merrell, 2001). It is therefore of interest to have a measure of moral disengagement also validated in elementary school-age children.

1.4 The Current Study

Our main goal was to validate the French version of the Moral Disengagement in Bullying Scale (hereafter Fr-MDBS) (Thornberg & Jungert, 2014) in a sample of 8- to 12-year-old children. First, we tested the Fr-MDBS' construct validity by examining its factorial structure. We tested five different models relying on both Bandura's theoretical framework and previous empirical research. Second, since higher absolute levels of moral disengagement have consistently been found in boys compared to girls (Gini et al., 2014), higher levels of moral disengagement in boys would also support construct validity. We thereby tested measurement invariance across gender. When scalar invariance is reached, we can meaningfully compare boys and girls (Dimitrov, 2010). Third, we assessed the predictive validity of the scale by examining the associations between moral disengagement and bullying perpetration, victimization, and bystander roles (i.e., defender and outsider). Relying on results from previous studies, we hypothesized that moral disengagement would be positively associated with both bullying perpetration and victimization (Killer et al., 2019). As we followed the recommendations to use both self- and peer-reports to evaluate bullying perpetration and victimization, we expected moral disengagement to be more strongly linked to bullying and victimization when the latter were self-reported (Gini et al., 2014; Killer et al., 2019). With regard to bystanders, we expected a negative association between moral disengagement and the defender role but a positive association between moral disengagement and the outsider role (Doramajian & Bukowski, 2015; Sjögren et al., 2020; Thornberg et al., 2017).

2. Method

2.1 Participants

An initial sample of 1,216 fourth, fifth and sixth graders participated in this study ($M = 10.15$ years; $SD = 1.02$; 52.4% girls). They came from 57 classrooms of nine French-speaking primary schools. These schools were diversified in terms of size, socioeconomic level, and geographical location. They were recruited on a voluntary basis in the context of a larger intervention study on school bullying. Passive parental consent was obtained prior to data collection through paper consent forms sent by the schools. Sixty-four students (5.2%) were absent during the data collection stage, and the parents of 37 pupils (3%) did not want their child to participate. These pupils, as well as eight recently arrived immigrant pupils (0.6%) who had difficulties reading French, were dropped from the analyses. The final sample was comprised of 1,107 pupils.

2.2 Procedure

The Fr-MDBS was obtained with a double-check procedure in collaboration with a certified English teacher from the Institute for Modern Languages of our university. Both the first author of this paper and the English teacher translated the scale in parallel. Discrepancies between the two translations were then discussed to reach agreement. Minor modifications

were made to simplify and adapt the items to French. These adjustments partly relied on a pilot study conducted with 38 fourth graders from a school outside the sample and are detailed in supplementary material (see Table S.1). The goal of this pilot study was to check that the youngest pupils of the present study (i.e., the fourth graders) would understand the items correctly and easily. In a first step, the 38 pupils were asked to help the researchers by testing a questionnaire that was intended for some other pupils of their age. They were specifically requested to report any unclear or difficult words or items. In a second step, a group of five pupils volunteered to examine the whole questionnaire together with the researchers to discuss the listed issues in depth and find the best alternatives. The final decisions were endorsed by the first author of this paper and the English teacher to ensure the accuracy and meaning of the items. The French version of the items is provided in supplementary material (see Table S.1).

Afterwards, pupils in the sample of the present study were asked to fill in confidential pen and paper questionnaires during regular class hours under the supervision of two researchers – a PhD student (first author of this paper) and a Master's student. Pupils were informed that their participation was not mandatory and that their answers were confidential. In the fourth graders' classrooms, the questionnaire was read aloud by a researcher during the completion to compensate for potential reading difficulties at this age. The fifth and sixth graders read and completed the questionnaire by themselves after receiving general instructions. The procedure was approved by the ethic committee of the Psychological Sciences Research Institute of our university.

2.3 Measures

2.3.1 Moral Disengagement

The 18 items of the Fr-MDBS represent seven of the eight mechanisms theorized by Bandura and colleagues (1996), as dehumanization (one item) and blaming the victim (two items) had been merged into one factor (victim attribution) in the original validation study (Thornberg and Jungert, 2014). Pupils were asked to indicate to which degree they agreed or disagreed with each statement (from 0 = *totally disagree* to 4 = *totally agree*). Factorial structure and internal consistency of the Fr-MDBS are presented in the results section. Reliability indices for the global index and seven subscales, as well as CFA results from Thornberg and Jungert's original validation study (2014) can be found in Appendix A.

2.3.2 Bullying Perpetration

Self-Report

Eleven items listed verbal (teasing, insulting), physical (hitting/kicking/slapping, biting/scratching/pulling hair), social (excluding, spreading rumors), material (racketing, hiding/damaging personal things) and cyber (insulting/scaring through the Internet, calling/texting, releasing pictures/videos) forms of bullying behaviors. Pupils had to indicate the number of times (from 0 = *zero times* to 4 = *four times or more*) they had engaged in each behavior toward another pupil over the last three months (ordinal alpha $\alpha = .87$). This scale

has previously been used among elementary school students (Grades 4-6; Tolmatcheff et al., 2020) and has a high internal consistency ranging from $\alpha = .84$ to $\alpha = .86$ (e.g., Galand et al., 2014; Tolmatcheff et al., 2020).

Peer-Nomination

Following the self-reported measure, each participant received a list with the names of all the children in the class. The instruction was inspired by Salmivalli et al. (1996), but unlike the original procedure, a single item captured the bullying role by specifying that the previous list of items (from the self-reported measure of bullying) were examples of bullying behaviors. Pupils were asked to “tick an unlimited number of classmates who are used to bullying¹ others, for instance, by doing some of the things described above”. The number of nominations received by each pupil was divided by the total number of pupils in the classroom minus one (as students were not allowed to nominate themselves), resulting in scores ranging from 0 to 1. The use of a single item is not unusual in peer-nomination research and its reliability arises from the multiple reporters instead of multiple items (Cillessen & Marks, 2017).²

2.3.3 Victimization

Self-Report

Eleven items listed the same common forms of bullying behaviors as in the self-reported bullying scale described above (Galand et al., 2014; Tolmatcheff et al., 2020). This time, pupils had to indicate the number of times (from 0 = *zero times* to 4 = *four times or more*) they had undergone the unpleasant experiences over the past three months, e.g., “another pupil has hidden or damaged my things on purpose.” (ordinal alpha $\alpha = .89$). This scale has previously been used among elementary school students (Grades 4-6; Tolmatcheff et al., 2020), with a high internal consistency ranging from $\alpha = .83$ to $\alpha = .89$ (e.g., Galand et al., 2014; Tolmatcheff et al., 2020).

Peer-Nomination

Pupils were asked to tick an unlimited number of classmates who used to be bullied by others. This single item (inspired by Salmivalli et al., 1996) captured the victim role by specifying that the previous list of items (from the self-reported measure of victimization) were examples of ways of being victimized by others. The number of nominations received by each pupil was divided by the total number of pupils in the classroom minus one, resulting in scores ranging from 0 to 1.²

¹ We did not use the French verb commonly used for bullying (*‘harceler’*) but instead paraphrased it. The wording “[...] les élèves qui embêtent souvent les autres” can be literally translated by “[...] the students who frequently harass others”.

² For a recent example of the use of a single item for bully/victim peer-nomination, see Garandeau et al. (2019).

2.3.4 Defender Behavior

We used the subset of items selected by Ruggieri et al. (2013) from the shortened version of the participant roles scale adapted by Sutton and Smith (1999) and originally developed by Salmivalli (1996). The three items investigated pupils' potential defending behavior when witnessing a bullying episode: "I ask an adult or peer for help", "I comfort the victim", "I try to help the victim". Pupils 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 $\alpha = .67$.

2.3.5 Outsider Behavior

Pupils were asked about behaving as an outsider when witnessing a bullying episode. We adapted three items based on the aforementioned shortened version of the participant roles scale (Sutton & Smith, 1999) to be formulated as self-reported (instead of peer-reported) statements. For instance, the description "Pretends not to notice what is happening" was changed into "I pretend not to notice what is happening". The two other items were: "I don't look in that direction or I stay away", and "I don't do anything". Children were asked to indicate how often they behaved as described in each item (0 = *never*, 1 = *seldom*, 2 = *sometimes*, 3 = *usually*, 4 = *always*). Ordinal alpha was $\alpha = .70$.

2.4 Data Analyses

We used Confirmatory Factor Analyses (CFAs) to compare the fit of five different models. **Model 1:** First, we tested a single factor solution. **Model 2:** Next, we tested a four-factor model with the four domains of moral disengagement as latent factors, as found by Pozzoli et al. (2012). **Model 3:** Then, we tested a seven-factor model, corresponding to the mechanisms, as found by Thornberg and Jungert (2014). **Model 4:** Congruent with Bandura's theoretical model, we tested a second-order factor model by adding a higher order factor to the four-factor model (i.e., Model 2). **Model 5:** Finally, we tested a second-order factor model by adding a higher order factor to the seven-factor model (i.e., Model 3). All CFAs were performed with MPlus version 8.4 and took the nested structure of the data into account by using the TYPE=COMPLEX option of the ANALYSIS command in conjunction with the CLUSTER option of the VARIABLE command. In this approach, the Maximum Likelihood Robust (MLR) method of estimation is used by default when all observed dependent variables are continuous. In MLR, missing data are handled using full-information maximum likelihood (FIML). Data were missing on all observed variables in seven cases (0.6%), which were not included in the analyses. The covariance coverage of data was above 98% for all the indicators. The Chi-square (χ^2) test has long been used as a measure of fit, although it is highly sensitive to sample size and can thus lead to model rejection even when discrepancies from expected values are trivial (e.g., Dimitrov, 2010). Because of the unsatisfactory nature of the χ^2 statistic for model fit assessment, diverse alternative indices have been developed (Schermele-Engel et al., 2003). We reported the root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMR), the comparative fit index (CFI), and the Tucker-Lewis index (TLI) as descriptive goodness-of-fit indices, along with the χ^2 test, as recommended by Hu and Bentler (1999). In addition, we reported the Akaike

Information Criterion (AIC) and Bayesian Information Criterion (BIC), which can also be used to compare non-nested models (Schermerle-Engel et al., 2003). For RMSEA and SRMR, values are considered as acceptable when they are equal to or lower than .08, and good when they are equal to or lower than .05. For CFI and TLI, values should be equal or higher than .90 to be acceptable, and equal or higher than .95 to be good (Little, 2013).

AIC and BIC values cannot be interpreted in isolation. Within a set of models for the same data, the lowest values indicate the best fitting model (Schermerle-Engel et al., 2003). However, it is common for AIC and BIC to disagree as they handle the tradeoff between sensitivity and specificity differently. The BIC rewards model parsimony more than AIC, and is usually preferred in large samples because its probability of selecting the true model increases with N (Dziak et al., 2020). As the χ^2 difference test applied to nested models is subject to the same shortcomings as the χ^2 test applied to any single model (i.e., trivial differences may become significant even in modest samples), it is not of use for comparing nested models in the present study (Schermerle-Engel et al., 2003). It should be noted that since RMSEA, SRMR, CFI, and TLI are absolute (as opposed to relative) indices of fit (Dimitrov, 2010), their values should not be directly used to compare competing models to each other (i.e., comparing a CFI of, for example, .98 to a CFI of .96, which both indicate a good fit, is meaningless). If both competing nested models are well fitting according to the conventional thresholds listed above, the more restricted model should not be rejected, provided that there is theoretical justification supporting this model (Byrne, 2005). It should be pointed out that this applies to the comparison of competing models only. Different recommendations prevail in the context of measurement invariance and are described below.

We then measured the reliability of the items grouped according to the selected model. Following Gadermann et al.'s recommendations (2012) for estimating reliability coefficients with ordinal Likert-type scales, we used the polychoric correlation matrix to compute ordinal alphas³. Conventionally, the required threshold value is .70.

Regarding measurement invariance across gender, we used a stepwise multi-group CFA analysis (MGCFA). Again, the nested structure of the data was taken into account by using the TYPE=COMPLEX option of the ANALYSIS command in conjunction with the SUBPOPULATION option of the VARIABLE command. The recommended forward approach to test factorial invariance across groups relies upon a sequence of comparisons between pairs of nested models with gradually increasing parameters constraints (Dimitrov, 2010). In the first step, baseline models were determined for both groups, then combined into a configural CFA model to test the model simultaneously across groups without equality restriction on any model parameters other than the number of factors and the patterns of factor loadings. (When applied to models with a second order-factor, each of the following steps must be tested for both the first-order and the second-order factors parameters.) In the second step, we tested *metric invariance*, that is to say invariance of the factor loadings. In the third

³ It should be noted that the empirical use of ordinal alpha is still subject to debate in the literature. Some authors advise to limit its use to a hypothetical tool. See, for example, Chalmers, R. P. (2018). On Misconceptions and the Limited Usefulness of Ordinal Alpha. *Educational and Psychological Measurement*, 78(6), 1056-1071. <https://doi.org/10.1177/0013164417727036>

step, we tested *scalar invariance*, which requires the intercepts to be set as equal across groups (Dimitrov, 2010). As mentioned above, the χ^2 difference test is known to reject the null hypothesis in large samples even when violations are minor in a pair of nested models. Specific recommendations have been developed for comparison of adjacent models in the context of measurement invariance. When the sample is large enough (>300), a CFI difference bigger than .01 supplemented by a difference in the RMSEA bigger than .015 or a difference in SRMR bigger than .03 indicates metric non-invariance (Chen, 2007). The same applies to the scalar invariance, except that the difference in SRMR should be larger than .01 (Chen, 2007). To our knowledge, these recommendations are applicable in the context of measurement invariance only, and there is to date no equivalent benchmarks for the comparison of competing models.

Next, using the IBM SPSS Statistics version 25 program, we compared boys and girls as an indicator of construct validity (available in Appendix B, see Table B.1). Then we computed inter-correlations between the factors emerging from the CFAs and the other variables of the study (available in Appendix B, see Table B.2). Finally, we carried out hierarchical regression analyses (Enter method, as recommended by Field, 2018) on bullying perpetration, victimization, defender behavior, and outsider behavior to explore the predictive validity of the scale. Preliminary analyses included collinearity tests. Variance inflation factors (VIF) ranged from 1 to 1.88, confirming the absence of collinearity. In the first step of the regressions, we entered age and gender as control variables as both are well-known and frequently examined predictors of bullying perpetration, victimization, and bystander roles (Hong & Espelage, 2012). In the second step, we entered the factors of moral disengagement emerged from the CFAs. As bullying perpetration and victimization are known to be strongly correlated (for a recent meta-analysis, see Walters, 2020), the regression analyses for self-reported bullying and victimization were both run a second time with the other variable entered in the first step as control variable along with age and gender. By default, SPSS uses pairwise (e.g., in correlation analysis) or listwise (e.g., in regression analysis) exclusion of missing values. As a consequence, sample sizes indicated in the tables slightly vary from one analysis to the other.

3. Results

3.1 Confirmatory Factor Analyses

Table 1 displays goodness-of-fit indices of the five tested models.

Table 1. Goodness-of-fit indices of Confirmatory Analyses of the five models

Models	S-B χ^2	df	RMSEA	90% CI for RMSEA		CFI	TLI	SRMR	AIC	BIC	
				LL	UL						
Model 1 (F1)	701.4*	135	.062	.057	.066	.78	.75	.056	62513.9	62784.1	Poor fit
Model 2 (F4)	575.6*	129	.056	.051	.061	.83	.79	.050	62354.5	62654.7	Poor fit
Model 3 (F7)	198.9*	114	.026	.020	.032	.97	.96	.029	61907.7	62282.9	Good fit
Model 4 (F4+1)	609.7*	132	.057	.053	.062	.81	.78	.057	62392.4	62677.6	Poor fit
Model 5 (F7+1)	255.5*	128	.030	.025	.035	.95	.94	.033	61950.4	62255.6	Good fit

N = 1100

* $p < .05$

F1 = single factor model; F4 = four first-order factor model (see Pozzoli et al., 2012); F7 = seven first-order factor model (see Thornberg & Jungert, 2014); F4+1 = second-order factor model with four first-order factors; F7+1 = second-order factor model with seven first-order factors

Both Model 3 (the seven-factor model, corresponding to the mechanisms) and Model 5 (adding a higher order factor to Model 3) indicated a good fit of the observed data. The small decrease in the fit of Model 5 is not surprising given that this model is less flexible (more constrained) than Model 3. Both models presented a good fit according to the conventional thresholds, hence adding a higher order factor did not significantly worsen the fit of the model. We contend that the higher order factor structure best represents the data for the following reasons: (a) both models demonstrated a good fit; (b) there is a strong theoretical justification to consider moral disengagement as a higher order construct encompassing lower order subdimensions; (c) the higher order model is more parsimonious (as supported by the lowest BIC); and (d) correlations among the seven first-order factors – provided in Appendix B (see Table B.2) are satisfying (Byrne, 2005). Standardized factor loadings for this model are displayed in Appendix C (Fig. C.1). All the estimated factor loadings were significant at $p < .001$, and ranged from .42 to .80.

3.2 Reliability

Ordinal Cronbach's alphas for the subscales were .64 (moral justification), .75 (euphemistic labelling), .75 (advantageous comparison), .68 (displacement of responsibility), .60 (diffusion of responsibility), .70 (distorting consequences), and .63 (victim attribution). For the global score of moral disengagement, ordinal Cronbach's alpha was .82. For a comparison with the reliability indices of the original scale (Thornberg & Jungert, 2014), see Table S.1 in supplementary material.

3.3 Measurement Invariance across Gender

A stepwise multi-group CFA analysis (MGCFA) was performed to test the invariance of the second-order factor model between girls and boys. As can be seen in Table 2, goodness-

of-fit indices revealed an acceptable (CFI, TLI) to good (RMSEA, SRMR) fit for both genders when tested separately. The acceptable to good fit of the configural model indicated the same number and pattern of factor loadings across genders. CFI differences smaller than .01 provided support for first-order and second-order factor loadings invariance (M1 and M2, respectively), thus achieving metric invariance. CFI differences provided support for scalar invariance as well, since setting first-order and second-order factor intercepts equal across groups (M3 and M4, respectively) did not worsen the model fit ($\Delta\text{CFI} > -.01$). This allowed us to meaningfully compare the means of the underlying factors across genders (Dimitrov, 2010).

Table 2. Testing for Factorial Invariance Across Genders

Model	χ^2	df	RMSEA	90% CI for		Δ RMSEA	CFI	Δ CFI	TLI	SRMR	Δ SRMR
				RMSEA							
				LL	UL						
Baseline girls ($N = 575$)	199.8*	128	.031	.023	.039	—	.942	—	.931	.040	—
Baseline boys ($N = 525$)	214.9*	128	.036	.027	.044	—	.933	—	.919	.044	—
M0 configural model	413.9*	256	.033	.027	.039	—	.938	—	.925	.042	—
M1 1st order- metric invariance	422.6*	267	.033	.027	.038	0	.939	.001	.930	.044	.002
M2 2nd order- metric invariance	433.6*	273	.033	.027	.038	0	.937	-.002	.929	.045	.001
M3 1st order- scalar invariance	467.5*	284	.034	.029	.040	.001	.928	-.009	.922	.047	.002
M4 2nd order- scalar invariance	473.2*	290	.034	.028	.039	0	.928	0	.924	.047	0

$N = 1100$

3.4 Gender Differences

The homogeneity of the variances' test indicated heterogeneity of variance for the global score of moral disengagement and all the mechanisms, the displacement of responsibility excepted. Therefore, we used an independent sample student *t*-test for the displacement of responsibility and Welch's *t*-test for the six remaining mechanisms and the global index. The results of the analysis (see Table B.1 in Appendix B) showed that boys reported to rely more on all the mechanisms, the displacement of responsibility excepted, and had a higher global score of moral disengagement than girls. Effect sizes were small to moderate, ranging from .11 to .38.

3.5 Moral Disengagement and Bullying Perpetration

Table 3 presents the results of the two-step hierarchical regression analyses with self-reported bullying (left) and peer-reported bullying (right) as the dependent variable.

Table 3. Regression Analysis for Bullying Perpetration

	Self-reported bullying				Peer-reported bullying			
	<i>B</i>	<i>SE B</i>	β	R^2	<i>B</i>	<i>SE B</i>	β	R^2
Step 1				.025				.068
Age	-.021	.015	-.041		-.013	.004	-.092**	
Gender (girls = 0, boys = 1)	.159	.031	.153***		.067	.008	.243***	
Step 2				.169				.087
Age	-.010	.014	-.020		-.011	.004	-.080**	
Gender (girls = 0, boys = 1)	.079	.030	.075**		.060	.008	.218***	
Moral justification	.045	.014	.100**		.008	.004	.064*	
Euphemistic labeling	.062	.015	.134***		.000	.004	-.003	
Advantageous comparison	-.013	.019	-.026		-.002	.005	-.012	
Displacement of responsibility	.013	.014	.028		.000	.004	.000	
Diffusion of responsibility	.055	.017	.106**		.006	.005	.047	
Distorting consequences	.049	.024	.078*		.003	.007	.020	
Victim attribution	.105	.022	.154***		.013	.006	.072*	

Note. Self-reported bullying (left): $F(2, 1086) = 13.9$, $p < .001$ for the first model; $F(7, 1079) = 24.5$, $p < .001$ for the second model; $\Delta R^2 = .145$. Peer-reported bullying (right): $F(2, 1094) = 39.6$, $p < .001$ for the first model; $F(9, 1087) = 11.5$, $p < .001$ for the second model; $\Delta R^2 = .019$.

* $p < .05$, ** $p < .01$, *** $p < .001$

Five of the seven moral disengagement mechanisms were significantly and positively correlated to self-reported bullying. Fourteen percent of variance in self-reported bullying is related to these five mechanisms, as indicated by the change in R^2 . When running the analysis a second time with victimization included in step 1, the change in R^2 dropped to 10.4% (which is not represented in the table for the sake of clarity). The two mechanisms that were not related to self-reported bullying were advantageous comparison and displacement of responsibility. Only two mechanisms had very small positive correlations with peer-reported bullying: moral justification and victim attribution. Two percent of variance in peer-reported bullying was related to these two mechanisms, as indicated by the change in R^2 .

Moral disengagement and Victimization

Two-step hierarchical regression analyses were performed on self-reported victimization (left) and peer-reported victimization (right) (Table 4). Two moral disengagement mechanisms were significantly and positively associated with self-reported victimization: moral justification and victim attribution. Three percent of variance in self-reported victimization is related to these two mechanisms, as shown by the change in R^2 . This proportion dropped to 1% when bullying is included in step 1 (which is not represented in the table for clarity purposes). Two mechanisms of moral disengagement presented very small significant correlations with peer-reported victimization; euphemistic labeling (negative), and victim attribution (positive). These two mechanisms explained 1.7% of variance in peer-reported victimization, as indicated by the R^2 change.

Table 4. Regression Analysis for Victimization

	Self-reported victimization				Peer-reported victimization			
	<i>B</i>	<i>SE B</i>	β	R^2	<i>B</i>	<i>SE B</i>	β	R^2
Step 1				.029				.035
Age	-.130	.023	-.168***		-.022	.003	-.188***	
Gender (girls = 0, boys = 1)	.041	.047	.026		.004	.007	.015	
Step 2				.062				.052
Age	-.117	.023	-.151***		-.021	.004	-.180***	
Gender (girls = 0, boys = 1)	.000	.048	.000		.004	.007	.015	
Moral justification	.050	.022	.074*		.000	.003	-.004	
Euphemistic labeling	.002	.023	.003		-.011	.004	-.105**	
Advantageous comparison	.023	.030	.030		-.003	.005	-.030	
Displacement of responsibility	.016	.022	.023		-.005	.003	-.044	
Diffusion of responsibility	-.015	.028	-.019		.005	.004	.043	
Distorting consequences	-.026	.038	-.027		.005	.006	.035	
Victim attribution	.154	.036	.150***		.012	.005	.076*	

Note. Self-reported victimization (left): $F(2, 1090) = 16.3, p < .001$ for the first model; $F(9, 1083) = 7.9, p < .001$ for the second model; $\Delta R^2 = .033$. Peer-reported victimization (right): $F(2, 1094) = 20.1, p < .001$ for the first model; $F(9, 1087) = 6.7, p < .001$ for the second model; $\Delta R^2 = .017$.

* $p < .05$, ** $p < .01$, *** $p < .001$

3.6 Moral disengagement and Bystanders' Behavior

To examine moral disengagement mechanisms associated with bystanders' behavior, two-step hierarchical regression analyses were carried out on defender behavior (left) and outsider behavior (right) (Table 5). Euphemistic labeling, distorting consequences, diffusion of responsibility, and victim attribution were negatively associated with defender behavior. Displacement of responsibility, on the other hand, presented a positive association, while moral justification and advantageous comparison were not related to defender behavior. Almost five percent of variance in defender behavior is related to these five mechanisms, as shown by the R^2 change. Euphemistic labeling, distorting consequences, diffusion of responsibility, and victim attribution were positively associated to outsider behavior, but moral justification, advantageous comparison, and displacement of responsibility were not related to it. The change in R^2 indicates that almost seven percent of variance in outsider behavior is related to these four mechanisms.

Table 5. Regression Analysis for Bystander's behavior

	Defender behavior				Outsider behavior			
	<i>B</i>	<i>SE B</i>	β	R^2	<i>B</i>	<i>SE B</i>	β	R^2
Step 1				.029				.019
Age	-.007	.028	-.008		.010	.025	.012	
Gender (girls = 0, boys = 1)	-.323	.056	-.171***		.236	.052	.137***	
Step 2				.076				.086
Age	-.011	.027	-.011		.016	.025	.019	
Gender (girls = 0, boys = 1)	-.261	.057	-.138***		.154	.051	.090**	
Moral justification	.030	.026	.037		-.006	.024	-.009	
Euphemistic labeling	-.064	.028	-.076*		.067	.025	.087**	
Advantageous comparison	.038	.036	.042		-.043	.032	-.052	
Displacement of responsibility	.067	.026	.079*		-.038	.024	-.050	
Diffusion of responsibility	-.077	.033	-.084*		.114	.030	.134***	
Distorting consequences	-.118	.045	-.104**		.120	.041	.116**	
Victim attribution	-.117	.042	-.095**		.093	.038	.083*	

Note. Defender behavior: $F(2, 1088) = 16.5, p < .001$ for the first model; $F(9, 1081) = 9.9, p < .001$ for the second model; $\Delta R^2 = .047$. Outsider behavior: $F(2, 1083) = 10.5, p < .001$ for the first model; $F(9, 1076) = 11.3, p < .001$ for the second model; $\Delta R^2 = .067$.

* $p < .05$, ** $p < .01$, *** $p < .001$

4. Discussion

This study aimed to validate the French version of the Moral Disengagement in Bullying Scale (MDBS; Thornberg & Jungert, 2014) in a sample of 8- to 12-year-old children. First, we examined the Fr-MDBS's construct validity by testing its factorial structure. Both the seven-factor model and its nested higher order-factor model had a good fit to the data. This supports the existence of an overarching dimension (moral disengagement) that encompassed the seven first-order factors, corresponding to the mechanisms. This result is important for both theory and research. From a theoretical perspective, the discrepancy between Bandura's conceptual model and the factorial structures of the developed scales was questioning either the theoretical model or its operationalization. The present study has addressed this issue by validating a factorial structure that is more congruent with the conceptual model. To our knowledge, only one study using the MDBS has demonstrated such a factorial solution to be better than a single dimension structure (Bjärehed et al., 2020), and none has demonstrated it to be as good as a first-order multi-dimension structure so far. From a research perspective, being limited to a single dimension of moral disengagement can lead to missing insight into the potential mechanisms involved in both bullies' and bystanders' behavior (Pozzoli et al., 2012). Yet, examining moral disengagement as a single dimension remains relevant given that moral disengagement mechanisms are conceptually believed to operate together and given the strong inter-correlations between them (Thornberg & Jungert, 2014). The validation of a second-order factor structure entitles researchers to examine either the

mechanisms or the global dimension of moral disengagement, according to their research goals.

Second, we tested the Fr-MDBS's measurement invariance across gender. As scalar invariance was reached, we could meaningfully compare boys and girls (Dimitrov, 2010). Gender differences followed the expected direction, which supports construct validity: Positive associations between being a boy and the use of all moral disengagement mechanisms were found, except for the displacement of responsibility. A specific subsection will be devoted to the findings related to this mechanism below.

Predictive validity of the Fr-MDBS was tested by examining associations between bullying perpetration, victimization, bystander roles (i.e., defender and outsider), and the factors of moral disengagement emerging from the CFAs. We expected bullying to be positively linked to moral disengagement factors (Gini et al., 2014; Killer et al., 2019). This study corroborated previous research, as most of the moral disengagement mechanisms were significantly and positively related to self-reported bullying, even when taking into account potential concurrent victimization. As in previous studies (Bjärehed et al., 2020; Thornberg & Jungert, 2014), both moral justification and victim attribution were significantly related to self-reported bullying. In addition, the use of euphemistic labeling, diffusion of responsibility, and distorting consequences was also related to self-reported bullying in our results. One explanation might be that the larger size of our sample enabled us to detect more significant associations. When considering peer-reported bullying, only moral justification and victim attribution were significantly related. This may lend support to a greater role of these two specific mechanisms in bullying behavior, which is in line with previous findings (Bjärehed et al., 2020; Thornberg & Jungert, 2014).

We hypothesized that victimization would be positively associated with moral disengagement (Killer, 2019). Our results confirmed this assumption, as children who tended to use moral justification and victim attribution as moral disengagement mechanisms were more likely to report being victimized, even when self-reported bullying perpetration was controlled for. Besides, victim attribution was also associated with peer-reported victimization. On the other hand, children who reported using more euphemistic labeling (i.e., using innocuous words to qualify bullying) were less likely to be designated by their peers as victims. This might be explained by a greater awareness among victims of the harmful effects bullying has on them. The association of victim attribution with both self- and peer-reported victimization is consistent with the tendency toward self-blame displayed by students experiencing chronic victimization (Graham & Juvonen, 1998; Tolmatcheff et al., 2019). In contrast with these results, Thornberg and Jungert (2014) did not find any mechanism associated with victimization when concurrent bullying perpetration was controlled. This is likely to be caused by insufficient statistical power due to the sample size, as suggested by the very small effect size of the associations found in our study as well as in meta-analyses (Killer, 2019).

Following recent recommendations (e.g., Hymel & Swearer, 2015; Volk et al., 2017), the present study used a combination of self-reported and peer-reported data for bullying

perpetration and victimization. Because of shared method variance, we expected stronger associations between moral disengagement and both bullying perpetration and victimization with self-reported data than with peer-reported data (Killer et al., 2019). Our results proved conclusive: moral disengagement mechanisms explained 12% more of the variance in self-reported bullying compared to peer-reported bullying, and 1.3% more of the variance in self-reported victimization compared to peer-reported victimization. It is noteworthy that expected associations, although weaker and fewer, can also be found with peer-reported measures.

Regarding bystanders' behavior, we assumed a negative association between moral disengagement and the defender role (Killer et al., 2019) but a positive association between moral disengagement and the outsider's role (e.g., Doramajian & Bukowski, 2015; Sjögren et al., 2020; Thornberg et al., 2017). These assumptions were mostly confirmed. As in Thornberg and Jungert's study (2014), defending was negatively predicted by the use of diffusion of responsibility and victim attribution. Furthermore, in our results, defending was also negatively predicted by using euphemistic labeling and distorting consequences. On the other hand, we found that children who reported using displacement of responsibility were more likely to report acting as defenders. Regarding the outsider role, we found a positive association with euphemistic labeling, distorting consequences, diffusion of responsibility, and victim attribution. These four mechanisms appear to play a particular role in bystander behavior as children who reported using them were both more likely to passively observe or walk away when bullying takes place and less likely to comfort the victim or stand up against the bully.

Displacement of responsibility presented an intriguing pattern throughout our results. This mechanism did not predict bullying perpetration (whether self- or peer-reported) or outsider behavior and was the only mechanism that was not used more by boys than by girls. In contrast, children who reported this mechanism were more likely to act as defenders. This result may be explained by children's unexpected reaction to the three corresponding items as observed by the researchers during the data collection phase. Many children reacted in an empathetic way, expressing compassion for the fictional children in the items that were having "family problems", had "parents who are mean or violent toward them", or had had "a poor upbringing", and agreed that these unfortunate children could not be blamed for their consequential bullying behavior. We can imagine that differently phrased items, measuring displacement of responsibility without eliciting the respondents' empathy for the fictional bullies, might yield different results. No significant link with defending was found in the original validation study (Thornberg & Jungert, 2014), but this could again be related to the small sample size. In contrast to these authors' and our results, Bjärehed et al. (2020) found that boys used displacement of responsibility more than girls. However, no significant associations with bullying perpetration were found in the regression analyses of any of the three studies.

4.1 Limitations and Future Directions

This study has some limitations. First, some of the subscales presented moderate ordinal alphas (e.g., diffusion of responsibility = .60), which weakens their reliability. Note that the original scale (Thornberg & Jungert, 2014) presented several low reliability indices (even below .60), as indicated in Appendix A. Yet, these rather low alphas can be explained by the small numbers of items in the subscales (Thornberg & Jungert, 2014). Second, as mentioned before, displacement of responsibility was unexpectedly interpreted by the children, which may have concealed this mechanism's actual contribution to bullying perpetration and bystanders' behavior. Future studies might test differently-worded items to better capture this moral disengagement mechanism. Some suggestions of items could be "If no adult is present to supervise the playground during recess, children cannot be blamed for bullying others" (inspired from Hymel et al., 2005) or "If someone got up on the wrong side of the bed, it is not his or her fault if he or she is being mean with other students". Third, we opted for behavior-based (as opposed to definition-based) self-reported measures of bullying perpetration and victimization (e.g., Brighi et al., 2012; Espelage et al., 2003), which do not specify the three definitional characteristics of bullying (repetition, intentionality, power imbalance). While this behavioral approach enables researchers to avoid some shortcomings of providing a definition or even using the word "bullying" in surveys (see Kert et al., 2010), it may also capture more generalized peer aggression and peer victimization. Fourth, as *dehumanization* and *blaming the victim* were originally merged into one factor (Thornberg & Jungert, 2014), the scale only taps into seven out of the eight theoretical mechanisms. This could imply either that the conceptual distinction between these two subdimensions does not empirically exist or that the items contained in the original pool could not capture them. Although the seven-factor structure already provides a rich picture of moral disengagement mechanisms, future studies should re-examine the existence of these two distinct factors. Finally, the directionality of the relationship between moral disengagement and bullying perpetration is not yet clearly established (Gini et al., 2014). While some studies suggest that moral disengagement predicts bullying behavior (e.g., Barchia & Bussey, 2011; Sticca & Perren, 2015; Troop-Gordon et al., 2019), others indicate the opposite (e.g., Obermann, 2013). The association might also be bidirectional, with moral disengagement enabling bullying perpetration and, in turn, repeated bullying perpetration making the use of the mechanisms easier (Gini et al., 2014; Mazzone et al., 2019). Longitudinal studies, and in particular longitudinal intervention studies that aim to decrease bullying perpetration by challenging students' use of moral disengagement mechanisms, are needed to clarify this issue (Gini et al., 2014).

4.2 Implications

As far as we know, only one study has assessed an intervention that explicitly targeted moral disengagement to help reduce bullying in primary schools, providing mixed but promising results (Wang & Goldberg, 2017). Several researchers have called for further research to determine the potential of this approach (Almeida et al., 2009; Hymel & Bonanno, 2014), notably given that such interventions could be a way of targeting both bystanders and

bullies (Obermann, 2011; Wang & Goldberg, 2017). The rationale for targeting moral disengagement in anti-bullying interventions is that outlining and invalidating the moral disengagement mechanisms should prevent children from using them. If they can no longer rely on these justifications, they will expose themselves to self-sanctions (guilt, shame), which, in turn, should discourage immoral behavior. In other words, bullying perpetration or refraining from defending the victim is expected to be less likely if children can no longer justify their behavior and therefore undergo unpleasant internal sanctions. Sound knowledge of the moral disengagement mechanisms could enable school psychologists and school staff in general to detect and invalidate them when dealing with a bullying situation.

Practically speaking, using storybooks displaying bullying situations to illustrate the use of moral disengagement mechanisms (e.g., Wang & Goldberg, 2017) seems to be a promising avenue to target moral disengagement. Bibliotherapy is often used by school psychologists and teachers with regard to conduct problems and can easily be integrated into the regular instruction curriculum. This makes it a particularly suitable approach for schools given their well-known time constraints (Wang & Goldberg, 2017). Moreover, integration of bullying prevention into daily instruction has been suggested as facilitating skill generalization (Leadbeater et al., 2016; Swearer et al., 2009).

Extending Thornberg and Jungert's work, the present study provides greater insight into the moral disengagement mechanisms related to bullying perpetration, victimization, and bystanders' roles, which can help to design more effective anti-bullying interventions. In that respect, interventions should address victim attribution in priority, given that this mechanism appeared to be related to all the investigated variables. By targeting victim attribution, interventions could prevent bullies from dehumanizing and blaming their victims, but also remind the victims that they are not responsible for their plight, exhort defenders to keep supporting the victims, and finally encourage outsiders to react when bullying occurs. Euphemistic labeling, diffusion of responsibility, and distorting the consequences were involved in both bullying perpetration and bystander (defender and outsider) roles, and are therefore worth targeting for anti-bullying interventions too. In particular, the role of diffusion of responsibility in dissuading bystanders from intervening is well already known (Salmivalli, 2010). Since moral justification was associated with both self-reported and peer-reported bullying, it could be targeted in selective interventions designed for bullies specifically. As highlighted in previous studies (Pornari & Wood, 2010; Thornberg & Jungert, 2014), when bullies use moral justification, they are likely to experience positive feelings for serving a moral purpose, which, in turn, will ease their bullying behavior. This mechanism, in particular, thus seems to perniciously reinforce bullying perpetration. Finally, advantageous comparison was not related to any of the examined variables. More studies are needed to demonstrate that this mechanism does not play any role in bullying dynamics, but this result suggests that anti-bullying interventions should not focus on it.

In conclusion, the validation of the Fr-MBDS empirically tested a second-order factor structure, thus narrowing the gap between Bandura's conceptual model of moral disengagement and its operationalization. This study contributes to enable researchers to validly examine either the mechanisms or the global dimension of moral disengagement,

depending on their research interests. Adding to the current literature, this study showed that more moral disengagement mechanisms are involved in bullying perpetration, victimization, defender behavior, and outsider behavior than was previously thought. This research also provided more detailed information about the respective contribution of these mechanisms to the bullying dynamics and brought useful insight into the mechanisms which can be prioritized for the design of anti-bullying interventions. Future intervention studies targeting moral disengagement in children should move beyond the single dimension of moral disengagement and investigate their effects on the distinct mechanisms in both bullies and bystanders.

Appendix A.

Reliability indices of the global index and seven factors of moral disengagement in the original validation study of the MDBS (Thornberg & Jungert, 2014): Moral justification ($r = .59$); Euphemistic labeling (Cronbach's $\alpha = .65$); Advantageous comparison ($r = .49$); Displacement of responsibility (Cronbach's $\alpha = .70$); Diffusion of responsibility ($r = .26$); Distorting consequences ($r = .56$); Victim attribution (Cronbach's $\alpha = .64$); Global index of moral disengagement (Cronbach's $\alpha = .84$).

Results of the Confirmatory Factor Analysis of the original validation study of the MDBS (Thornberg & Jungert, 2014): Satorra-Bentler scaled $\chi^2(114, N=372) = 118.56, p < .366$, CFI = .99, RMSEA = .01, 90% CI .00 - .03

Appendix B.

Table B.1. Gender differences in Moral Disengagement mechanisms

Moral Disengagement variables	Total		Girls		Boys		95% Confidence Interval				
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>t</i>	<i>p</i> -value	Lower	Upper	Cohen's <i>d</i>
Moral justification	1.72	1.16	1.54	1.09	1.9	1.21	-5.19 ^a	<.001	0.50	0.23	.32
Euphemistic labeling	1.05	1.13	0.91	1.03	1.2	1.21	-4.21 ^a	<.001	0.42	0.15	.26
Advantageous comparison	1	1.05	0.85	0.97	1.17	1.11	-5.16 ^a	<.001	0.45	0.20	.31
Displacement of responsibility	1.75	1.13	1.7	1.11	1.82	1.14	-1.74	.082	0.25	0.01	.11
Diffusion of responsibility	0.88	1.02	0.72	0.90	1.04	1.11	-5.21 ^a	<.001	0.44	0.20	.32
Distorting consequences	0.82	0.84	0.69	0.79	0.96	0.86	-5.38 ^a	<.001	0.37	0.17	.33
Victim attribution	0.61	0.77	0.53	0.71	0.69	0.83	-3.54 ^a	<.001	0.26	0.07	.22
Moral disengagement (global)	0.93	0.66	0.81	0.60	1.05	0.69	-6.32 ^a	<.001	0.31	0.16	.38

Note. a = Welch's *t*

N = 1107

* *p* < .05, ** *p* < .01, *** *p* < .001

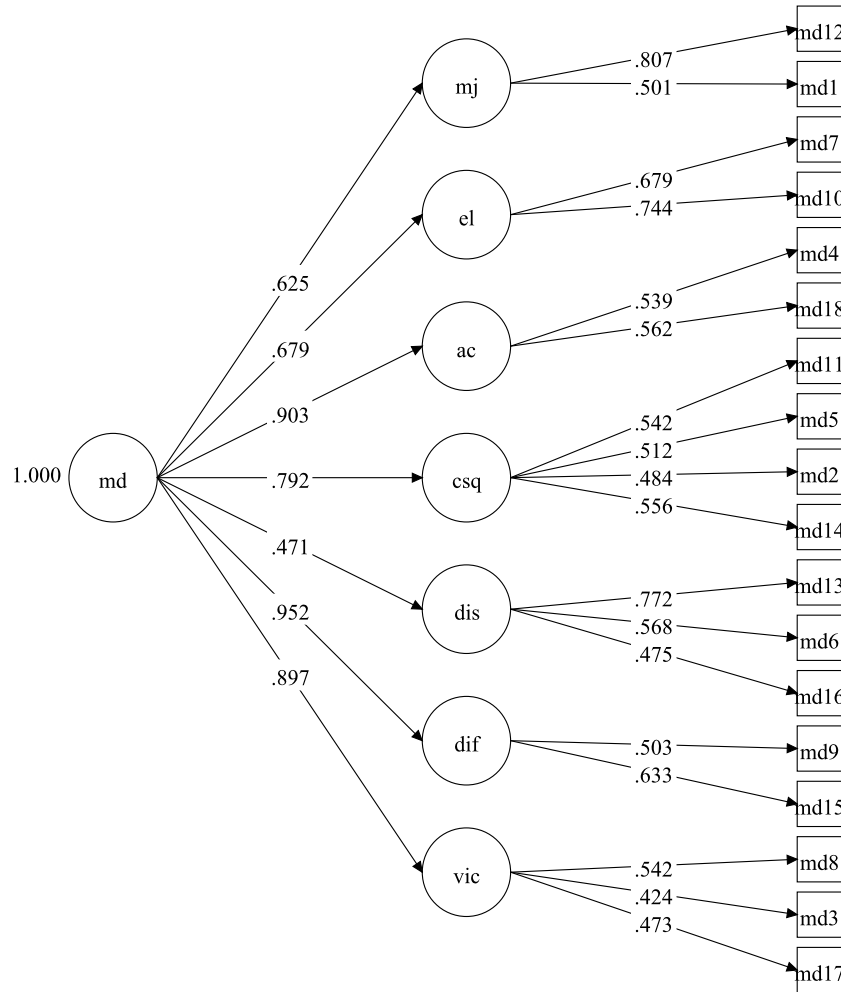
Table B.2. Inter-correlations between all variables

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Age	—														
2. Gender (girls = 0, boys = 1)	.01	—													
3. Self-reported bullying	-.04	.15***	—												
4. Peer-reported bullying	-.09**	.25***	.33***	—											
5. Self-reported victimization	-.17***	.02	.44***	.31***	—										
6. Peer-reported victimization	-.18***	.02	.15***	.35***	.41***	—									
7. Defender	-.01	-.17***	-.10***	-.03	.15***	.05	—								
8. Outsider	.01	.14***	.24***	.06*	.10**	.08**	-.30***	—							
9. Moral justification	-.02	.16***	.25***	.14***	.12***	-.00	-.06	.10**	—						
10. Euphemistic labeling	.02	.13***	.27***	.08**	.06*	-.08**	-.15***	.18***	.29***	—					
11. Advantageous comparison	-.07*	.16***	.21***	.10**	.09**	-.00	-.11***	.13***	.29***	.37***	—				
12. Displacement of responsibility	-.08**	.05	.16***	.06*	.08*	-.02	.01	.04	.25***	.20***	.23***	—			
13. Diffusion of responsibility	-.10**	.16***	.29***	.14***	.10**	.05	-.16***	.22***	.32***	.33***	.37***	.27***	—		
14. Distorting consequences	.00	.16***	.25***	.11***	.07*	.01	-.17***	.20***	.29***	.36***	.65***	.23***	.37***	—	
15. Victim attribution	-.09**	.11***	.30***	.15***	.18***	.08**	-.17***	.19***	.26***	.29***	.32***	.20***	.20***	.47***	—
16. Moral disengagement (global)	-.09**	.19***	.39***	.17***	.16***	.02	-.17***	.23***	.59***	.61***	.65***	.59***	.59***	.67***	.63***

$N = 1107$

* $p < .05$, ** $p < .01$, *** $p < .001$

Appendix C.



Note. md = moral disengagement; mj = moral justification; el = euphemistic labeling; ac = advantageous comparison; csq = distorting the consequences; dis = displacement of responsibility; dif = diffusion of responsibility, vic = victim attribution

Figure C.1. Standardized Factor Loadings of Model 5.

Supplementary material

Table S.1. Items of the original MDBS and Fr-MDBS

Item 1: It's okay to harm another person a couple of times a week if you do that to protect your friends	Item 1 : « On peut faire mal à quelqu'un si on fait ça pour protéger nos ami(e)s »
Item 2: Teasing a person a couple times a week is no big deal because you don't really hurt the person	Item 2 : « Ce n'est pas grave de se moquer de quelqu'un de temps en temps parce que ça ne fait pas vraiment mal »
Item 3: If people are weird, it is their own fault if they get bullied	Item 3 : « Si un(e) élève est bizarre, alors c'est de sa faute si il ou elle se fait embêter par les autres »
Item 4: Teasing a person a couple of times a week is no big deal because it's much worse to give the person a beating every week *	Item 4 : « Ce n'est pas si grave de se moquer de quelqu'un parce que c'est pire de le ou la frapper »
Item 5: People who get teased don't really get too sad about it	Item 5 : « Ça ne rend pas les gens vraiment tristes quand on se moque d'eux »
Item 6: If students are having a hard time at home, they can't be blamed for bullying others **	Item 6 : « Si des élèves ont des problèmes familiaux, alors ce n'est pas de leur faute s'ils embêtent les autres à l'école après »
Item 7: There's nothing wrong with name-calling a person a bit every day as long as you just do it as a joke	Item 7 : « Il n'y a rien de mal à insulter quelqu'un tous les jours tant que c'est pour rire »
Item 8: It's okay to bully those who are not like others	Item 8 : « On peut embêter ceux ou celles qui sont différent(e)s »
Item 9: If my friends begin to bully a classmate, I can't be blamed for being with them and bullying that person too *	Item 9 : « Si mes ami(e)s embêtent un(e) autre élève, ce n'est pas de ma faute si je l'embête aussi »
Item 10: Saying mean things to a certain person a couple of times a week doesn't matter. It's just about joking a little with the person	Item 10: « Ce n'est pas grave de dire des choses méchantes à quelqu'un de temps en temps. C'est juste pour rigoler un peu avec la personne »
Item 11: Nobody is harmed by being bullied a bit sometimes	Item 11 : « Se faire parfois un peu maltraiter par les autres, ça ne rend pas malheureux »
Item 12: It's okay to hurt a person a couple of times a week if you do that in order to help your friends	Item 12 : « On peut embêter quelqu'un de temps en temps si on fait ça pour aider nos ami(e)s »
Item 13: If students have parents who do bad things toward them, it's not the students' fault if they then bully other students **	Item 13: « Si des enfants ont des parents qui sont méchants ou violents avec eux, alors ce n'est pas de leur faute s'ils embêtent les autres à l'école après »
Item 14: It surely won't hurt you to be teased a bit now and then	Item 14 : « Ça ne rend pas triste de recevoir des moqueries de temps en temps »
Item 15: A student can't help that he/she is bullying another student if all his/her friends are doing it	Item 15: « Si tous/toutes mes ami(e)s embêtent un(e) élève, je ne peux pas m'empêcher de l'embêter aussi »

Item 16: If a child has got a poor upbringing, it's not the child's fault if he/she is teasing and beating a person every week	Item 16 : « Si un(e) enfant a été mal éduqué(e) par ses parents, ce n'est pas de sa faute si il ou elle se moque ou frappe les autres »
Item 17: If you can't be like everybody else, you have to blame yourself if you get bullied	Item 17 : « Si quelqu'un n'est pas comme tout le monde, alors c'est de sa faute si il ou elle se fait embêter par les autres »
Item 18: Teasing a person now and then every week is not so bad if you compare it to hitting and kicking a person every day	Item 18 : « Ce n'est pas si grave de rigoler de quelqu'un chaque semaine comparé à taper quelqu'un tous les jours »

* Minor modifications to simplify and shorten the item:

Item 4 was simplified into "Teasing a person is no big deal because it's much worse to give the person a beating".

Item 9 was simplified into "If my friends bully a classmate, I can't be blamed for bullying that person too."

** Minor modifications to adjust the item to French:

Item 6: An exact translation in French of "having a hard time at home" means nothing to the pupils, especially of this age. By contrast, they are very familiar with the notion of "having family problems", which was used instead.

Item 13: An exact translation in French of "parents who do bad things toward them" sounds peculiar and is likely to suggest a sexual implication. We used instead "parents who are mean or violent toward them".

Chapter 3

The Effectiveness of Moral Disengagement and Class Norms as Anti-Bullying Components

A Randomized Controlled Trial

This three-armed randomized controlled trial tested the effects of intervening on moral disengagement and class norms as anti-bullying components within separate interventions among 1,200 Belgian elementary students (48% boys, aged 9 to 12) from 57 classes in nine schools. 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 class norms decreased bullying ($\beta = -.18$), but did not influence the intended mediator (students' perception of their classmates' anti-bullying attitudes). The findings support the use of a moral disengagement component in anti-bullying programs. Testing the mechanisms of anti-bullying programs is crucial to improving bullying prevention.

This chapter is based on:

Tolmatcheff, C., Galand, B., Roskam, I., & Veenstra, R. The effectiveness of moral disengagement and class norms as anti-bullying components: A randomized controlled trial. *Child Development (submitted)*

The Effectiveness of Moral Disengagement and Class Norms as Anti-Bullying Components

A Randomized Controlled Trial

1. Introduction

Bullying is a situation of systematic, unprovoked, peer-to-peer abuse in the context of a power imbalance (Gini et al., 2011). Short- and long-term negative consequences have been documented not only for victims, but also for bullies and bystanders (Nishina & Juvonen, 2005; Salmivalli & Peets, 2018; Savahl et al., 2019). Several anti-bullying programs have been developed to tackle this universal and pervasive phenomenon. Although the findings of these evaluations have been mixed overall (Gaffney et al., 2019), some programs have shown positive effects on bullying and/or victimization (e.g., KiVa, Kärnä et al., 2011; ViSC, Yanagida, et al., 2016). However, it is still unclear what the causal mechanisms of effective antibullying programs are (Eisner & Malti, 2012; Saarento et al., 2015).

First, anti-bullying programs are often complex and combine numerous components at different levels of prevention (van Versveld et al., 2021). Because these components are generally assessed in combination, we do not know how each of them contributes to the effectiveness of programs. It is important to determine the effect of each component separately to understand how anti-bullying programs work (Menesini & Salmivalli, 2017). Second, most researchers do not assess the variables that they aim to manipulate to ultimately influence the desired outcome. Instead, they tend to focus on evaluating the outcomes and overlook the assessment of the hypothesized mediators (Saarento et al., 2015). To improve our knowledge and understanding of programs' causal mechanisms, we need to examine not only the effects of anti-bullying components on the program outcomes but also whether the effects flow through the expected mediators (Menesini & Salmivalli, 2017; Saarento et al., 2015). Although assessing the mediators is sometimes used as a way to identify programs' active components, the need for further innovative research to better understand programs' causal mechanisms has been pointed out (Eisner & Malti, 2012). Instead of evaluating whole packages of interventions, researchers have been encouraged to isolate promising anti-bullying components based on prior research and theory, and then to investigate their respective effects (Eisner & Malti, 2012).

Our study examines two promising anti-bullying components: intervening on moral disengagement and class norms (e.g., Almeida et al., 2009; Menesini et al., 2015). Most students disapprove of bullying and consider it immoral (e.g., Gini et al., 2011; Salmivalli, 2001). Nevertheless, many students bully others and few bystanders defend or support victims (Salmivalli, 2001). Both moral disengagement and class norms have been proposed

as rationales to explain this discrepancy between students' private attitudes and their actual behaviors, and have been suggested as promising directions for anti-bullying prevention (e.g., Almeida et al., 2009; Menesini et al., 2015). The goal of our study was to isolate these two anti-bullying components and to examine their distinct effects within separate interventions, while verifying that the effects flowed through the expected mediators, that is, moral disengagement and class norms. We assessed the two different interventions that exclusively targeted either moral disengagement or class norms in a sample of elementary school students (Grades 4-6) through a three-armed randomized controlled trial.

Many theories underlying anti-bullying programs have been based on cross-sectional studies. Intervention studies are not only an ideal means of testing these theories and providing possible insights for program improvement — they also help to expand our knowledge of the psychosocial developmental processes involved in bullying (Eisner & Malti, 2012). Our research design, in particular, provides direct information about the causal influence of moral disengagement and class norms on bullying, as each of the two interventions is designed to target and manipulate only one of these mediators.

1.1 Intervening on Moral Disengagement as an Anti-Bullying Component

Moral disengagement refers to selective disengagement by individuals from their own moral standards. Growing up, 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, 1986), 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. Students may, for example, argue that the bullying serves a moral and worthy purpose (e.g., “I do it to help my friends”), or negate it through sanitized and innocuous language (e.g., “It’s just a joke”). The eight moral disengagement mechanisms are listed in the Measures section.

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 provide support to the victim or stand up to the bully (Yun, 2019) – 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; Sjögren et al., 2020; Thornberg et al., 2017). 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).

Moral disengagement has been suggested as a promising avenue for prevention (e.g., Almeida et al., 2009; Hymel & Bonanno, 2014; Wang & Goldberg, 2017). Challenging students' use of moral disengagement mechanisms in connection with bullying should inhibit their ability to deactivate internal controls when transgressing their moral principles. This should thus leave students no longer protected from self-sanctions like guilt and shame, which, in turn, are assumed to deter immoral behaviors. In sum, 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.

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 and writing activities to prevent bullying among elementary children. A pilot study of the BLP-MD revealed a significant decrease in both moral disengagement and victimization following an intervention on third-grade classes (Wang & Goldberg, 2017). Appendix A presents more information about the BLP-MD and how we adapted it for this study's purposes.

1.2 Intervening on Class 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 (Pozzoli et al., 2012; Veenstra et al., 2018). 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 (Prentice & Miller, 1996). However, individuals' subjective perceptions of norms are not necessarily accurate (Tankard & Paluck, 2016). Young people tend to overestimate the prevalence of risky or problem behaviors and attitudes among their peers (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).

Patterns of misperception have been observed toward many risky behaviors, including school bullying in both elementary and middle schools (Perkins et al., 2011; Sandstrom & Bartini, 2010). In one study, the perceived pro-bullying norm was twice as high as the actual self-reported attitudinal norm (Perkins et al., 2011). As a result of pluralistic ignorance, bullies may freely engage in bullying, thinking that others approve of it, whereas bystanders may not express their disapproval, thinking they are alone in their opposition, or even support the bullying because they mistakenly feel social pressure to do so (Perkins et al., 2011; Sandstrom et al., 2013). Class norms have been found to influence students' behavior in bullying, regardless of their personal attitudes (Salmivalli & Voeten, 2004).

Norm-based interventions aim to correct people's misperceptions of their peers' attitudes and behaviors (e.g., Dillon & Lochman, 2019). The rationale behind norm-based

interventions is that providing group members with valid information about the norm can modify their perception of it. This alteration, in turn, makes a behavior more or less desirable and eventually induces a change of behavior (Miller & Prentice, 2016). In the case of pluralistic ignorance, interventions should attempt to correct biased attitudinal norms, referring to the perceived degree of (dis)approval of a behavior within the group (also known as *injunctive* norms), rather than biased behavioral norms, referring to the perceived prevalence of the behavior itself (or *descriptive* norms) (Miller & Prentice, 2016). 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 attitudinal class norms towards bullying should discourage bullies' behaviors and increase bystanders' willingness to intervene.

The second intervention used in our study is based on the Survey of Bullying at Your School, which is an anti-bullying intervention using students' self-reported data to show them the discrepancy between what they perceive to be other students' norms and students' actual norms towards bullying. A pilot study of the project in Grade 6-8 students revealed a significant decrease in both perceived pro-bullying norms and bullying and victimization (Perkins et al., 2011). Appendix A presents more information about the Survey of Bullying at Your School project and how we adapted it for this study's purposes.

1.3 The Present Study

This study aimed to assess and compare, within separate interventions, the effects of two anti-bullying components (intervening on moral disengagement and class norms), while verifying that the effects flowed through the expected mediators. A three-armed randomized controlled trial with a moral disengagement intervention condition, a class norm intervention condition, and a control group was conducted in a sample of elementary school students (Grades 4-6). The two interventions exclusively targeted either the moral disengagement component or the class norms component 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). Appendix A presents the two interventions in detail. 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 non-specific 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 whole study took place over the school year 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 one hour a week within a month after the training. Each class teacher was provided with material including a detailed description of each session and how it should unfold, supplementary material, and, for the control group only, videos.

We assumed that challenging students' use of moral disengagement mechanisms would restrain their use of the mechanisms. Our first hypothesis was thus that intervening on moral disengagement would decrease moral disengagement in students (Hypothesis 1). In parallel, we assumed that correcting the biased perception of attitudinal class norms regarding bullying would decrease the discrepancy between the actual classroom norm and the students' perception of that norm. Our second hypothesis was thus that intervening on class norms would increase students' perceptions of anti-bullying class norms (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).

2. Method

2.1 Sampling and Design

Schools were recruited on a voluntary basis. The recruitment drive was launched in September 2018 and broadcast among schools through various contacts in the field. Using G*Power 3.1, we obtained a minimum sample size of 967 students to reach 80% power for one-way repeated measures multivariate analyses of variance (MANOVAs). We needed a minimum of 48 classes with an average number of 20 students per class (OECD, 2013). Figure 1 presents a flowchart of the recruitment and retention of intervention and control schools. Appendix A provides detailed information about the power calculation, recruitment, eligibility criteria, randomization process, timeline, and data collection. Nine 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. Demographic characteristics of the sample are provided in Appendix A.

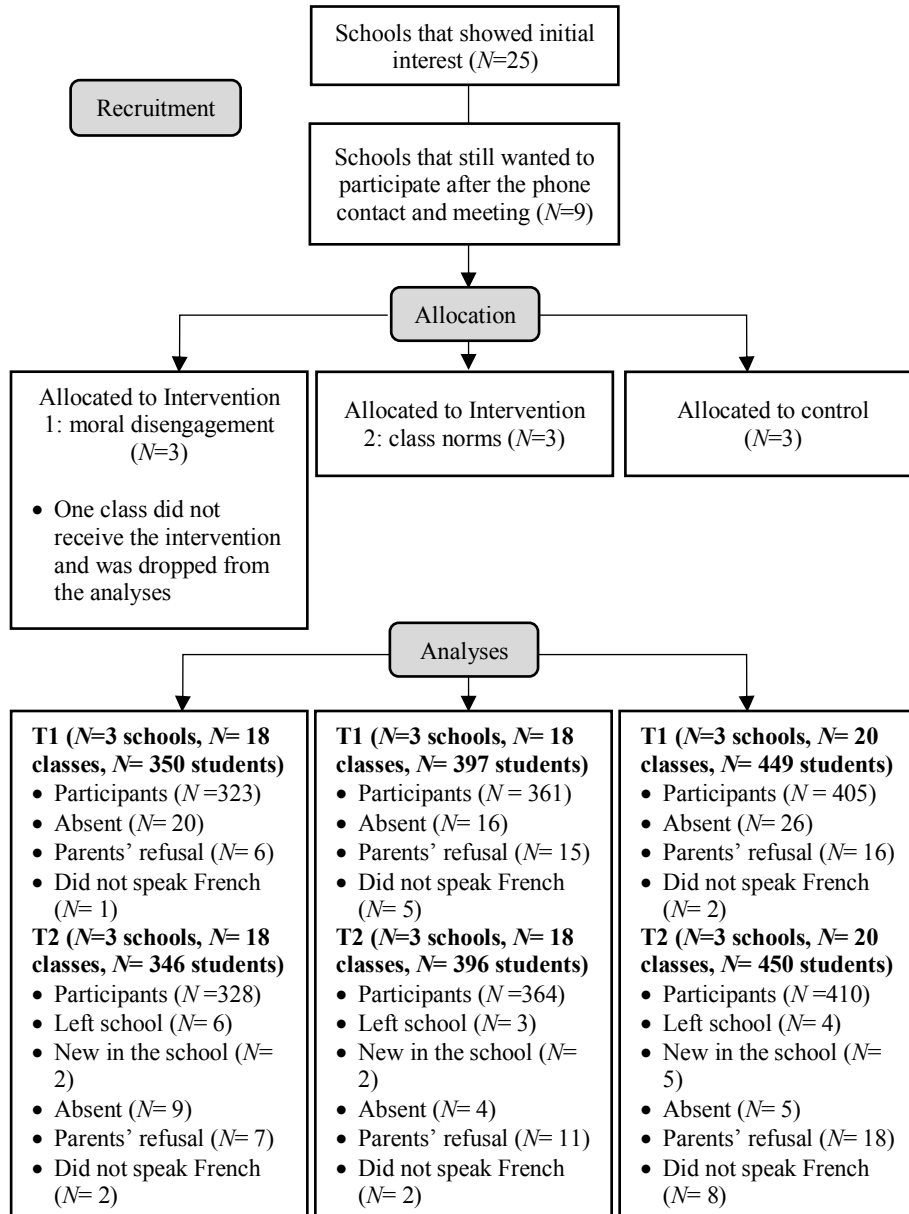


Figure 1 Flowchart of recruitment and retention of intervention and control schools

2.2 Data Collection and Participants

Data were collected in the fall of 2018 and the spring of 2019, 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 to two research team members, who were available to answer questions. In a five-minute 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 Appendix C for the CONSORT Checklist. Table A1 (in Appendix A) 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 described in Appendix A.

2.3 Measures

2.3.1 Moral Disengagement

We used the French-version of the Moral Disengagement in Bullying Scale (Fr-MDBS, Authors, 2021), adapted from Thornberg and Jungert (2014). The scale contains the moral disengagement mechanisms theorized by Bandura (1986): moral justification, euphemistic labeling, advantageous comparison, responsibility displacement, responsibility diffusion, disregarding or distorting the consequences, dehumanization and blaming the victim. 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 (Authors, 2021). 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), e.g., “*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: $\chi^2(70, N=1094) = 120.7, p < .001$, CFI = .97, TLI = .97, RMSEA = .026, 95% CI: .018-.033, and 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.

2.3.2 Perceived Attitudinal Class Norm

We translated into French eight items from Salmivalli and Voeten (2004) measuring attitudes toward bullying, e.g., “*Bullying may be fun sometimes*” (from 0 = totally disagree to 4 = totally agree). Following the original pilot study procedure for measuring the perceived attitudinal 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 $\alpha = .80$ at T2. As four of the six items were reversed, a method-specific factor was added to the measurement model of the scale to take account of this shared specific variance. The results of the CFA showed a good fit: $\chi^2(5, N=1091) = 9.5, p = .089$, CFI = .99, TLI = .98, RMSEA = .029, 95% CI: .000-.056, and SRMR = .013.

2.3.3 Bullying

We used a 10-item scale intended for primary school students (Authors, 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; 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 three months, e.g., “I have hidden or damaged another student’s things on purpose.” 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: $\chi^2(45, N=1097) = 34.8, p = .21$, CFI = .99, TLI = .99, RMSEA = .014, 95% CI: .000-.028, and SRMR = .025. Ordinal Cronbach’s alphas were $\alpha = .85$ (T1) and $\alpha = .83$ (T2) for the global construct, $\alpha = .74$ (T1) and $\alpha = .75$ (T2) for indirect bullying, $\alpha = .81$ (T1) and $\alpha = .81$ (T2) for direct bullying, and $\alpha = .85$ (T1) and $\alpha = .84$ (T2) for cyber bullying. In our sensitivity analyses, we examined the interventions’ effects on the three specific facets of bullying.

2.3.4 Defender Behavior

We used three items selected by Ruggieri et al. (2013) from the shortened version of the participant roles scale (Sutton & Smith, 1999), which was originally developed by Salmivalli and colleagues (1996). The items explored students’ potential defending behavior when witnessing a bullying episode: “I ask an adult or peer for help”, “I comfort the victim”, “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.

2.3.5 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 (instead of peer-reported) statements, like in the defender behavior items. For example, the description “Pretends not to notice what is happening” was turned into “I pretend not to notice what is happening”. The two other items were: “I don’t look in that direction or I stay away”, and “I don’t do anything”. Children were asked to indicate how often they behaved as described in each item (from 0 = “never” to 4 = “always”). Ordinal alpha was $\alpha = .70$ at T1 and $.79$ at T2.

2.4 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 attitudinal class norm (negatively skewed). Therefore, we first performed a square root transformation on these variables. In addition, we ran non-parametric Kruskal-Wallis tests on the non-transformed variables to check the results’ consistency. Table 1 provides the results of these preliminary analyses.

We performed structural equation modeling (SEM) with MPlus version 8.4. Standard error computations used a sandwich estimator to account for the non-independence of observations caused by the hierarchical structure of the data. The amount of variability in bullying and victimization is known to be substantially larger between classes than between schools (Kärnä et al., 2011). Examination of the intraclass coefficients (ICC) of all dependent variables (provided in Appendix B, see Table B2) confirmed that the between-school variability was negligible. Therefore, only the class level was taken into account in the analyses. In this case, the default estimation with complex survey data is MLR (maximum likelihood robust), which provides parameter estimates that are robust to non-normality and non-independence of observations, and is also a leading technique for handling missing data, using full-information maximum likelihood (FIML) (Enders, 2010). 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 post-test 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). Latent difference variables represent interindividual differences in true (error-free) intraindividual change over time and can be used like any other latent variable in an SEM (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 for correlations between the residuals corresponding to the same indicator over time (Geiser, 2013).

As LC models require at least strong factorial invariance, we tested measurement invariance across time for all dependent variables beforehand. 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 (>300) in the context of measurement invariance. Weak non-invariance 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 non-invariance, 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 versus 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 & Preacher, 2014). The final model included (a) regression paths between the two dummy-coded binary predictors (moral disengagement and class norms components) and the two mediating latent difference variables (moral disengagement and perceived attitudinal class norm), (b) regression paths between these two mediating latent difference variables and the three outcome latent difference variables: bullying, defender behaviors, and outsider behaviors, and (c) regression paths between the two binary predictors and the three outcome latent difference variables, corresponding to the direct effects of the interventions.

To assess the indirect effects' significance, we used the asymmetric bootstrap method, which does not assume the normality of the sampling distribution of the indirect effect. Two hundred bootstraps were required to obtain the 95% confidence intervals.

3. Results

Table 1 displays the descriptive statistics (means and standard deviations) of the mediating and outcome variables. Appendix B (see Table B1) provides correlations between all latent variables at T1. Despite the randomization process, preliminary analyses pointed to significant between-group differences for some variables of interest at baseline (detailed in Table 1). Students in the class norms (Intervention 2) condition initially reported slightly less bullying behaviors than the other groups. Students in the moral disengagement (Intervention 1) condition initially reported slightly more outsider behaviors than the other groups. Finally, students in the control group reported slightly more perceived anti-bullying attitudinal class norms than students in the class norms (Intervention 2) 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.

Table 1. Descriptive Statistics and Univariate Analyses of Variance at Baseline

Outcome variables	Intervention 1 moral disengagement						Intervention 2 class norms						Control group						Baseline differences		
	T1			T2			T1			T2			T1			T2			<i>F</i> ratio	<i>p</i>	η²p
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>			
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	0.80	328	2.84	0.79	366	2.81 ^a	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 ^{a,b}	0.46	382	0.41	0.45	405	0.52 ^b	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,b}	0.93	329	0.82	0.92	365	0.71 ^a	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 same subscripts differ at the $p = .05$ level in Bonferroni post hoc test.

Results from non-parametric Kruskal-Wallis tests were consistent with ANOVAs on the square root transformed variables.

* $p < .05$, ** $p < .01$, *** $p < .001$

Figure 1 is a path diagram of the structural equation model. The model includes latent variables: MORAL DISENGAGEMENT COMPONENT, CLASS NORMS COMPONENT, CHANGE IN MORAL DISENGAGEMENT, CHANGE IN PERCEIVED ANTI-BULLYING CLASS NORMS, CHANGE IN BULLYING BEHAVIORS, CHANGE IN OUTSIDER BEHAVIORS, and CHANGE IN DEFENDER BEHAVIORS. It also includes observed variables: State 1 (Y11, Y21, Y1i) and State 2 (Y12, Y22, Yi2). Path coefficients are shown for all relationships, with standardized coefficients in parentheses. Error terms are labeled epsilon.

Latent Variables:

- MORAL DISENGAGEMENT COMPONENT
- CLASS NORMS COMPONENT
- CHANGE IN MORAL DISENGAGEMENT
- CHANGE IN PERCEIVED ANTI-BULLYING CLASS NORMS
- CHANGE IN BULLYING BEHAVIORS
- CHANGE IN OUTSIDER BEHAVIORS
- CHANGE IN DEFENDER BEHAVIORS

Observed Variables:

- State 1 (Y11, Y21, Y1i)
- State 2 (Y12, Y22, Yi2)

Path Coefficients:

- MORAL DISENGAGEMENT COMPONENT to CHANGE IN MORAL DISENGAGEMENT: $-.46^*$
- CLASS NORMS COMPONENT to CHANGE IN MORAL DISENGAGEMENT: $(-.00)$
- MORAL DISENGAGEMENT COMPONENT to CHANGE IN PERCEIVED ANTI-BULLYING CLASS NORMS: $(.06)$
- CLASS NORMS COMPONENT to CHANGE IN PERCEIVED ANTI-BULLYING CLASS NORMS: $(-.11)$
- CHANGE IN MORAL DISENGAGEMENT to CHANGE IN BULLYING BEHAVIORS: $.33^*$
- CHANGE IN MORAL DISENGAGEMENT to CHANGE IN OUTSIDER BEHAVIORS: $-.10^*$
- CHANGE IN MORAL DISENGAGEMENT to CHANGE IN DEFENDER BEHAVIORS: $-.20^*$
- CHANGE IN PERCEIVED ANTI-BULLYING CLASS NORMS to CHANGE IN BULLYING BEHAVIORS: $.22^*$
- CHANGE IN PERCEIVED ANTI-BULLYING CLASS NORMS to CHANGE IN OUTSIDER BEHAVIORS: $-.18^*$
- CHANGE IN PERCEIVED ANTI-BULLYING CLASS NORMS to CHANGE IN DEFENDER BEHAVIORS: $(.08)$
- CHANGE IN BULLYING BEHAVIORS to CHANGE IN OUTSIDER BEHAVIORS: $(.19)$
- CHANGE IN OUTSIDER BEHAVIORS to CHANGE IN DEFENDER BEHAVIORS: $(-.08)$
- CHANGE IN BULLYING BEHAVIORS to CHANGE IN DEFENDER BEHAVIORS: $(-.10)$
- CHANGE IN OUTSIDER BEHAVIORS to CHANGE IN DEFENDER BEHAVIORS: $(-.02)$
- CHANGE IN DEFENDER BEHAVIORS to CHANGE IN BULLYING BEHAVIORS: $(-.05)$
- CHANGE IN DEFENDER BEHAVIORS to CHANGE IN OUTSIDER BEHAVIORS: $(-.02)$

Measurement Model:

- State 1 to Y11: λ_{11}
- State 1 to Y21: λ_{21}
- State 1 to Y1i: λ_{1i}
- State 2 to Y12: λ_{12}
- State 2 to Y22: λ_{22}
- State 2 to Yi2: λ_{i2}

Error Terms:

- ϵ_{11} , ϵ_{21} , ϵ_{1i} for State 1
- ϵ_{12} , ϵ_{22} , ϵ_{i2} for State 2

74

3.1 Moral Disengagement Component

The results indicated that intervening on moral disengagement significantly decreased moral disengagement in students ($\beta = -0.46$). Change in moral disengagement, in turn, was positively linked to change in bullying ($\beta = 0.33$) and outsider behaviors ($\beta = 0.20$) — 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 ($\beta = -0.10$) — thus a decrease in moral disengagement was associated with an increase in defender behaviors. The indirect effect on the outcome variables flowing through moral disengagement was significant in all three cases (bullying 95% CI $[-0.15, -0.05]$, outsider behaviors 95% CI $[-0.08, -0.02]$, and defender behaviors 95% CI $[0.00, 0.06]$). The results also displayed a slight but significant positive direct effect on the change in bullying behaviors ($p = .020$), as shown in Figure 2. However, this effect was not robust, as the asymmetric bootstrap 95% CI included zero $([-0.04, 0.21])$.

3.2 Class Norms Component

Perceived attitudinal class norms did not mediate the effects of the class norms intervention. More specifically, the class norm intervention did not yield change in the hypothesized mediator, which, in turn, did not predict change in any of the outcome variables. Accordingly, none of the indirect effects were significant. However, the class norms intervention had a small robust direct negative effect on bullying: 95% CI $[-0.31, -0.06]$ (see Figure 2).

3.3 Sensitivity Analyses

Sensitivity analyses yielded similar conclusions. Change in moral disengagement predicted change in indirect ($\beta = 0.30$), direct ($\beta = 0.27$), and cyber ($\beta = 0.17$) bullying. The corresponding indirect effects were significant for indirect (95% CI $[-0.22, -0.09]$) and direct bullying (95% CI $[-0.22, -0.08]$), and only marginally significant for cyber bullying (95% CI $[-0.16, 0.00]$). Intervening on moral disengagement had a small significant direct effect on indirect bullying ($\beta = 0.23$), but not on direct or cyber bullying. In addition, the small significant direct effect of the class norms intervention related to both indirect ($\beta = -0.17$) and direct bullying ($\beta = -0.17$), 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.

4. Discussion

Because most intervention studies assess anti-bullying programs' components in combination and omit to test the hypothesized mediators, our understanding and knowledge of the causal mechanisms of these programs remain limited. Accurately testing hypotheses and theories underlying anti-bullying programs is, however, crucial for the improvement of bullying prevention (Menesini & Salmivalli, 2017; Saarento et al., 2015). Intervention studies offer an ideal research design for both testing how anti-bullying programs work and expanding our knowledge of the psychosocial developmental factors involved in bullying.

This study aimed to assess, within separate interventions, the distinct effects of two promising anti-bullying components: intervening on moral disengagement and class norms (e.g., Almeida et al., 2009; Menesini et al., 2015) –, while examining whether these effects flowed through the expected mediators. We assumed that intervening on moral disengagement would decrease students' use of moral disengagement mechanisms (Hypothesis 1). Our results supported this first hypothesis. It is noteworthy that the moral disengagement component had no effect on the other mediator, the perceived class norms. This supports the idea that the intervention's design successfully isolated moral disengagement and did not incidentally elicit another anti-bullying component. We then assumed that the change in this mediator would, in turn, lead to a change in bullying (Hypothesis 1a) and bystander behaviors (Hypothesis 1b). These assumptions were also met, as the reduction in moral disengagement led to a decrease in bullying and outsider behaviors, and an increase in defending. The three corresponding indirect effects were significant, confirming the mediating role of moral disengagement in the intervention. In terms of practical significance in the domain of bullying prevention, it is noteworthy that these effects' magnitude is larger than that found in mediation analyses of other anti-bullying programs such as KiVa. For instance, at the student level, KiVa's standardized effects of anti-bullying attitudes, perceived proportion of defenders, and perceived teacher's attitude toward bullying on self-reported bullying ranged from 0.02 to 0.09 in absolute values (Saarento et al., 2015). In our intervention, moral disengagement had a standardized effect of 0.37 on self-reported bullying. Sensitivity analyses revealed that the intervention's indirect effects were significant for indirect and direct bullying, but only marginally significant for cyber bullying. This might be because of the very low prevalence of cyber bullying among elementary students (e.g., Monks et al., 2012).

Regarding the class norms component, our hypotheses were refuted. That is, the intervention did not lead to a change in the hypothesized mediator – the perceived anti-bullying attitudinal class norms (Hypothesis 2), and change in this variable was not related to change in any of the measured outcomes (Hypotheses 2a and 2b). However, the class norm intervention had a direct effect, decreasing both indirect and direct bullying behaviors.

A possible explanation for this result is that our measure of perceived attitudinal class norm failed to capture the intended mediator. Students in our sample were younger (Grades 4 to 6) than those from the original study (Grades 6 to 8; Perkins et al., 2011). Asking students to report what they believe that most of their classmates would answer, instead of their own

opinion, is a classical way of assessing the perceived injunctive norm (e.g., Prentice & Miller, 1993; Shelton & Richeson, 2005). However, it might have been too difficult a cognitive exercise, especially in the youngest as younger children's reasoning is more self-centered (Astington, 1993). Live observations during the data collection could support this hypothesis, as students sometimes reported some confusion about this part of the questionnaire to the researchers. However, this measurement procedure has already successfully been used in students as young as seventh graders (Dillon & Lochman, 2019).

An alternative 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 attitudinal 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 correctly implemented.

Finally, this result could imply that the effect did not flow through the expected mediator. It may be that by revealing that general approval toward a pro-bullying norm was illusory, the intervention did not lead students to cognitively reconstruct their estimate of the norm, but simply weakened the perceived social pressure to conform (Miller & Prentice, 2016; Prentice, 2008; Schroeder & Prentice, 1998). It may also be that change in students' bullying behaviors was mediated by another type of norm. For instance, popular students have been suggested as powerful class norm setters (Dijkstra et al., 2008; Veenstra & Lodder, 2022). Popular students' perception of the class attitudinal norm toward bullying might thus have primarily mediated bullying change. The subgroup of friends may also be seen as more influential than the whole class group (e.g., Paluck & Shepherd, 2012). Teachers also convey social norms among their students, both explicitly and implicitly (Veenstra et al., 2014). Finally, it is also possible that the intervention did not impact a 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' perception of their classmates' attitudes.

4.1 Strengths and Limitations

Strengths of this study are the relatively large sample with high retention rates, the use of latent change 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. This last point, in particular, is important for evaluation of anti-bullying programs (Volk et al., 2017). Effects on bullying can easily be missed in intervention studies because the intervention (predictor) is likely to have a stronger influence on the proximal variable it manipulates (mediator) than on the distal behavioral outcome (e.g., bullying) it aims to ultimately influence. As a result, researchers may conclude that there is no effect because of a lack of statistical power, whereas a mediation analysis would be able to detect the indirect effect flowing through the mediator (Rucker et al., 2011). A non-significant total effect when

mediation is present may also occur because another effect in the *opposite* direction is obscuring the relevant effect, or because measurement of the distal outcome is moderately reliable while that of the mediator is highly reliable – thus increasing power (Rucker et al., 2011).

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 schools' initial average levels on some chosen variables such as bullying or anti-bullying attitudes (e.g., Huitsing et al., 2020). 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, as mentioned above, the perceived attitudinal class norm measure might not have been well suited to the age of the students in our sample. Another way of measuring students' perception of their classmates' attitudes regarding bullying in the literature is to present them with different situations (e.g., "a classmate joining in the bullying", "a classmate telling the teacher about the bullying") and ask them how their classmates would react in each case (e.g., "the others would admire him/her", "the others would show disapproval of him/her"). This measure has been used in students of the same age as in our sample (Salmivalli & Voeten, 2004). However, we originally discarded it because the answer categories were nominal and did not allow for straightforward interpretation. Third, because all measures were self-reported, results may have been partially inflated because of shared method variance across the mediators and outcome variables. Alternatively, we could have analyzed peer reports of bullying and bystander roles, although there is no consensus about which of self- and peer-reported data are the most suitable method to assess bullying (Hymel & Swearer, 2015; Volk et al., 2017). However, classical peer-reported measures of bullying (e.g., "Who are the students who bully others in your class?") are reputational nominations, which are typically stable constructs (e.g., see the meta-analysis of Jiang & Cillessen, 2005). While peer reports are correctly framed as measures of reputation, it is not surprising that they are not sensitive to the frequency with which a specific behavior occurs (Gazelle et al., 2015). For this reason, reputational nominations may not be the most suitable for detecting objective behavioral change over a short period. Nevertheless, including both forms of reports may help compensating for their respective biases (Hymel & Swearer, 2015; Volk et al., 2017). Fourth, because this study was not replicated across different sociocultural groups (e.g., in other countries), our results may not generalize to other contexts. 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 cultural context does not threaten our findings' replicability. Nonetheless, both Belgium and the US 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 (Goncy et al., 2015), and that these fluctuations are likely to influence programs' outcomes (Durlak & Dupre, 2008). 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.

4.2 Implication and Future Directions

This intervention study provides sound evidence supporting the involvement of moral disengagement not only in bullying, but also in outsider and defender behaviors. It also has important implications for bullying prevention as it demonstrates that preventing students from relying on moral disengagement is an effective way of targeting both bystanders and bullies. Importantly, this study shows that students' use of moral disengagement mechanisms can be easily and quickly reduced through a relatively light (five hours) intervention delivered by class teachers. The use of children's stories to illustrate the moral disengagement mechanisms seems to be a suitable approach that can easily be integrated into the regular instruction curriculum (Wang & Goldberg, 2017). 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 (Authors, 2021).

Because some studies have pointed to differentials in the involvement of moral disengagement mechanisms in bullying and bystanders behaviors (e.g., Thornberg & Jungert, 2014; Authors, 2021), future intervention studies should examine the effects of the distinct mechanisms. Moreover, anti-bullying interventions targeting moral disengagement could be refined and improved over time based on which mechanisms need to be prioritized in light of these empirical results (Authors, 2021). Because the intervention had a marginally significant indirect effect on cyber bullying, future studies may investigate whether moral disengagement can also be effective in tackling cyber bullying. Such interventions will need to focus on cyber bullying-related stories, and may be more useful in high school students (with some adaptation to the students' age), as relatively few elementary students are concerned with cyber bullying (e.g., Monks et al., 2012).

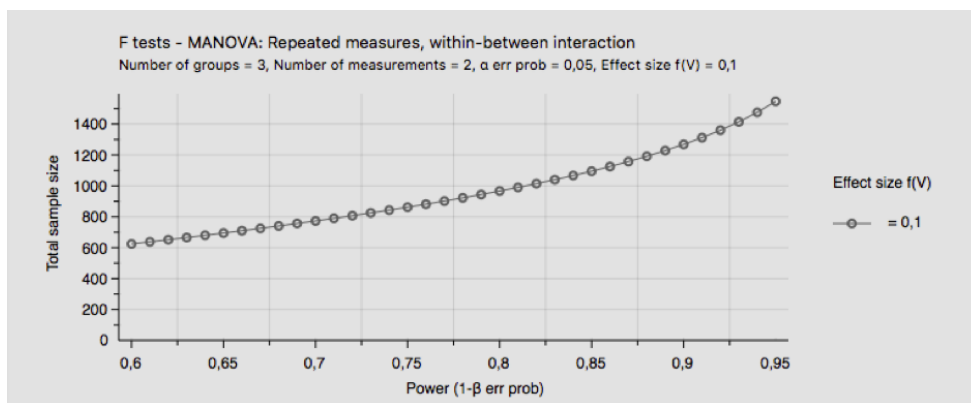
Given previous evidence suggesting an implication of class norms in bullying (e.g., Pozzoli & Gini, 2010; Rigby & Johnson, 2006; Salmivalli & Voeten, 2004; Sandstrom & Bartini, 2010; Sandstrom et al., 2013) and the direct effect on bullying found in our study, further intervention studies are needed to determine the potential of this component in bullying prevention. These studies should include and test several alternative mediators in order to clarify which of them play(s) a role in mediating change in bullying — notably different types of norm such as popularity norm, friend norm, and teacher norm. In addition, this pattern of results highlights the importance of using mediation analysis for assessing the effectiveness of an intervention to not only ensure that the intervention succeeded in changing the desired outcome variable, but also that the effects flowed through the hypothesized mediator(s). Finally, future studies should carefully monitor the implementation of anti-bullying interventions by teachers and examine its influence on the outcomes.

This study isolated two anti-bullying components – intervening on moral disengagement and class norms – and examined their respective effects within separate interventions, while verifying that these effects flowed through the expected mediators. Although the class norm intervention decreased bullying, this effect did not flow through students’ perception of their classmates’ attitudes toward bullying, thus leaving open the question of the processes involved. Further research is therefore needed to clarify the causal path of our class norm intervention. Our findings represent additional empirical evidence for the involvement of moral disengagement in both bullying and bystander behaviors and indicate that moral disengagement can be easily reduced in students. This study provides strong support for the integration of a moral disengagement component into anti-bullying programs and suggest that intervening on the class norms may be a promising anti-bullying component to further explore, provided that future studies clarify the underlying process conveying change in bullying.

Appendix A

Power Calculation

Using G*Power 3.1, we obtained a minimum sample size of 967 students to reach 80% power for one-way repeated measures multivariate analyses of variance (MANOVAs). We needed a minimum of 48 classes with an average number of 20 students per class (OECD, 2013). There are two same-grade classes in most Belgian elementary schools (two 1st grade classes, two 2nd grade classes, and so on). We wanted to recruit nine schools (three in each condition) in order to have a sample that would still be large enough if a school dropped out.



School Recruitment and Eligibility Criteria

Schools were recruited on a voluntary basis. The recruitment drive was launched in September 2018 and broadcast among schools through various contacts from the field. A letter describing the research project presented both the costs (the number of hours required for data collection, the training session, and the intervention) and the benefits that schools could expect. Each school would get a detailed assessment of their bullying situation and a

day of anti-bullying training for the whole teaching team the next year for free. In addition, in the two experimental conditions, 4th, 5th, and 6th grade teachers would also receive a specific training day and personalized supervision throughout the implementation of the interventions for free. Twenty-five schools expressed an interest in the project. These schools were contacted by phone by the first author to give them a broad outline of the project. Out of the 25 schools, three did not match the eligibility criteria (they had a preschool or 1st to 3rd grade section only, or less than five teachers in Grades 4-6), seven said they lacked time and/or had too many concurrent projects, two were hesitant, and four did not follow up after the phone contact. Nine school principals confirmed their interest immediately during the call. The first author met these schools' Grade 4-6 teachers between mid-October and mid-November to present the collaboration project. Teachers were informed about the randomization process. At the end of the meeting, they filled in a short confidential form to indicate whether they were willing to join the project. A minimum of five teachers per school was required to seal the agreement. All agreed to participate in the project ($N = 57$).

Demographic Characteristics

These schools differed in school size, ranging from 94 to 213 students in Grades 4-6. They were also diverse in terms of geographical location — covering four out of the six provinces of the French-speaking part of Belgium. Finally, they were diverse in terms of socioeconomic level: their official socioeconomic indices varied from 1 to 19. A school's socioeconomic index (ISE) is an official figure calculated annually and ranging from 1 to 20, based on five socioeconomic indices of their students' area of residence — per capita income, level of education, unemployment rate, professional activities, and comfort of accommodation. Each school was then randomly assigned to one of the three conditions by the first author using a draw procedure. To avoid inequitable group sizes, we randomly assigned the three biggest schools to each of the three conditions. Demographic characteristics of the three groups are detailed in Table A1.

Specific Culturally-Related Constraints in the Data Collection

In the French-speaking part of Belgium, schools are not at all aware of confidential demographic information such as the socioeconomic level of students' families. They only have access to the school's average socioeconomic index (ISE), as described in the previous section. Another culturally-related constraint in the demographic data collection relates to students' ethnic background: asking students about their race or ethnicity would be considered inappropriate, especially in elementary schools. As an illustration, there is no official census of students' ethnic background beyond their nationality in the French-speaking part of Belgium (Jacobs & Rea, 2015). However, researchers regularly indicate that students from non-Belgian ethnic backgrounds are present in particularly large numbers in the Region of Brussels, corresponding to Schools #2, #6, and #7 (Sacco et al., 2016).

Table A1. Demographic Information at T1

	Intervention on moral disengagement	Intervention on class norms	Control group
Boys (%) at T1	49.7%	49.1%	44.3%
Age <i>M (SD)</i> at T1	10.13 (1.06)	10.28 (0.98)	10.05 (1.02)
Grade			
4 th graders	38.1%	29.1%	35.7%
5 th graders	31.8%	24.1%	30.2%
6 th graders	30.1%	46.9%	34.1%
Geographical location (Province)	Liège ¹ Brussels ⁶ Namur ⁸	Brussels ² Namur ³ Namur ⁴	Brabant wallon ⁵ Brussels ⁷ Brabant wallon ⁹
School size (number of students in Grades 4-6)	98 ¹ 168 ⁶ 108 ⁸	129 ² 131 ³ 137 ⁴	142 ⁵ 213 ⁷ 94 ⁹
School ISE	7 ¹ 1 ⁶ 10 ⁸	9 ² 12 ³ 19 ⁴	14 ⁵ 1 ⁷ 15 ⁹

Note. Schools are indicated by superscript numbers (e.g., school number one was located in the Province of Liège, had 98 Grade 4-6 students, and had a socioeconomic index of seven).

Although there is no census of students' ethnic background in this part of Belgium, researchers regularly indicate that students from non-Belgian ethnic backgrounds are present in particularly large numbers in the Region of Brussels, corresponding to schools #2, #6, and #7 (Sacco et al., 2016).

Timeline

Teachers from Grades 4 to 6 attended the training day between mid-January and the end of February 2019. Schools began the intervention within a month after the training. The pre-test data collection took place between mid-November and mid-December 2018. The post-test data collection took place between the very end of March and mid-May. This somewhat longer period for the second data collection was because of the two-week spring break. On average, there was a time lapse of 15 weeks between the two waves.

Data Collection

Parental consent had been asked prior to data collection through passive consent forms sent by the schools. Students did not participate if parents objected to their participation or if they themselves did not want to participate. Students completed confidential paper questionnaires during class hours under the supervision of one to two research team members, who were available to answer questions. In a five-minute 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. For follow-up purposes, they were asked to create an identification code by joining the first two letters of their first name and the first two letters of their surname to the number of the day of their birthday (e.g., MA+DU+14). In fourth graders' classes (including two multi-grade classes with both 4th and 3th graders), the questionnaire was read aloud by a researcher to compensate for possible reading difficulties of students of this age. Non-response rates due to students' absences were higher at T1 than at T2 because we left empty

questionnaires with pre-stamped and pre-addressed envelopes with the teachers at T2. This way, students who had missed the data collection for some reason could participate within the following two weeks. The whole procedure was approved by the ethics committee of the psychological sciences research institute of the university of the first author. See Appendix C for the CONSORT Checklist.

Teacher Training

The teachers' training day in the two intervention groups was divided into two parts: the first part was devoted to bullying in general and the second part to the specific intervention (moral disengagement or class norms, depending on the group). The part on bullying aimed to provide general knowledge and raise awareness about the subject among teachers. Teachers were sensitized to characteristics, forms, participants' roles, social dynamics, prevalence, and consequences of bullying. The second part of the training started with a quick presentation of the theoretical and empirical foundations of the intervention. Next, teachers were instructed how to use the material. Then, they were immersed in an accelerated simulation of the five intervention sessions to familiarize them with the content and enable them to gain a feel for the expected mechanisms of change for themselves. Finally, various challenging situations that could occur during the sessions were envisaged (e.g., What if a student show resistance to or makes fun of the proposed activities?) to discuss the most appropriate reactions and exchange tips. Close monitoring was provided during the intervention period by means of exchanges in newly created WhatsApp or Facebook groups.

The teachers' training day in the control group started with a reminder of the whole research project's design and of the importance of the control group for the evaluation process. Then, the main issues relating to the world environmental crisis were presented to highlight the purpose of delivering this intervention to their students. Next, teachers were given an explanation of how to use the material; the sessions were described and the teachers were immersed in an accelerated simulation of the five intervention sessions. Close monitoring was also provided for teachers of the control group over the weeks of intervention.

Intervention 1: Moral Disengagement Component

The original BLP-MD is a five-session class-wide intervention that uses storybooks displaying bullying situations to prevent bullying among elementary children. It includes story reading, discussions with students, writing activities, and role play. A pilot study of the BLP-MD revealed a significant decrease in both moral disengagement and victimization in intervention 3rd grade classes (Wang & Goldberg, 2017). We did not use the BLP-MD storybooks in our study for two reasons. First, we thought they would be perceived as "too childish" by the students in our Grade 4-6 sample. Second, we judged that the bullies portrayed within these stories were rather stereotypical and presented as "the big mean bully that everyone is afraid of." This simplistic view of bullying does not reflect the increasingly recognized diversity of profiles among and motives of bullies (e.g., Peeters et al., 2010; Authors, 2018). We therefore created new stories of around 1,800 words each, which were adapted to the age of the students in our sample and designed to represent various bullying

situations (e.g., same-sex or cross-sex bullying, provocative or non-provocative victims, traditional or cyber bullying, bullying extended to the whole class or restricted to a few students, prejudice-related bullying, bullying stemming from some conflict). A summary of the five stories is provided below. The stories were intended to illustrate the use of moral disengagement mechanisms by the bullies. They did not contain other anti-bullying components such as bystanders' strategies for supporting the victim, because our goal was to assess the effects of the moral disengagement component. As such, the stories did not describe a resolution of the situation for the victims at the end. Instead, students were invited to identify the moral disengagement mechanisms (presented as a number of "little lies" on which the bullies relied to avoid guilt), which were then invalidated by the teachers. A typical session unfolded as follows: First, the teacher read the story aloud before handing out the written version of the story. Then, students were invited to answer individual questions (inserted in the written story) to reflect on the situation (e.g., "*When [name of the bully] says that 'it was just a joke', do you think it is true?*"). The teacher then chose one or two volunteers to share their answers with the others. Next, the teacher invited students to identify collectively the mechanism(s) used by the bullies in the story. Once identified, students were invited to take a minute of individual reflection to think about their own experience of using these mechanism(s). Then, the teacher reminded the students that moral disengagement mechanisms are false justifications that do not make the harmful behavior more acceptable. Students were finally invited to write the (simplified) name of the mechanism(s) discovered during the session on an individual summary sheet. 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 remember to avoid the mechanisms, committing to stop using the moral disengagement mechanisms in a confidential letter to themselves that would only be reopened by the students at the end of the school year).

Our study's goal was to assess anti-bullying components separately. For that reason, we focused only on students' use of moral disengagement mechanisms. In contrast to the BLP-MD, we did not use role play, which can enhance empathy (Bagès et al., 2020). Thus, we tried as much as possible not to manipulate any other component.

Summary of the Five Stories

(Story 1 – euphemistic labeling, advantageous comparison) At the beginning of the school year, Lumi finds out her two friends are no longer in the same class as her. Two girls in her class start to whisper and laugh every time she answers the teacher's questions – Lumi has no idea why. During a group work session, the two girls reject Lumi and make fun of her. When the teacher notices, the two bullies claim that "it was just a joke" and that "it is not that bad compared to insulting her." They prevent Lumi from making new friends and mock a classmate who tries to defend her. One day, Lumi ends up running out of the classroom, sobbing.

(Story 2 – moral justification) Leon used to be bullied in his previous school. He now has two new friends in his new school. Leon likes everyone in the classroom but Luc, who is a bit of a "weirdo". During a class trip, nobody wants to team up with Luc. The class teacher

decides to assign him to Leon and his friends' group. They plot to lose him in the museum they are visiting. Leon ends up violently pushing Luc off a trampoline, and the teacher catches the boy just before he badly hurts himself. Leon is punished but keeps repeating to himself that what he did was right because he did not want his friends' class trip to be ruined by Luc's presence.

(Story 3 – dehumanizing the victim, blaming the victim) Nobody has ever made fun of Cerise because of her cleft lip until a newly arrived student, Aël, starts calling her “the rabbit.” The other students progressively start using this nickname as well. During a recess, Aël and some followers run after her, trying to force her to eat a carrot. Cerise stops eating her lunches because other students pretend to be disgusted watching her eating. The class teacher informs Cerise's parents, and Aël gets summoned to the principal's office. He claims that Cerise is the one to blame for hanging around when she knows she is going to be teased. He even adds that she just has to be like everyone else.

(Story 4 – diffusion and displacement of responsibility) Laïs, a popular student, is very bossy with her classmates, especially since her parents split up. A boy called Seny is quite annoyed by her behavior. One day, Laïs decides to organize a beauty contest in the class. Seny refuses to participate. The situation escalates, and Seny, upset, tells Laïs that she is ugly. She replies by calling him a loser. The next morning, he finds ‘loser’ written on his table, and the other boys don't want to play basketball with him anymore. Someone throws a paper ball at him in the class. Seny then decides to apologize to Laïs, but she says it is too late and that she is not the only one calling him a loser – everybody does. He even receives an insult letter signed by all his classmates. When a friend of Laïs asks her why she doesn't go easier on him, she replies that it is not her fault if she is not being nice as her parents are divorcing.

(Story 5 – minimizing or distorting the consequences) Menea goes for a week-long school trip to the sea. All the girls in the class find one of the boys, Mehdi, very handsome. Two of them, in particular, follow him everywhere. On the last evening of the trip, Mehdi invites Menea for a dance during the party organized by the teachers. After returning home, Menea starts receiving insulting and threatening messages from an unknown phone number. She keeps receiving messages from three different numbers throughout the following week. One of the bullies wonders if they are going too far, but the two others retort that the messages will not kill Menea and that she is probably even happy to get some attention. At a certain point, Menea receives dozens of messages every day. She feels worse and worse. One day, she cannot take it anymore and bursts into tears in the middle of the class.

Intervention 2: Class Norms Component

The “Survey of Bullying at Your School” project is a five-session class-wide intervention using students' online self-reported data to highlight the discrepancy between perceived and actual pro-bullying norms. A pilot study of the project in Grade 6-8 students revealed a significant decrease in both perceived pro-bullying norms and bullying and victimization (Perkins et al., 2011). In the original project, social norm messages about bullying were created by the research team based on the online pre-test data. These messages were displayed on posters in the schools to provide students with feedback about the actual norms,

which were widely misperceived. We modified this procedure to let students discover their actual norms about bullying by themselves, in a more entertaining, instant, and offline format, better suited to our participants' age. During the first session, each student received ten small sheets with yes/no questions. The sheets referred to five very common bullying behaviors — bugging, excluding, making fun, insulting, and pushing/hitting. Five questions related to the perceived pro-bullying class norm and five other questions related to the corresponding personal pro-bullying attitude. In other words, the same behavior (e.g., excluding someone) was evaluated by each student in terms both of the perceived attitudinal class norm (e.g., “*Most of my classmates think that excluding someone is OK*” – yes/no) and personal attitude (e.g., “*I think that excluding someone is OK*” – yes/no). After the questions had been anonymously answered, the sheets were gathered in, then divided by type of behavior. Students were split into five work groups that received the sheets for both perceived attitudinal norm and personal attitude related to one of the behaviors. The composition of the groups had been decided on carefully based on the pre-test data and in collaboration with the class teachers prior to the intervention. Suspected bullies were mixed with prosocial students and key roles in the group were attributed to students who were not involved in bullying and were well liked. Each group was asked to count the number of yes and no answers related to the perceived norm and the personal attitude. The final counts of each group were displayed in colors on two parallel posters to visualize the discrepancy between students' perceived attitudinal class norm and their actual class norm, in the form of the sum of pro- and anti-bullying personal attitudes. For instance, the perceived norm poster might show that 12 out of 20 students thought that most classmates regarded excluding someone as OK, whereas the actual norm poster would show that only two out of 20 students thought it was OK. This discrepancy was underlined by the teacher and the *actual* norms of the class were emphasized. In the following sessions, groups were invited to create by themselves a slogan to describe their class's actual norm (e.g., “*In our classroom, we want to accept everyone*”). The slogans were supposed to emphasize the desired prosocial behavior rather than the condemned antisocial behavior (e.g., “*We accept everyone*” rather than “*We do not reject anyone*”). 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.

Appendix B

Table B1. Inter-correlations between Latent Variables at T1

Variables	1	2	3	4	5	6	7
1. Moral disengagement	—						
2. Perceived anti-bullying class norms	-.30***	—					
3. Bullying behaviors	.55***	-.24***	—				
4. Indirect bullying	.54***	-.24***	—	—			
5. Direct bullying	.47***	-.17**	—	.83***	—		
6. Cyber bullying	.29***	-.13***	—	.46***	.37***	—	
7. Outsider behaviors	.39***	-.18**	.37***	.35***	.33***	.13**	—
8. Defender behaviors	-.27***	.17**	-.19***	-.20***	-.15**	-.02	-.49***

N = 1166

* $p < .05$, ** $p < .01$, *** $p < .001$

Table B2. Intraclass Coefficients

Outcome variables	T1	
	School	Classroom
Moral disengagement	.005	.030
Perceived attitudinal class norm	.006	.037
Bullying (self-reported)	.029	.058
Defender behavior	.018	.015
Outsider behavior	.004	.040

N = 1166

Note. ICCs were computed from the average score of observed variables.

Table B3. Testing for Factorial Invariance Across Time

Model	χ^2	df	RMSEA	90% CI for RMSEA		Δ RMSEA	CFI	Δ CFI	TLI	SRMR	Δ SRMR
				LL	UL						
Moral disengagement											
M0 configural model	657.5*	321	.030	.027	.033	—	.933	—	.922	.037	—
M1 weak invariance 1 st	667.9*	328	.030	.027	.033	0	.933	0	.922	.038	.001
M2 weak invariance 2 nd	679.7*	334	.030	.027	.033	0	.932	-.001	.923	.039	.001
M3 strong invariance 1 st	700.2*	341	.030	.027	.033	0	.929	-.003	.921	.039	0
M4 strong invariance 2 nd	731.8*	347	.031	.028	.034	.001	.924	-.005	.917	.040	.001
N = 1163											
Perceived attitudinal class norm											
M0 configural model	46.4*	38	.014	0	.026	—	.996	—	.992	.018	—
M1 weak invariance	51.4*	46	.010	0	.023	-.004	.997	.001	.996	.021	.003
M2 strong invariance	63.6*	50	.015	0	.026	.005	.993	-.004	.990	.023	.002
M3 strict invariance	78.1*	56	.018	.006	.028	.003	.988	-.005	.986	.032	.009
N = 1161											
Bullying											
M0 configural model	199.9*	145	.018	.011	.024	—	.979	—	.973	.042	—
M1 weak invariance	191.9*	158	.014	.004	.020	-.007	.987	.008	.985	.035	-.007
M2 strong invariance	205.5*	165	.015	.006	.021	.001	.985	-.002	.982	.036	.001
M3 strict invariance	209.0*	175	.013	.003	.019	-.002	.987	.002	.986	.039	.003
N = 1163											
Defender behaviors											
M0 configural model	13.8*	5	.039	.015	.064	—	.992	—	.976	.018	—
M1 weak invariance	22.9*	7	.044	.025	.065	.005	.986	-.006	.969	.027	.009
M2 strong invariance	41.4*	9	.056	.039	.073	.012	.971	-.015	.951	.030	.003
M2P partial strong invariance	25.9*	8	.044	.026	.063	0	.984	-.002	.970	.028	.001
N = 1166											
Outsider behaviors											
M0 configural model	15.3*	5	.042	.019	.067	—	.986	—	.958	.027	—
M1 weak invariance	21.5*	7	.042	.023	.063	0	.980	-.006	.958	.035	.008
M2 strong invariance	23.7*	9	.038	.019	.056	-.004	.980	0	.967	.035	0
N = 1164											

Appendix C

CONSORT 2010 checklist of information to include when reporting a randomized trial

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomized trial in the title	☒
	1b	Structured summary of trial design, methods, results, and conclusions (for specific guidance see CONSORT for abstracts)	☒ (Within the 120 words-limit)
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	☒ See introduction, specifically <i>Intervening on Moral Disengagement as an Anti-Bullying Component</i> and <i>Intervening on Class Norms as an Anti-Bullying Component</i>
	2b	Specific objectives or hypotheses	☒ See <i>The present study</i> section
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	☒ See <i>Sampling and Design</i> and Fig.1 (Flowchart of recruitment and retention), and <i>Appendix A</i>
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	Not applicable
Participants	4a	Eligibility criteria for participants	☒ See <i>Appendix A</i>
	4b	Settings and locations where the data were collected	☒ See <i>Appendix A</i>
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	☒ See <i>Appendix A</i> and <i>The Present Study</i>
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	☒ See <i>The Present Study</i> , <i>Data collection</i> , and <i>Data Collection and Participants and Measures</i>
	6b	Any changes to trial outcomes after the trial commenced, with reasons	Not applicable
Sample size	7a	How sample size was determined	☒ See <i>Sampling and Design</i> and <i>Appendix A</i>
	7b	When applicable, explanation of any interim analyses and stopping guidelines	Not applicable
Randomization: Sequence generation	8a	Method used to generate the random allocation sequence	☒ See <i>Appendix A</i>
	8b	Type of randomization; details of any restriction (such as blocking and block size)	☒ See <i>Appendix A</i>

Opening the black box of anti-bullying programs

Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	☑ See <i>Appendix A</i>
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	☑ See <i>Appendix A</i>
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	☑ Blinding was not applied. Schools, teachers, and researchers were aware of the intervention or control status of schools.
	11b	If relevant, description of the similarity of interventions	☑ See <i>Appendix A</i>
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	☑ See <i>Analytical Strategy</i>
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	☑ See <i>Sensitivity Analyses in the Results</i>
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analyzed for the primary outcome	☑ See <i>Figure 1</i> (flowchart)
	13b	For each group, losses and exclusions after randomization, together with reasons	☑ See <i>Figure 1</i> (flowchart)
Recruitment	14a	Dates defining the periods of recruitment and follow-up	☑ See <i>Appendix A</i>
	14b	Why the trial ended or was stopped	Not applicable
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	☑ See <i>Table 1</i> for baseline comparison of the dependent variables and <i>Appendix A (Table A1)</i> for gender, age, grade, geographical location, school's size and socioeconomic level
Numbers analyzed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	☑ See <i>Figure 1</i> (flowchart)
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	☑ See <i>Figure 2</i> + bootstrap 95% confidence intervals of the indirect effects are provided in the <i>Results</i> section.
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	Not applicable
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	☑ See <i>Sensitivity Analyses in the Results</i>
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	☑ See <i>Results</i>
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	☑ See <i>Discussion (Limitations)</i>
Generalizability	21	Generalizability (external validity, applicability) of the trial findings	☑ See <i>Discussion (Limitations)</i>

Chapter 3. The Effectiveness of Moral Disengagement and Class Norms

Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	☒ See <i>Discussion</i>
Other information			
Registration	23	Registration number and name of trial registry	Not available
Protocol	24	Where the full trial protocol can be accessed, if available	Not available
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	☒

Chapter 4

Anti-Bullying Interventions Delivered by Teachers

**Examination of the Influence of the Implementation
Fidelity and Quality in a Randomized Controlled Trial**

Anti-Bullying Interventions Delivered by Teachers

Examination of the Influence of the Implementation Fidelity and Quality in a Randomized Controlled Trial

1. Introduction

Given the high prevalence of bullying in schools and its detrimental consequences for the children involved, an increasing number of anti-bullying programs have been developed over the past three decades (Low et al., 2014). Although meta-analyses have shown that anti-bullying programs can be effective, they do not consistently translate into positive outcomes (Bradshaw, 2015; Cross & Barnes, 2014). In most studies, a global overall effect of a given program is examined (Kallestad & Olweus, 2003). A closer look may reveal that program effects vary strongly between classrooms, and, to a lesser degree, between schools. These fluctuations are likely to be related to differential implementation, referring to how the programs are put into practice (Kärnä et al., 2011). This is especially true when programs are delivered by the teachers, which is often the case in bullying prevention (Goncy et al., 2015).

An explanation for the limited and/or mixed results of anti-bullying programs might thus be that not all teachers delivered the program as intended by the developers (Salmivalli et al., 2013). It is widely recognized that implementation strongly influences school-based prevention program's outcomes (Durlak & Dupre, 2008). A proper implementation of (essential parts of) a program is critical for any effects to occur (Kallestad & Olweus, 2003; Salmivalli et al., 2013). Nonetheless, for long, school bullying evaluation studies had rarely assessed implementation, or only at a very general level (Salmivalli et al., 2013).

As researchers have called for rigorous assessment of implementation in anti-bullying programs (Goncy et al., 2015), more evaluation studies nowadays report an index of implementation. However, most of them have only used it to ensure that a sufficient implementation level was reached, or as a possible explanation for mixed or non-significant results, or else to highlight practical issues related to the implementation and suggest directions for improvement (e.g., Axford et al., 2020; Clarkson et al., 2019; Gage et al., 2019; Nickerson et al., 2019). Few studies have examined how implementation is associated with the program outcomes (Haataja et al., 2014). This was the goal of the present study, which examined the influence of implementation on the effects of two different anti-bullying components delivered by elementary teachers within separate interventions in a randomized controlled trial.

1.1 Influence of implementation in anti-bullying programs

A main issue surrounding the assessment of anti-bullying program implementation concerns the unfortunate lack of consensus regarding the vocabulary and definitions in the implementation literature (Berkel et al., 2011; Durlak & Dupre, 2008). First, implementation science is a wide research area that encompasses the many parameters and factors that support or impede subsequent implementation, the issue of program sustainability over time, and the delivery process of the program content to the recipients (Cross & Barnes, 2014; Durlak & Dupre, 2008). This latter aspect is critical to interpret the effects of anti-bullying interventions accurately (Low et al., 2014). However, this aspect of implementation at the program users' level often remains concealed in the "black box" of implementation (Leadbeater et al., 2018). This is the focus of the present paper. In the following, we thus refer to the delivery process of the intervention content from the teachers to their students as 'implementation'.

Different theoretical frameworks from the prevention literature have provided useful insights on the relevant implementation features (e.g., Berkel et al., 2011; Domitrovich et al., 2012; Durlak & Dupre, 2008; Fixsen et al., 2005). However, as aforementioned, how implementation features have been defined, operationalized, and assessed varies among studies. In addition, our knowledge of which implementation features are decisive in influencing outcomes of interventions is still limited (Leadbeater et al., 2018), notably when it comes to the combined effects of different implementation features (Durlak & Dupre, 2008; Schultes et al., 2014). Below, we attempted to clarify how evaluation studies of anti-bullying programs have examined the influence of implementation on the outcomes so far. To do so, we conducted a systematic literature review on evaluation studies of anti-bullying programs that have assessed implementation and related it to the study outcomes. We wanted to select the relevant literature in the field of bullying prevention and implementation, and searched the electronic databases of Scopus, PsycInfo, and Eric for their high quality peer-reviewed research in education sciences and psychology. Appendix A details our search. Figure A1 provides a Prisma flow chart (Silva et al., 2017) displaying the identification and selection process of the articles. Table 1 presents which implementation features have been assessed and how they have been operationalized, as well as how their impact on the program outcomes has been examined for each of the 22 articles contained in the final selection.

Examination of these papers confirmed the terminological fuzziness that surrounds implementation (Berkel et al., 2011; Durlak & Dupre, 2008). The same terms have been conceptualized in various ways, and conversely, various terms have been used to designate the same concept. We organized the different concepts into four main categories, corresponding to the most used implementation features: *fidelity* (the extent to which the delivered intervention corresponds to the prescribed content and procedure), *dosage* (how much of the prescribed program has been delivered), *quality* (how well the program has been delivered), and *responsiveness* (recipients' attitudes toward the program, interest, and participation) (Durlak & Dupre, 2008). In the present selection, almost a third of the papers assessed at least two different implementation features. Although the different implementation features might be important to fully capture implementation (Berkel et al.,

2011), three quarters of the studies assessed fidelity ($N=16$), and 40% of them, dosage ($N=9$). Quality was only assessed in a fifth of the selected papers ($N=5$), and student responsiveness was the most rarely assessed ($N=3$).

The operationalization of fidelity (also referred to as adherence or completion) most commonly relied on checklists of activities or core components prescribed by the program or specific lesson. The fidelity index was then expressed as a percentage or as a sum of delivered elements. Fidelity was mainly self-reported by the teachers (or in some cases, by the program trainers), although observational data have been found to be more reliable than self-reports and are recommended if possible (Durlak & Dupre, 2008; Hirschstein et al., 2007).

Dosage (also referred to as duration, intensity, or participation) was mainly operationalized either as the number of intervention sessions or as the duration of the sessions. Regarding the former, the focus was either on how many sessions were delivered by the teacher or were actually attended by each student. Unless determined by the research design (e.g., comparison of two different dosages of the intervention), dosage was reported by the teachers or trainers.

When teachers deliver interventions, the quality of delivery has been suggested as particularly important (Berkel et al., 2011; Durlak, 2010; Goncy et al., 2015). The general influence of teachers on student behavior and peer relationships has been referred to as the “invisible hand” of the teacher (Bierman, 2011; Farmer et al., 2011). In the present selection, the time of lesson preparation, reported by teachers, was sometimes used as a proxy of quality (also referred to as integration or treatment intensity). A few papers assessed teachers’ generalization efforts (also referred to as integration), that is, to what extent teachers foster and coach students to apply the intervention principles in everyday life, beyond the prescribed lessons (Haataja et al., 2014). However, whereas this concept was considered as quality by some authors, it was seen as fidelity by others. The most comprehensive operationalization of quality relied on independent observational ratings of instructional clarity, classroom management, emotional tone, as well as student responsiveness (Hirschstein et al., 2007). This is in line with the idea that quality includes not only technical but also relational dimensions (Goncy et al., 2015; Haataja et al., 2014). Interestingly, these authors directly integrated student responsiveness in their measure of quality, as reflecting teachers’ ability to engage students (Hirschstein et al., 2007).

An issue connected to the dimensions of fidelity and quality is the role of adaptation from the providers. Although most researchers perceive adaptation as a threat to fidelity, both quantitative and qualitative studies suggest that some adaptation leads to better implementation and program outcomes (Durlak & Dupre, 2008). In this perspective, adjustments made by teachers to best fit the local context and their students’ needs can be seen as contributing to the quality (Cross & Barnes, 2014; Goncy et al., 2015). To find the right balance between fidelity and adaptation, the core components that should not be modified for the program to remain effective need to be identified (Durlak & Dupre, 2008).

The operationalization of student responsiveness (also referred to as engagement) varied a lot. We found different combinations of student liking of the program (self-reported),

student engagement, classroom management (teacher-reported), and students' absorption of the intervention principles (teacher- and parent-reported). This latter aspect, however, is not in line with common definitions, which underline student attitudes, interest, and involvement, but do not necessarily include a learning dimension (Low et al., 2014). It is noteworthy that similar dimensions can be found in both quality and responsiveness assessments (e.g., classroom management, student engagement), highlighting the close links between these two implementation features.

There was considerable variation in how studies examined the association between implementation and program outcomes. Most of them used either longitudinal or pre-/post-test research designs, but some used correlational and adjacent cohort longitudinal research designs. Sample sizes ranged from one to a thousand schools and the study period ranged from four weeks up to six years. Although data were often collected in the context of a larger randomized controlled trial, only data from the intervention group were analyzed in most studies. Half of the studies used implementation data as continuous scores, which has more statistical power than categorical scores (Durlak & Dupre, 2008). The type of statistical analyses varied a lot, from simple correlation analyses to sophisticated multilevel structural equation modeling. The overall pattern of results was mixed, with a fifth of the studies reporting results in the expected direction, another fifth reporting no significant results, and half studies reporting results in the expected direction only for some of the variables or subgroups of students (e.g., grade- or sex-specific). This heterogeneous pattern of results highlights the need for more studies examining the association between implementation and anti-bullying intervention outcomes, using high-quality research designs and statistical method.

Table 1. Study characteristics, operationalization, assessment, analysis, and main results of the association between implementation features and program outcomes

References	Sample	Implementation features	Operationalization	Assessment method	Analytical strategy and main results
KiVa (Kärnä et al., 2011)					
1 Salmivalli et al., 2005 (pilot intervention based on same principles as KiVa)	N=1,220 students at baseline; N=48 classes; N=16 schools Grades 4-6 Finland	fidelity and quality combined (<i>fidelity</i>)	• fidelity and quality: teacher reports collected four times a year were coded (what was done and how well it was done, from 0 to 3) by two independent raters (curriculum based class-level work and handling acute cases of bullying) Schools were divided into a low- and a high-implementation group.	teacher reports subsequently coded by raters following predefined criteria	Multilevel (logistic) regressions were conducted for each student outcome variable with the implementation level dummy coded and contrasted against the control group, controlling for sex. There was a significant effect of the program on student attitudes and efficacy beliefs toward bullying for schools with a high level of implementation. Regarding bullying and victimization, effects were found in fourth graders only, mainly in the group with a high level of implementation. Results regarding bystander behaviors were mixed.
2 Kärnä et al., 2011	N=749 teachers; N=425 elementary schools ; N=113 secondary schools Grades 1-6 and 7-9	dosage	Number of hours devoted to implementing the lessons + number of bullying cases processed by the KiVa team divided by the total number of students in the school	teacher self-reports	Correlation analyses were computed at the school level to examine whether the dosage was associated with the outcomes of the program, expressed as bullying and victimization difference scores between T2 and T1. The number of hours of implementation was associated with reductions in both bullying and victimization. The number of bullying cases was associated with reductions in bullying but not victimization.

3	Haataja et al., 2014	N=7,413 students; N=417 classes/ teachers; N=76 schools Grades 1-6 Finland	fidelity (<i>adherence</i>) dosage (<i>duration</i>) quality	• fidelity: dichotomous checklist (<i>used/did not use</i>) of the core activities of the 10 lesson plans – fidelity was expressed as a percentage. • dosage: duration (in minutes) of the lessons • quality: time of lesson preparation (in minutes) was used as a proxy All measures were averaged across the 10 lessons.	teacher self-reports	Multilevel and Latent Difference Score models (thus controlling for baseline scores) were used to examine the influence of the implementation features on the change in bullying and victimization at the classroom level, controlling for age and sex. Fidelity and quality (but not dosage) were associated with reductions in victimization at the classroom level. No significant effects were found for classroom-level bullying.
4	Johander et al., 2021	N=38,931 students; N=1,101 schools Grades 4-9 Finland	fidelity (<i>adherence</i>)	Five dummy-coded variables representing either recommended (<i>confronting, non-confronting, case-specific</i>) or not recommended (<i>own, unspecified</i>) intervention methods. + How often the KiVa team organized follow-up discussions (<i>no, occasionally, in all case</i>)	teacher self-reports	Multilevel regression analyses were conducted to test the effects of the method used and organizing follow-up discussions on the KiVa team- and student-perceived effectiveness of the intervention, controlling for the level of schooling (primary/secondary) and number of years in KiVa. Both the KiVa teams and students perceived the interventions to be less effective when the schools used their own adaptation or unspecified intervention method and did not organize follow-up discussions systematically.
Steps to Respect (Committee for Children, 2001) & Second Step (Committee for Children, 2008)						
5	Low et al., 2014	N=1,424 students; N=128 teachers; N=17 schools Grades 3-5 U.S.A.	fidelity (<i>adherence</i>) responsiveness (<i>engagement</i>)	All features were assessed using the web-based Program Implementation Checklists (PIC; Low, Brown, & Smith, 2008) • fidelity: percentage of activities outlined in each lesson (averaged across lessons) • responsiveness: four items (from 0=<i>not at all</i> to 3=<i>a lot</i>) assessing classroom management, to what extent students were engaged, distracted, and have absorbed the objectives of the lesson.	teacher self-reports	Multilevel models were used to test whether a) some class- and school-level factors predicted implementation, and b) implementation predicted student outcomes, controlling for baseline levels of the outcome variables and both student and school demographic characteristics. Class-level student climate and school percentage of students eligible for free/reduced-price lunch predicted student responsiveness. Class-level student attitudes toward bullying predicted fidelity. Fidelity was not associated with any of the student outcomes. Responsiveness was significantly associated with student support, student attitudes toward defending, and victimization.

6	Hirschstein et al., 2007	N=549 students; N=36 classes Grades 3-6 U.S.A.	fidelity quality quality-integration	• fidelity: dichotomous checklist of activities for each lesson • quality: assessment of clarity, student responsiveness, classroom management, and emotional tone (<i>poor, good, excellent</i>) + assessment of teachers' generalization efforts: Social-Emotional Learning Checklist - Bullying Report (SELC-BR; Hirschstein, Edstrom, & Frey, 2000) (from 1=<i>never</i> to 4=<i>four or more times in the past week</i>)	direct observation teacher self-reports (quality-integration)	Multilevel models were used to assess the influence of implementation on student outcomes, with pretest scores, grade and sex entered as fixed-effect variables. Student interpersonal skills reported by teachers were associated with fidelity. Quality was associated with more self-reported victimization and difficulty responding to bullying. Teachers' generalization efforts were globally associated with less observed victimization and aggression in older students, and less encouragement of bullying.
7	Polanin & Espelage, 2015	N=3,616 students at baseline; N=36 schools; N=66 Grade 6 and 78 Grade 7 teachers Grade 6 (T1 and T2) to 7 (T3) U.S.A.	dosage (<i>treatment intensity</i>) quality (<i>treatment intensity</i>)	Online teachers' survey responses were aggregated at the school level for each lesson, then averaged across the lessons for each school • dosage: lesson duration (one item) - quality: preparation time, three items (<i>yes/no</i>) assessing extra-investment (additional resources or information, additional financial resources, outside consultation)	teacher self-reports	Following meta-analytic procedure, meta-regression models were used to examine the relation between the effect sizes of the combined student outcomes (related to bullying and victimization) and the implementation variables. In sixth graders, the average effect size was significantly and positively predicted by preparation time and outside consultation, but dosage (lesson duration) negatively predicted the average effect size. No significant associations were found in seventh graders.

Viennese Social Competence program (ViSC; Spiel & Strohmeier, 2011)						
8	Schultes et al., 2014	N=370 teachers Austria	fidelity teacher responsiveness	At the school level (trainer reports): • fidelity: percentage of training units delivered • responsiveness: percentage of staff attendance to the trainings At the individual level (teacher reports): two items (post-test) assessing teachers' involvement into the program implementation (being a ViSC teacher, being a member of the ViSC school team)	ViSC trainer reports (school level) teacher self-reports	Multilevel models were used to examine the effects of implementation variables both at the individual- and school-level on teacher outcomes. Fidelity was associated to the change in teachers' self-efficacy. Teachers' individual responsiveness (being a ViSC teacher) was associated with the change in teachers' support of the victims.
9	Doğan et al., 2017	N=642 students; N=26 classes; N=6 schools Grade 5 Turkey	dosage	Two program dosages were compared: school and class level intervention versus school intervention only	/	Multilevel growth modelling was conducted to examine cross-level interactions between time and program dosages. There were no significant differences in effects on bullying, victimization, physical, relational, and verbal aggression and victimization between the two program dosages.
Olweus Bullying Prevention Programme (OBPP; Olweus, 1993)						
10	Black & Jackson, 2007	N=3,741 students (year 1); 1,598 students (year 4) Grades k-8 U.S.A.	fidelity	dichotomous checklist (<i>did / did not fulfill</i>) based on the 14 BPP core components – fidelity index was expressed as a percentage	direct observation	Correlations between fidelity (during year 4) and change in a) student self-reported bullying (Grades 3-8); b) observed bullying incident density were performed. Both correlations were not significant.

11	Ossa et al., 2020	N=5,759 students at baseline; N=23 schools; Grades 5-9 Germany	fidelity (<i>completion</i>)	fidelity: a comparison was made between schools that fully implemented the program (completer) and schools that canceled implementation within the 18-months period (non-completer).	/	A multilevel logistic regression analysis was conducted with program completion as a possible moderator. A significant reduction in victimization over time was observed for completer schools but not for non-completer schools. No interaction between fidelity (completion) and time was found for the number of bullies.
----	-------------------	--	-----------------------------------	--	---	---

WITS Program (Walk Away, Ignore, Talk it Out, and Seek Help) (www.witsprogram.ca)

12	Leadbeater et al., 2016	N=1,118 students at baseline; N=105 teachers at baseline; N=16 schools; N=1,113 parents Grades 1-4 (T1 to 3-6 (T4 and T5) Canada	fidelity (<i>adherence</i>) quality-integration (<i>integration/quality</i>) responsiveness	All items are dichotomous (yes/no) - fidelity: core activities of the program (five items) - quality-integration: teachers' reinforcement and integration of the program strategies into everyday practices (five items) - responsiveness: students' use of the program strategies at school and home (five items) Items reported by teachers (4), parents (3), and students (8) were combined into a single index for each dimension	multi-informant (self-)reports (teachers, parents, students)	Longitudinal path models were used to test whether a) fidelity and quality (T2-T3), mediated by student responsiveness, predicted change in the hypothesized mediators of the intervention at T4, and b) this change, in turn, was associated with improvement in outcome variables at T5, controlling for baseline scores (T1) of all variables. Both fidelity and quality were significantly associated with greater student responsiveness. Responsiveness, in turn, was associated with an increase in parent- and teacher-reported prosocial leadership and social responsibility. Change in these mediators, in turn, was associated with improvement in all student outcomes (victimization, aggression, and emotional problems)
----	-------------------------	--	---	--	--	--

13	Leadbeater et al., 2018	N=1,135 students at baseline; N=16 schools; N=70-75 teachers (T2-T5); N=1,131 parents at baseline Grades 1-3 (T1) to 3-5 (T4 and T5) Canada	fidelity integration children's use of program strategies (CUoPS)	All items are dichotomous (yes/no) - fidelity: core activities of the program (four items) - integration: teachers' reinforcement and integration of the program strategies into everyday practices (five items) - CUoPS: students' use of the program strategies at school and home (five items) - Items reported by teachers (3), parents (3), and students (8) were combined into a single index for each dimension	multi-informant (self-)reports (teachers, parents, students)	Auto-regressive cross-lag models including within-time correlations were used to examine the effects of the implementation variables on each student outcome, controlling for demographic variables (sex, SES, age). Integration and CUoPS were positively associated with social responsibility within-time. Fidelity and CUoPS were negatively associated with emotional problems. There was a significant effect of integration on both outcomes across time mediated by CUoPS. Prosocial leadership was associated with integration within time. No within-time or cross-lagged associations were found for victimization and aggression.
Creating a Peaceful School Learning Environment program (CAPSLE; Twemlow et al., 2001) & Gentle Warrior Program (component of CAPSLE)						
14	Twemlow et al., 2008	N=254 students Grades 3-5 U.S.A.	dosage (<i>participation</i>)	dosage: the number of sessions attended by the students (from 0 to 3)	records from the instructor	Multilevel regression analyses were conducted to examine the association between dosage and change in each student outcome variable, controlling for the baseline level of each variable and sex, grade, and ethnicity. Dosage was significantly associated with change in aggression and defending, but in boys only. These effects were respectively partially and fully mediated by change in empathy.
15	Biggs et al., 2008	N=122 students; N=16 classes/ teachers; N=3 schools Grades 3-4 (T1) to 4-5 (T2, T3) U.S.A.	fidelity-integration (<i>adherence</i>)	- CAPSLE Use scale: frequency of use of the program components in the classroom management (four items, from 0=<i>rarely/never</i> to 5=3+ <i>times/day</i>) (average score) (observation ratings of one of the items were used to establish criterion validity)	teacher self-reports (+ direct observation for criterion validity)	Fidelity was used in regression analyses to predict students' outcomes while controlling for the initial level (previous year) of each outcome variable. After one year of implementation, fidelity significantly predicted change in students' self-reported empathy for victims and attitudes toward aggression, and in peer-reported bystanders' defending behaviors. After two years, fidelity significantly predicted change in defending only.

School-Wide Positive Behavioural Interventions and Supports (PBIS; Chapman & Hofweber, 2000; McIntosh et al., 2011).

16	Kelm et al., 2014	<i>N</i> =200 elementary students; <i>N</i> =one school	fidelity	The Schoolwide Evaluation Tool (SET; Sugai et al., 2001) provided an overall score (%) of critical features implemented	direct observation ^b	Descriptive changes in behavioral outcomes from partial implementation (Year 1) to full implementation (Year 2) were examined.
		Grades 4 and 7 only for safety and bullying		Results from two consecutive years were compared, assuming that a greater fidelity was reached Year 2 (>80%) compared with Year 1 (<80%)		The number of office discipline referrals decreased by more than a half. The number of school suspensions or expulsions decreased by a half. Student-reported feeling of safety increased more than in other schools from the district. Student-reported victimization decreased more than in other schools from the district.
		Canada				

School-based drama program

17	Joronen et al., 2012	<i>N</i> =134 students; <i>N</i> =eight classes; <i>N</i> =two schools	dosage (<i>intensity</i>)	Two program dosages were compared: eight or more intervention sessions as intended by the program (<i>N</i> =60, 3 intervention class out of 4) versus less than eight intervention sessions	(not specified)	Repeated measures analyses of variance (ANOVA) were used to test whether there was an interaction between the group (intervention/control) and time on social relationships in the class.
		Grades 4-5				The interaction effect was marginally significant when considering the whole sample, but significant when considering the high-intensity intervention subsample only. No effects of the intervention were found for bullying or victimization, controlling for sex and baseline score.
		Finland				

Boston vs. Bullies Program					
18	Greif Green et al., 2020	N=338 students; N=19 classes; N=7 facilitators (teachers); N=6 schools Grade 5 U.S.A.	fidelity (<i>adherence</i>) responsiveness (<i>engagement</i>)	- fidelity: dichotomous checklist (<i>did/did not implement</i>) of components for each of the four lessons – summed to create a fidelity index - responsiveness^a: three items assessing how much students liked the program, learned, (<i>not at all, a little, a lot</i>) and whether they would recommend it (<i>yes/no</i>); one item assessing student engagement (from 1=<i>not at all engaged</i> to 4=<i>very engaged</i>) All measures were averaged across the four lessons.	fidelity: teacher self-reports responsiveness student self-reports teacher ratings (one item) Linear regression models were used among students of the intervention group to test whether fidelity, and both student- and teacher-reported responsiveness were associated with each outcome variable at T2 while controlling for the corresponding baseline (T1) score. Fidelity was associated with decreased self-reported victimization, bullying, and fighting. No significant association was found with teacher ratings of student engagement. Liking the program was associated with more perceived adult response to bullying, feeling of responsibility to intervene, anti-aggression attitude, and less bullying. Learning from the program was associated with reductions in cybervictimization. Being likely to recommend the program was associated with more perceived adult response to bullying.
Safe School Ambassadors (SSA) bystander education program					
19	Pack et al., 2011	N=19 schools U.S.A.	global implementation (<i>fidelity</i>)	a 10-item program implementation index included the number and accountability of active student “ambassadors”, frequency and attendance to meetings, and principal’s and staff’s support	school-reports Based on the implementation index, the upper-half of intervention schools was selected for correlation analysis between implementation and discipline indicator. A higher level of implementation of the program was associated with a decrease in the number of discipline incidents.
Promoting Positive Peer Relationships (P3R; Faull et al., 2008)					
20	Renshaw & Jimerson, 2012	N=636 students; N=12 classes; N=one teacher; N=one school Grades 7-8 U.S.A.	dosage (<i>implementation duration</i>) fidelity ^d	dosage: two dosages of the program (five versus eight lessons) were compared. /	ANCOVA analyses were conducted to examine whether there was an interaction effect between the group (intervention versus control) and dosage, controlling for baseline levels of student outcomes. There were no significant interaction effects for student attitudes toward bullying or perceived school bullying support.

School-based life skills program Lions Quest (Wilms & Wilms, 2011)						
21	Matischek-Jauk et al., 2018	N=363 students; N=16 classes; N=three schools Grades 5-8 Austria	global implementation (<i>adherence</i> , <i>quality</i> , <i>dosage/exposure</i>) responsiveness ^c	The intervention classes were classified into three levels of implementation (low, average, high), based on both teacher self-reports (number and topic of the sessions, adaptations if any, general outcome, classroom management) and data from the semi-structured interviews.	teacher self-reports interviews	Multilevel analyses were conducted. The three implementation levels were dummy coded and contrasted against the control group, controlling for baseline levels of the outcome variables. There was only a medium effect of the intervention on one dimension of the classroom climate for the low implementation group. There were small and medium to large effects on both the classroom climate and bullying for the average and high implementation group, respectively. No effect was found on psychosocial health for any of the three levels of implementation.
Setu (सेतु)						
22	Sharma et al., 2020	N=174 students; N=five classes; N=two schools Grade 8 India	dosage fidelity (<i>adherence</i>) ^c	dosage: students were classified into two groups based on whether they attended all four sessions or not.	teacher self-reports	T-tests were conducted to test for differences between the two dosage groups in outcome variables (bullying, aggression, victimization) at baseline, post-test, and follow-up. No significant difference was found between the students who did or did not attend all sessions.

^a In addition, nine items assessed students' experience with the program (from 1=*strongly disagree* to 4=*strongly agree*) but were not used in the analyses.

^b Additional interviews were conducted but were not used in the analyses.

^c Responsiveness was assessed through focus groups with students but was not used in the analyses.

^d The percentage of components delivered by lesson (fidelity) was assessed by the P3R Implementation Fidelity Checklists (P3R-IFC, teacher self-reports) but was not used in the analyses. In addition, direct observation was used to establish criterion validity of this measure of fidelity.

^e Components delivered were logged by the providers to assess fidelity but this was not used in the analyses.

1.2 Moral Disengagement and Class Norms as Anti-Bullying Intervention Components

Our study is part of a larger research project aiming to evaluate the effects of two anti-bullying program components, selected on the basis of prior research and theory: moral disengagement and class norms. There is a discrepancy between students' private attitudes toward bullying and their actual behaviors. Whereas most of them consider bullying immoral and disapprove it, many bully others and few of them would stand up for the victims (Hymel et al., 2010; Salmivalli, 2001). Both the moral disengagement and class norms frameworks provide a rationale to explain this paradox and have been suggested as promising directions for anti-bullying prevention (e.g., Almeida et al., 2009; Menesini et al., 2015).

This apparent paradox between students' private attitudes toward bullying and their actual behaviors can be analyzed through the lenses of Bandura's model of moral disengagement (Hymel & Bonnano, 2014). This framework is based on the idea that individuals can selectively disengage from the moral principles that normally guide their behavior, which they integrate over the course of their development. Following these principles leads to self-satisfaction and a sense of self-worth, whereas transgressing them leads to self-sanctions such as guilt and shame. However, internal controls can be selectively disengaged through cognitive processes used to reinterpret the situation. Relying on these so-called moral disengagement mechanisms, one can behave in an immoral way without undergoing self-sanctions. In the context of bullying, students may, for example, argue that their act is not that serious when compared with an even worse one (e.g., "It is not like if I had punched him/her in the face") or diffuse the responsibility among several people (e.g., "Everybody does it anyway") (Bandura, 1986). Moral disengagement has been found to be positively associated with bullying in cross-sectional studies (see Killer et al., 2019 for a recent meta-analysis) and is a promising avenue for prevention (e.g., Almeida et al., 2009; Hymel & Bonanno, 2014; Wang & Goldberg, 2017). If students can no longer use moral disengagement mechanisms, they should not be able to deactivate internal controls when transgressing their moral principles. As a result, they would undergo the unpleasant self-sanctions that they were trying to escape, which should in turn discourage their immoral behavior.

Social norms, referring to what is typical or appropriate in a group or a situation, are the second explanatory framework for the paradox between students' private attitudes toward bullying and their actual behavior (Prentice & Miller, 1996). Norms shape, constrain, and maintain behaviors of group members over and above individual attitudes (Pozzoli et al., 2012; Veenstra, 2021). However, what individuals perceive to be the norm in a context is 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 regarding bullying (Dillon & Lochman, 2019; Perkins et al., 2011; Sandstrom & Bartini, 2010; Sandstrom et al., 2013). In one study, the perceived pro-bullying norm was twice as high as the actual self-reported attitudinal norm (Perkins et al., 2011). Class norms influence students' behavior in bullying, regardless of their private attitudes (Pozzoli & Gini,

2010; Rigby & Johnson, 2006; Salmivalli & Voeten, 2004; Sandstrom & Bartini, 2010; Sandstrom et al., 2013). As a result, most of them privately disapprove of bullying and pro-bullying attitudes but believe that the others approve of it. This biased norm, in turn, may enhance bullying as students mistakenly feel support or even social pressure to do so (Perkins et al., 2011; Sandstrom et al., 2013). Interventions targeting norms aim to rectify people's misperceptions of their peers' attitudes or behaviors (e.g., Dillon & Lochman, 2019). Providing group members with correct information can change their perception of the norm, making a behavior or attitude more or less desirable. This change of perception is likely to induce a change of behavior (Miller & Prentice, 2016). Thus correcting students' misperception of their classmates' attitude toward bullying should discourage bullies' behavior.

1.3 The present study

The goal of our study was to examine the combined influence of implementation fidelity and quality on the outcomes of two different anti-bullying components delivered by elementary teachers within separate interventions in a randomized controlled trial. The two interventions exclusively targeted either moral disengagement or the class norms as mediator of change in bullying. We used an active control group, in which teachers implemented an intervention of similar duration and intensity but on a topic unrelated to bullying (climate and environmental issues). This allowed us to control for changes stemming from non-specific treatment factors such as improvement of relationships between teachers due to the project collaboration, which can lead to spurious intervention effects (Lilienfeld et al., 2014). Each group of teachers attended a training day delivered between January and February 2019. Five lessons were then delivered by the class teachers to their students for one hour a week after the training.

The main effects of the two interventions were examined in [Authors, 2021] using a mediation path model and contrasting each intervention group against the active control group. Intervening on moral disengagement decreased students' use of moral disengagement mechanisms, which, in turn, decreased bullying behaviors. On the other hand, intervening on class norms did not yield change in students' perception of their classmates' anti-bullying attitudes but had a direct decreasing effect on bullying. However, implementation was not taken into account in these analyses. One possible reason for the absence of effect of the class norm component on the intended mediator might be that fluctuations in implementation by the teachers concealed the potential effect of the intervention under adequate implementation conditions. How implementation fidelity and quality influenced change in student outcomes for both interventions was the focus of the present paper. Therefore, only data from the two intervention conditions were analyzed. Dosage was not considered because there was minimal variability (only three teachers out of 33 had delivered less than five sessions of intervention) (Haataja et al., 2014). Following some prior measures of quality (Hirschtein et al., 2007), student responsiveness was included in the measure of quality as reflecting the teachers' ability to get students interested and involved in the lessons.

Assessing both fidelity and quality fits recommendations from the implementation literature (e.g., Berkel et al., 2011; Goncy et al., 2015). We expected that higher levels of either fidelity or quality would yield greater change in the outcomes. More specifically, we expected that higher levels of fidelity or quality would lead to a more pronounced reduction in moral disengagement in the moral disengagement intervention group (hypothesis 1a), and a more pronounced increase in perceived anti-bullying class norms in the class norm intervention group (hypothesis 1b). We assumed that fidelity and quality would interact (i.e., multiplicative moderation; Hayes, 2018) in influencing change in the outcomes (hypothesis 2). Finally, we hypothesized that the influence of fidelity and quality would differ between the two interventions (hypothesis 3). Indeed, different anti-bullying components are likely to differentially react to implementation conditions. That is, different levels of fidelity and quality may be needed for different components to be effective. Because intervening on class norms did not yield change in the relevant mediator [Authors, 2021], we specifically expected that this intervention would more likely produce the intended change in the class norm intervention group under optimal implementation conditions (e.g., in the case of both high fidelity and quality).

1.4 Teachers training and interventions

Information on the training and intervention content, outlines, and rationales can be found in [Authors, 2021]. The teachers' training day was divided into two parts: the first was general, provided general knowledge and raised awareness about bullying; the second was specific to the intervention group. This second part covered the related theoretical framework, the use of the intervention material, and an accelerated simulation of the lessons. Each class teacher was provided with detailed lesson plans and required material for the activities. A close follow-up was provided during the intervention period to sustain the implementation process.

The moral disengagement intervention was based on the Bullying Literature Project – Moral Disengagement version (BLP-MD) that uses stories and writing activities to prevent bullying among elementary children (Wang & Goldberg, 2017). We created stories adapted to our sample's age, which represented a variety of bullying situations and illustrated the use of moral disengagement mechanisms by the bullies. Challenging students' use of moral disengagement mechanisms was assumed to restrain their use of the mechanisms, which, in turn, would lead to a decrease in bullying.

The class norms intervention was inspired by the Survey of Bullying at Your School project, which uses students' self-reported data to reveal the discrepancy between the perceived and actual injunctive class norms about bullying, and involves social norm messages displayed on posters (Perkins et al., 2011). We made students discovering their norm misperceptions in an entertaining and instant format suited to their age. Students had to create the social norm messages by themselves to better involve them in the intervention. Correcting the biased perception of the attitudinal class norm regarding bullying was assumed to decrease the students' norm misperception, which, in turn, would lead to a decrease in bullying.

2. Method

2.1 Sampling and Design

A call for recruitment was broadcasted among elementary schools in the context of a research project on bullying prevention. Schools were recruited on a voluntary basis. Detailed information about the power calculation, recruitment, eligibility criteria, randomization process, timeline, and data collection, as well as demographic characteristics of the sample can be found in [Authors, 2021]. Nine schools of various size, geographical location, and socioeconomic index agreed to participate in the project and were randomly assigned to either one of the two intervention conditions or the control group. The original sample comprised 1,216 fourth to sixth graders from 57 classes at baseline. Implementation data was collected in intervention schools only ($N = 6$). Figure S1 presents a flowchart of the recruitment and retention of intervention and control schools and is available in supplementary material.

2.2 Data Collection and Participants

Data were collected in the fall of 2018 and the spring of 2019, with an average of 15 weeks between the two waves. Passive consent forms had been sent to parents prior to data collection. Students did not participate in case of parental objection or when they themselves did not want to participate. Confidential paper questionnaires were filled in during class hours under the supervision of one to two research team members, who were available to answer questions. The survey was introduced to the students as an investigation of “children’s well-being at school” and a few basic rules to ensure confidentiality and free participation were reminded. The whole procedure was approved by the ethic committee of the research institute of the first author. The CONSORT Checklist can be found in [Authors, 2021]. At pre-test, the intervention 1 sample comprised 350 students (49.7% boys, M age = 10.13, SD = 1.06) and the intervention 2 sample comprised 397 students (49.1% boys, M age = 10.28, SD = 0.96).

Implementation data was collected by two trained raters who attended in parallel one specific lesson in each intervention classroom ($N = 34$). We could not observe implementation in two classes from the moral disengagement condition because the teachers were absent on the day of observation. All sessions were rated by the principal researcher, who developed the interventions and trained the teachers, and 85% were also independently rated by the second judge to check the interrater agreement. Once we ensured that acceptable interrater reliability was achieved for both fidelity and quality (see next section), only ratings from the principal researcher were used in the analyses because of her better knowledge and understanding of the interventions and theoretical frameworks.

2.3 Measures

2.3.1 Implementation measures

Fidelity

We developed a checklist of the activities intended in the lesson plan of the specific session that would be observed for both interventions (Hirschstein et al., 2007). Because the two interventions differ in the number of activities, we created 13 items for the moral disengagement and 11 items for the class norm intervention. Items were rated on a 3-point Likert scale, from 0 = *not implemented* to 2 = *fully implemented*. Examples of items for the moral disengagement intervention are “*Teacher explains the logic underlying moral disengagement*” and “*Teacher makes the students looking for the moral disengagement mechanism used by the bully in the story.*” Examples of items for the class norm intervention are “*Teacher reminds to the students the discrepancy between their perceived and actual class norms about bullying behaviors (from the previous lesson)*” and “*Teacher asks the positive leader from each group to read aloud the newly created norm-based message to the whole class.*” Based on the 28 overlapping observations, we computed weighted kappa coefficients to assess the degree of agreement between the two judges (Landis & Koch, 1977). Interrater reliability ranged from substantial ($K = .75$) to almost perfect ($K = .96$) in the moral disengagement intervention, and from moderate ($K = .56$) to almost perfect ($K = 1$) in the class norm intervention (Landis & Koch, 1977). Following recommendations for estimating reliability coefficients with ordinal Likert-type scales, we used the polychoric correlation matrix to compute ordinal Cronbach’s alphas (Gadermann et al., 2012) for the two interventions: $\alpha = .75$ for the moral disengagement and $\alpha = .83$ for the class norm intervention.

Quality of the delivery

We used eight items to assess the four dimensions of the quality scale (Hirschstein et al., 2007): formal aspects of the delivery (clarity, fluency), classroom management (discipline, frame), relational aspects (listening, no judging), and students’ engagement (interest, involvement). A ninth item related to teachers’ ability to adapt the lesson to their students’ needs, that is, small adjustments that do not jeopardize the lesson’s core elements. Items were rated on a 3-point Likert scale, from 0 = *poor* to 2 = *excellent*. Interrater reliability (based on the 28 overlapping observations) ranged from moderate ($K = .53$) to substantial ($K = .71$) (Landis & Koch, 1977). As this measure of quality is supposed to be independent from the specific intervention content, we computed a common ordinal alpha for both interventions, which was $\alpha = .87$.

2.3.2 Students outcomes

Moral Disengagement

We used the French-version of the Moral Disengagement in Bullying Scale (Authors, 2021), adapted from Thornberg and Jungert (2014). The scale taps into the moral

disengagement mechanisms theorized by Bandura (1986): moral justification, euphemistic labeling, advantageous comparison, responsibility displacement, responsibility diffusion, disregarding or distorting the consequences, dehumanization, and blaming the victim. In the present study, we used only 14 out of the 18 items (that is, two items per mechanism) to shorten the questionnaire. This scale 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 research interests (Authors, 2021). Because the intervention aimed at reducing general moral disengagement (rather than specific mechanisms), we used the global index. Students were asked to indicate to which degree they agreed or disagreed with each statement (from 0 = *totally disagree* to 4 = *totally agree*), e.g., “*If people are weird, it is their own fault if they get bullied.*” Ordinal alpha α was = .84 at T1 and .87 at T2. The result of the confirmatory factorial analysis (CFA) showed a good fit for the higher-order structure: $\chi^2(70, N=1094) = 120.7, p < .001$, CFI = .97, RMSEA = .026, 95% CI: .018-.033, and 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 attitudinal anti-bullying class norm

We translated into French eight items from Salmivalli and Voeten (2004) measuring attitudes toward bullying, e.g., “*Bullying may be fun sometimes*” (from 0 = *totally disagree* to 4 = *totally agree*). We followed the original Survey of Bullying at Your School project procedure for measuring the perceived attitudinal class norm, and asked the students to indicate what they thought that most of their classmates would answer, instead of their own opinion (Perkins et al., 2011). Two items were deleted because they were too difficult for the students to understand. The six remaining items had ordinal Cronbach’s alphas of $\alpha = .78$ at T1 and $\alpha = .80$ at T2. As four of the six items were reversed, a method-specific factor was added to the measurement model of the scale to take account of this shared specific variance (Geiser, 2013). The CFA result showed a good fit: $\chi^2(5, N=1091) = 9.5, p = .089$, CFI = .99, RMSEA = .029, 95% CI: .000-.056, and SRMR = .013.

Bullying

A 10-item scale intended for primary school students (Authors, 2020) 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; calling or texting; releasing pictures or videos) bullying behaviors. Students had to indicate how often (from 0 to 4 = *4 times or more*) they had had each behavior toward another student over the last three months e.g., “*I have hidden or damaged another student’s things on purpose.*” 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 demonstrated a good fit: $\chi^2(45, N=1097) = 34.8, p = .21$, CFI = .99, RMSEA = .014, 95% CI: .000-.028, and SRMR = .025. Ordinal Cronbach’s alphas were $\alpha = .85$ (T1) and $\alpha = .83$ (T2) for the global construct, $\alpha = .74$ (T1) and $\alpha = .75$ (T2) for indirect bullying, $\alpha = .81$ (T1) and $\alpha = .81$ (T2)

for direct bullying, and $\alpha = .85$ (T1) and $\alpha = .84$ (T2) for cyber bullying. In our sensitivity analyses, we examined the interventions' effects on the three specific facets of bullying.

2.4 Analytical strategy

We first used the IBM SPSS Statistics version 25 program to compute the descriptive statistics and inter-correlations between fidelity and quality for both intervention conditions. Apart from these preliminary analyses, we used structural equation modeling (SEM) with MPlus version 8.4. The intraclass coefficients (ICC) for bullying was .058 at the class level at the pretest. Because our sample included only six schools (McNeish & Stapleton, 2014), and our research interest lay in implementation discrepancies between teachers rather than between schools, only the class level was taken into account in the analyses. Besides, the amount of variability in bullying is known to be larger between classes than schools (Kärnä et al., 2011), and the ICC at the school level was .029, which is below the conventional threshold of .05 often used in practice (Hedges & Hedberg, 2007). MLR (maximum likelihood robust) is the default estimation with complex survey data in this case, and provides parameter estimates robust to non-normality and non-independence of observations. MLR is also a leading technique for handling missing data, using full-information maximum likelihood (FIML) (Enders, 2010). In the final model, data were missing on all observed dependent variables in 18 cases (2.6%), which Mplus automatically excluded from the analyses. The covariance coverage of data was above 92% for all the indicators.

We used latent change (LC) models to model change between the pre- and post-test of both mediators and bullying. LC models rely on the decomposition 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). Latent difference variables represent interindividual differences in true intraindividual change over time (corrected for random measurement error) and can be used like any other latent variable in a SEM (Geiser, 2013). Because indicators share a specific variance with themselves over time in longitudinal analyses, the assumption of uncorrelated residuals is likely to be violated. To avoid this issue, we allowed the residuals corresponding to the same item to correlate over time (Geiser, 2013).

We first tested measurement invariance across time for all endogenous variables because LC models require at least strong factorial invariance. We progressively constrained factor loadings, intercepts, and residual variances to be equal, and tested the fit of each nested model against the less restrictive one (Geiser, 2013). Specific recommendations have been developed for comparing pairs of nested models in large samples (>300) in the context of measurement invariance because the traditional Chi-square (χ^2) difference test is highly sensitive to sample size and can thus lead to model rejection even when discrepancies are negligible (Chen, 2007). Weak non-invariance is indicated by a CFI difference bigger than .01 in combination with a difference in the RMSEA bigger than .015 or a difference in SRMR bigger than .03. The same applies to strong non-invariance, with the exception that the difference in SRMR should be bigger than .01 (Chen, 2007).

Next, we used a moderated mediation model to examine how fidelity and quality influenced the outcomes of interest. Figure 1a represents this model in the form of a conceptual diagram, in which the effect of the dummy coded variable X (intervention) is influenced by a primary moderator W (fidelity), and this moderation effect of W is itself dependent on a secondary moderator Z (quality). Both moderators were standardized prior to the analysis. Fidelity was standardized in each group separately. To estimate the three-way interaction between fidelity, quality, and the intervention, we introduced the different terms representing the conditional effects and interactions of the three variables, as illustrated in the statistical diagram in Figure 1b. Because the model includes XWZ, the six other terms should be included, even if they are not statistically significant (Hayes, 2018). Importantly, their regression coefficients are conditional effects and interactions – that is to say that they are conditioned on the other variables being zero. For instance, XZ coefficient estimates the conditional interaction between the intervention (X) and quality (Z) when fidelity (W) is equal to zero (Hayes, 2018). We tested the significance of the three-way interaction for the effect of X (intervention) on both mediators, as well as on the direct effect on bullying. A significant three-way interaction term indicates that the effect of X varies across levels of Z, W, and/or the interaction of Z and W (Hayes, 2018). In case of a non-significant three-way interaction term, we removed it and tested whether the two-ways interaction terms were significant. If not, we removed them as well and tested for the significance of the main effects of fidelity and quality.

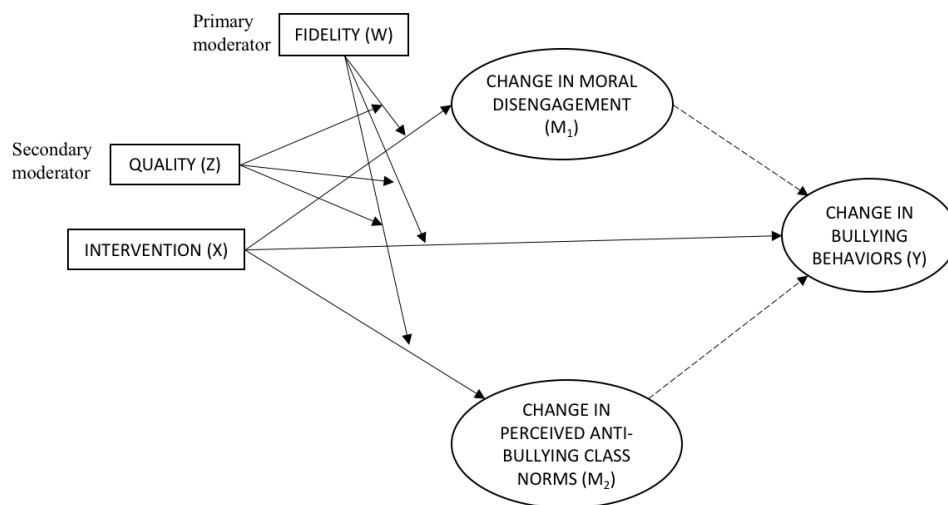
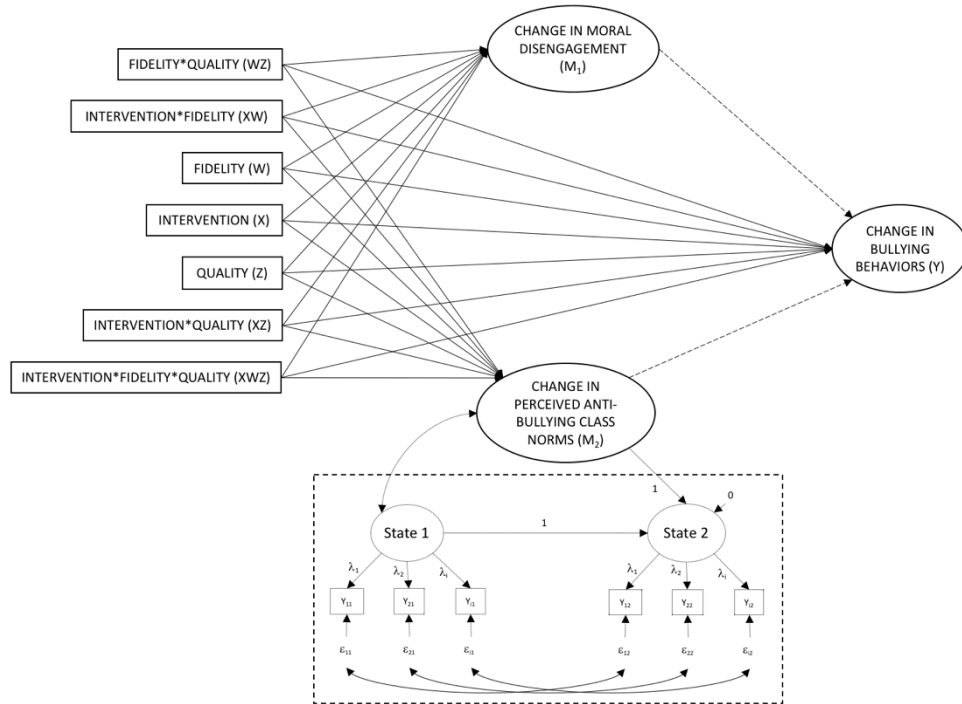


Figure 1a. Conceptual diagram of the moderated mediation model



Note. Each latent variable (indicated by an ellipse) displayed on the statistical diagram implies a latent change model with autocorrelated errors. For the sake of clarity, the LC model of perceived anti-bullying class norms is presented, whereas similar LC models of moral disengagement and bullying are hidden.

Figure 1b. Statistical diagram of the moderated mediation model

Next, we used a pick-a-point approach to probe the significant interactions. We selected the values of the mean (0) and a standard deviation below (−1) and above the mean (+1) for both fidelity and quality. Then, we estimated the conditional effects for the nine possible combinations of fidelity and quality levels, and required an inferential test for each of the combinations. In the case of a significant interaction effect for a mediator, we also estimated the indirect effect of X on bullying flowing through the relevant mediator for all the possible combinations. In our sensitivity analyses, we estimated the indirect effect of X on indirect, direct, and cyber bullying where appropriate.

All effects were contrasted against the other intervention condition and thus reflect average differences between the two groups. For instance, a significant three-way interaction in a mediator means that the average difference between the two intervention groups in this mediator depends on W and/or Z. To examine the influence of the moderators in a group independently, we plotted the absolute expected values of the relevant mediator(s) in the

relevant group at various levels of W (fidelity) and Z (quality). Plots are useful tools to help visualize and interpret the influence of moderators (Hayes, 2018). Finally, to go beyond face validity, we requested inferential tests of the slope differences that could be observed on the plot. To do so, we used either high or low levels of fidelity or quality as the reference for the statistical significance of conditional effects provided in Mplus.

3. Results

Table 2 presents the descriptive statistics and inter-correlations between fidelity and quality for both intervention conditions. Preliminary analyses revealed moderate to large correlations between fidelity and quality. The correlation was only significant in the social norm intervention condition but this is likely to be caused by insufficient statistical power due to the sample size ($N=16$).

Table 2. Descriptive statistics and inter-correlations between implementation features

	<i>M</i>	<i>SD</i>	Min.	Max.	Percentiles			Correlations	
					25	50	75	<i>r</i>	<i>p</i> -value
Moral disengagement									
fidelity	1.14	0.41	0.46	1.85	0.73	1.23	1.44	.44	.085
quality	1.37	0.49	0.56	2.00	0.86	1.56	1.77	—	
Social norm									
fidelity	1.42	0.44	0.40	2.00	1.25	1.50	1.70	.63	<.01
quality	1.31	0.48	0.44	2.00	0.86	1.33	1.78	—	
<i>N</i> = 34									

Strong factorial invariance was reached for all three dependent variables, which allows for a meaningful interpretation of latent mean change over time (Geiser, 2013). Details of the measurement invariance testing can be found in [Authors, 2021]. The model demonstrated acceptable to good model fit: $\chi^2(2076, N=663) = 2958.86$, $p < .001$, CFI = .90, RMSEA = .025, 95% CI: .023-.027, and SRMR = .047. The percentage of variance explained by the full model (as indicated by the R^2) was 23.9 % for the change in moral disengagement, 15.9 % for the change in perceived anti-bullying class norms, and 43.1 % for the change in bullying.

Results of the moderated mediation model indicated that the three-way interaction term was marginally significant for the effect of intervention (X) on change in moral disengagement ($\beta = 0.17$, $p = .06$). Probing the interaction showed that intervening on moral disengagement had both a significant effect on change in moral disengagement and a significant indirect effect on bullying through this mediator for all the combinations of fidelity and quality levels. Table 3 reports the unstandardized regression coefficient and associated p -value for the nine conditional effects on change in moral disengagement and conditional indirect effects on bullying flowing through this mediator.

Table 3. Regression Coefficient and Significance of the Conditional Effects

Combinations	Effect on change in moral disengagement		Indirect effect on bullying	
	β	p -value	β	p -value
low fidelity – low quality	–0.42	<.01	–0.14	.02
low fidelity – medium quality	–0.70	<.01	–0.25	.02
low fidelity – high quality	–0.99	<.01	–0.36	.02
medium fidelity – low quality	–0.51	<.01	–0.16	<.01
medium fidelity – medium quality	–0.62	<.01	–0.22	.01
medium fidelity – high quality	–0.74	<.01	–0.28	.01
high fidelity – low quality	–0.60	.02	–0.18	<.01
high fidelity – medium quality	–0.54	<.01	–0.19	.01
high fidelity – high quality	–0.49	<.01	–0.20	<.01

$N=663$

A closer look at the plot of absolute expected values of change in moral disengagement revealed that intervening on moral disengagement clearly decreased students' moral disengagement, regardless of the fidelity and quality levels (see Figure 2). Visual interpretation of the plot suggests that when fidelity was low, the magnitude of the decrease in moral disengagement was conditional upon the level of quality. Greater quality tended to amplify the reduction. An inferential test indicated that this effect was marginally significant, and that an increase of one standard deviation in quality yielded a decrease of $\beta = -0.11$ in moral disengagement ($p = .05$). As soon as fidelity reached a medium level, however, the effect of quality was no longer significant. On the other hand, when quality was low, it seems that the magnitude of the decrease in moral disengagement varied as a function of fidelity. An inferential test indicated that this effect was significant, and that an increase of one standard deviation in fidelity yielded a decrease of $\beta = -0.11$ in moral disengagement ($p < .01$). Again, as soon as quality reached a medium level, the effect of fidelity was no longer significant.

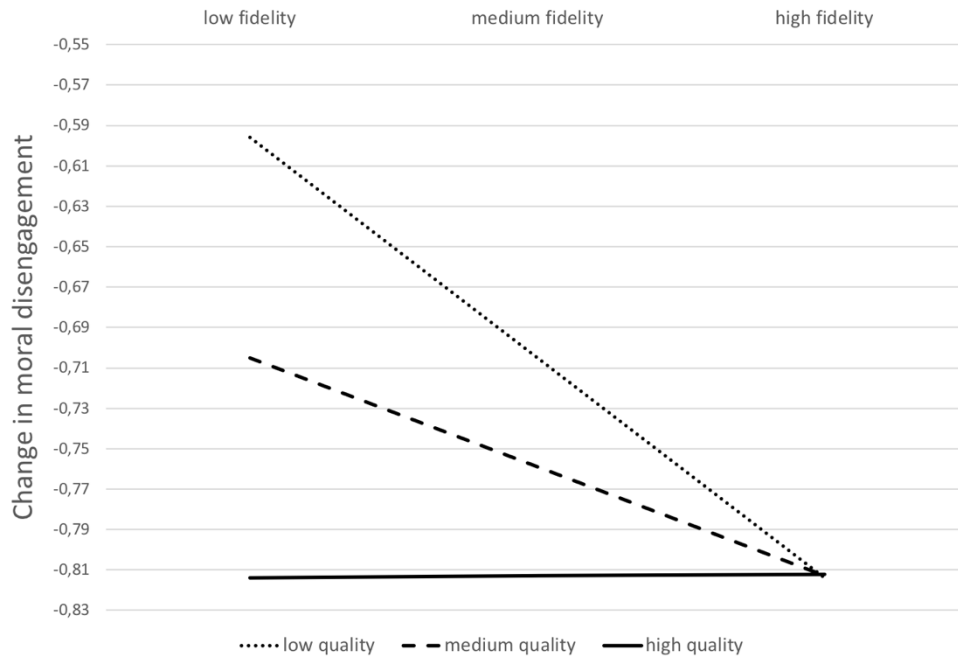


Figure 2. Expected values of change in moral disengagement following the moral disengagement intervention

By contrast, there was no significant main or interaction effects of fidelity or quality on the effect of intervention (X) on change in perceived anti-bullying class norm. There was no main or significant interaction for the direct effect of intervention (X) on bullying either.

3.1 Sensitivity Analyses

Sensitivity analyses revealed that intervening on moral disengagement had a significant indirect effect on indirect and direct (but not cyber) bullying through change in moral disengagement for all the combinations of fidelity and quality levels. When a high fidelity was combined to a low quality, the effect was only marginally significant for direct bullying ($p = .07$). Table 4 presents the regression coefficient and associated p -value for the nine conditional indirect effects on indirect, direct, and cyber bullying flowing through change in moral disengagement.

Table 4. Regression Coefficient and Significance of the Conditional Indirect Effects

Combinations	Indirect bullying		Direct bullying		Cyber bullying	
	β	<i>p</i> -value	β	<i>p</i> -value	β	<i>p</i> -value
low fidelity – low quality	–0.13	.02	–0.09	.03	–0.08	.23
low fidelity – medium quality	–0.21	.02	–0.16	.03	–0.14	.23
low fidelity – high quality	–0.30	.02	–0.22	.04	–0.20	.23
medium fidelity – low quality	–0.15	.01	–0.11	.03	–0.10	.23
medium fidelity – medium quality	–0.19	<.01	–0.14	.02	–0.12	.22
medium fidelity – high quality	–0.22	<.01	–0.16	.02	–0.15	.23
high fidelity – low quality	–0.18	.04	–0.13	.07	–0.12	.26
high fidelity – medium quality	–0.17	<.01	–0.12	.03	–0.11	.24
high fidelity – high quality	–0.15	.02	–0.11	.05	–0.10	.27

N=663

4. Discussion

Limited or inconsistent results of anti-bullying programs are likely to be explained by variations in how these programs are implemented (Salmivalli et al., 2013). Assessing implementation is critical to interpret results of anti-bullying interventions (Low et al., 2014). The importance of implementation fidelity and quality, in particular, have been emphasized by the literature in bullying prevention, especially when interventions are delivered by teachers (Berkel et al., 2011; Durlak, 2010; Goncey et al., 2015). Nevertheless, in the domain of bullying prevention, few evaluation studies have associated implementation and program outcomes (Haataja et al., 2014). In addition, findings from our literature review revealed a high heterogeneity of measures, methods, and results among these studies, highlighting the need for further investigation. This paper aimed to examine the influence of fidelity and quality on the outcomes of two promising anti-bullying components delivered by elementary teachers within separate interventions in a randomized controlled trial. Both intervening on moral disengagement and class norms to prevent bullying rely on relevant theoretical frameworks and sound empirical foundations (e.g., Almeida et al., 2009; Menesini et al., 2015). The main effects of these two components have been assessed by contrasting them against a control group in a previous study, which, however, did not take implementation into account (Authors, 2021). Briefly summarized, findings of this previous study revealed that intervening on moral disengagement reduced students' moral disengagement, which, in turn, reduced self-reported bullying. By contrast, intervening on class norms did not increase student perceived anti-bullying class norms but had a direct decreasing effect on self-reported bullying.

We hypothesized that greater fidelity or quality would amplify the expected change in each mediator in the relevant intervention group (hypothesis 1a and 1b) and that fidelity and quality would interact in influencing the outcomes (hypothesis 2). Our results partially supported these first two hypotheses for the moral disengagement intervention. In this group, we could see that a high quality was related to a greater reduction in moral disengagement, regardless of the fidelity. When fidelity was low, a better quality significantly predicted a

greater reduction in moral disengagement. Thus, how well teachers delivered the moral disengagement intervention to their students played an important role, especially when they implemented the lessons less rigorously. This encompassed the clarity and fluency of the delivery, their management of the classroom, and their ability to get their students interested and involved in the lesson, as well as to adapt the lesson unfolding to their needs. When the quality was low, however, a higher level of fidelity significantly increased the reduction in moral disengagement, and seemed to compensate the lack of quality in a way. This impact of fidelity faded as soon as the quality level increased, suggesting that when quality was medium or higher, fidelity no longer significantly influenced the change in moral disengagement.

Nevertheless, it is important to note that regardless of the fidelity and quality levels, students who received the moral disengagement intervention systematically decreased in their use of moral disengagement mechanisms. In addition, the average difference in moral disengagement change as well as the indirect effect on bullying (flowing through this mediator) between the students who did and did not receive the moral disengagement intervention was significant for all combinations of fidelity and quality levels. More precisely, the indirect effect was significant for indirect and direct but not cyber bullying, as indicated by our sensitivity analyses. For the perceived anti-bullying class norm, there were no significant effects, meaning that the average difference in perceived class norm change did not vary at any levels of fidelity and quality.

Finally, we assumed that the impact of the implementation variables would differ between the two interventions. That is, we postulated that the influence of the implementation variables on moral disengagement change in the moral disengagement intervention group might be different than that on perceived anti-bullying class norms in the class norm intervention group (hypothesis 3). This third hypothesis is supported by our results in a way, as fidelity and quality interacted in influencing moral disengagement change in the corresponding intervention group whereas this was not the case for class norm change.

Contrary to our expectations, however, was the absence of significant average differences in class norm change between the groups at any combined levels of fidelity and quality. In other words, even under ideal conditions of implementation, that is, when teachers implemented the class norm intervention with high fidelity and quality, there was no significant average difference in perceived class norm change between students who received the relevant intervention or not. This result is important as it rules out implementation as potential explanation for the absence of effect on the intended mediator.

An alternative explanation might be that while our intervention actually impacted bullying through the expected mediating path, our mediator measure failed to capture students' perception of their classmates' attitudes accurately. For students to report what they thought that most of their classmates would answer, instead of their own opinion, might have been too difficult for students of the age of our sample (Grades 4-6) [Authors, 2021]. This explanation, however, does not seem very likely, as the same measurement procedure has successfully been used in students as young as seventh graders (Dillon & Lochman, 2019).

It is also possible that the intervention did not impact students' perception of their classmates' anti-bullying attitudes as hypothesized. Instead, it may have reduced bullying through another mediator. For instance, the class norm component may have elicited another type of norm such as popularity norm (popular students' perception of the class attitudinal norm), friend norm (students' perception of their friends' rather than all classmates' attitudes), or teacher norm (students' perception of their teacher's attitudes) [Authors, 2021].

4.1 Strengths and Limitations

Strengths of this study are the relatively large sample with high retention rates, the use of latent change models, the control for the nested structure of the data, and the use of observational data for assessing the fidelity and quality. In addition, the use of a multiplicative moderation model allows for considerably more flexibility than an additive multiple moderation, which constraints the interaction between X and W to be independent of Z (Hayes, 2018).

Despite its strengths, this study also has some limitations. First, implementation data was collected during only one lesson for each teacher. In a few studies (e.g., Hirschstein et al., 2007), implementation data has been collected for each lesson and then averaged across the lessons. However, such a procedure would have been too time-consuming with respect to the available resources for this study. An alternative could have been to combine teacher self-reports with observational ratings to ensure criterion validity (e.g., Biggs et al., 2008; Renshaw & Jimerson, 2012).

Second, although observational data are generally seen as more reliable than self-reports (Durlak & Dupre, 2008; Hirschstein et al., 2007), teachers can be resistant to it (Dusenbury et al., 2004). They might not feel comfortable delivering the lesson in front of raters (who sometimes are also the program trainers). Although we were careful to build a relationship of trust and non-judgment with the teachers, being observed while delivering the lesson was nevertheless stressful for some of them, which might have impaired the delivery to some extent.

Third, a limitation faced by many studies assessing implementation is the difficulty to use data from a control group in the analysis. For instance, it would not have been possible to use a moderation model including the control group in the present study because there was no implementation data for this group – thus any data from the control group would have been automatically excluded from the analysis. Some studies have created categorical implementation variables (e.g., low, moderate, high) to be able to contrast each level with the control group (e.g., Matischek-Jauk et al., 2018; Salmivalli et al., 2005). However, cut-offs used for categorization are always arbitrary and have less statistical power than continuous variables (Durlak & Dupre, 2008). We suggest that using an active control group (in which an intervention of same duration and intensity but unrelated to bullying is delivered, as in our design) might be an interesting direction to solve this issue. Although being unrelated to bullying, implementation of the 'control intervention' could be assessed as well, thus allowing a relevant comparison in a sophisticated statistical design using continuous

moderators. As far as we know, this has never been done before in anti-bullying intervention studies linking implementation to student outcomes, presumably because of the high cost in time and human resources.

Fourth, because our measure of the perceived attitudinal class norm has not been previously validated in elementary students, we cannot guarantee its validity in our sample (Grades 4-6). Because of this uncertainty, we cannot definitively exclude that change in bullying was actually mediated by the intended mediator. However, as a similar measure has been successfully used in seventh graders (Dillon & Lochman, 2019), we tend to favor other possible explanations for our findings such as alternative mediating paths.

Finally, our adaptation of the class norm intervention implied a greater involvement of the teachers compared with the original study (Perkins et al., 2011). Whereas teachers have a crucial role in shaping student relationships and social climate (e.g., Colpin et al., 2021; Demol et al., 2020; 2021; Kallestad & Olweus, 2003), this may also have accentuated a bias of social desirability in students when reporting their attitudes toward bullying. Based on our observations during the implementation data collection, not all students were fully convinced by their classmates' answers during the intervention. Despite the guaranteed anonymity, students were sometimes skeptical, and suggested that some of their classmates may have been dishonest in reporting their anti-bullying attitudes. Although limited, this unexpected phenomenon may have jeopardized the impact of the message conveyed by the intervention to some extent.

4.2 Implications and future directions

This study highlights the relevance of both assessing the implementation of anti-bullying interventions or programs and analyzing how it relates to the outcomes. Assessing implementation is important not only for estimating the intervention feasibility or providing additional support if needed along the way (Haataja et al., 2014). It also offers unique information about the effects of interventions at different levels of implementation. Besides, it is likely that different anti-bullying components require different levels of implementation fidelity and quality. It is thus crucial to determine the specific implementation conditions under which each component is effective. Implementation of anti-bullying interventions by teachers in natural contexts inevitably fluctuates (Goncy et al., 2015). For a given intervention to be effective even under suboptimal implementation conditions represents a strong asset. At least, ensuring that the intervention does not produce any iatrogenic effect in that case should be a necessary condition for any intervention to be further disseminated. Besides, clarifying that limited or null impact of an anti-bullying intervention is not due to variations in its implementation is important to extend our understanding of how anti-bullying interventions work and our knowledge of the effective components.

In terms of cost-effectiveness ratio, our findings provide strong support for the integration of a moral disengagement component into anti-bullying programs as it demonstrated that it reduced moral disengagement and had an indirect effect on bullying even when poorly implemented. The fact that a higher fidelity seemed to weaken the negative effect of a poorer

quality is also a strength. Provided that teachers sufficiently follow the intervention instructions, the intervention reduces moral disengagement, even if teachers lack other skills such as classroom management and relational skills. Our findings also support the use of bibliotherapy (i.e., the use of reading materials to assist personal problem solving) to convey anti-bullying interventions in elementary students. The use of children stories may have facilitated the implementation of the moral disengagement intervention by the teachers because of their familiarity with this approach and its easiness to be integrated into the regular instruction curriculum (Wang & Goldberg, 2017).

As the class norm component had no effect on students' perceived attitudinal class norm even under optimal conditions of implementation, it rules out the possibility that the absence of effect on the intended mediator be caused by fluctuations in implementation by the teachers. Future studies should therefore test several alternative mediators in order to clarify which of them play(s) a role in mediating change in bullying. In particular, we suggest that different types of norm such as popularity norm, friend norm, and teacher norm may be relevant targets.

As aforementioned, we suggest that future studies of anti-bullying interventions develop parallel interventions (unrelated to bullying) to be used in an active control group. This way, implementation data could be collected in the control group as well, which would ensure meaningful comparison of the influence of implementation variables on student outcomes (e.g., bullying) between the groups. Albeit entailing additional costs (e.g., development of the alternative intervention, training of the control teachers, implementation assessment in the control classes), we argue that such a research design could be a promising direction for evaluation studies of anti-bullying interventions. Regarding quality, in particular, this would enable researchers to distinguish the part of the effects of quality that is specific to the intervention from the part that is not (i.e., that occurs whether teachers implement the anti-bullying intervention or not). That is, to determine what are the beneficial effects of teachers' more personal qualities and transversal skills that are reflected in the quality measure on students' outcomes.

This study investigated the combined influence of implementation fidelity and quality on the effects of two different anti-bullying components. Results highlighted several assets of the moral disengagement component, which seemed effective in reducing students moral disengagement (and, indirectly, bullying) even under suboptimal implementation conditions. A high quality appeared to be beneficial in any case. However, when the quality was low, then fidelity mattered. The class norm component had no effect of the hypothesized mediator regardless of the fidelity and quality levels, thus ruling out implementation as possible reason for this absence of effect. Rather, our results suggest that alternative mediating paths should be tested to explain how this intervention reduces students' bullying behaviors. We advocate for the systematic assessment of implementation in evaluation studies to expand our understanding of how anti-bullying program work and knowledge of their effective components.

Appendix A.

The starting keywords were ‘bullying’, ‘intervention’ or ‘prevention’ or ‘program’, and ‘implementation’. As some studies may only refer to specific features rather than general implementation, we added ‘dosage’ or ‘exposure’ or ‘delivery’ or ‘responsiveness’ or ‘fidelity’, as well as four possible synonyms for fidelity: ‘adherence’ or ‘compliance’ or ‘integrity’ or ‘replication’. As ‘quality’ and ‘competence’ are very generic terms, they were not included in the equation. We chose to add the keyword ‘school’ to filter articles related to workplace and prison bullying out. We used an asterisk to include any form (e.g., noun, verb, gerund, or plural) of the key words, using for instance ‘interven*’ and ‘prevent*’ instead of intervention and prevention. Our search terms had to be present in the title, abstract, or keywords. We limited our results to peer-reviewed articles published in English in a scientific journal. Books, dissertations, book chapters, or other formats were not considered. All articles published until May 2021 were examined, with no restrictions linked to the geographical region or sociocultural origin of the sample. We selected articles that met the inclusion criteria based on title and abstract. In case of doubt, we browsed the full-text. Articles were excluded when they did not evaluate an intervention or program aiming at reducing bullying in school-aged children or adolescents. They had to target non-clinical students, as opposed to students with disabilities or special needs (e.g., students with autism). Studies were excluded if they did not assessed at least one implementation feature or if they did not use this data in their main analyses. The search resulted in 723 hits in Scopus, 363 in PsycInfo, and 166 in Eric. After removing duplicates in the different databases (362 articles) there were 881 articles left. Out of these articles, 844 articles did not meet the inclusion criteria. An additional 12 articles were excluded after reading the content of the publication, when it appeared that they did not, in fact, assessed implementation, or only use the implementation index to ensure that a certain level of implementation was reached, but did not associate implementation data with program outcomes. Furthermore, an article used implementation data as a dependent variable and two articles removed the schools with a low level of implementation from the analyses, and were therefore excluded. The final selection contained 22 articles.

Supplementary Material

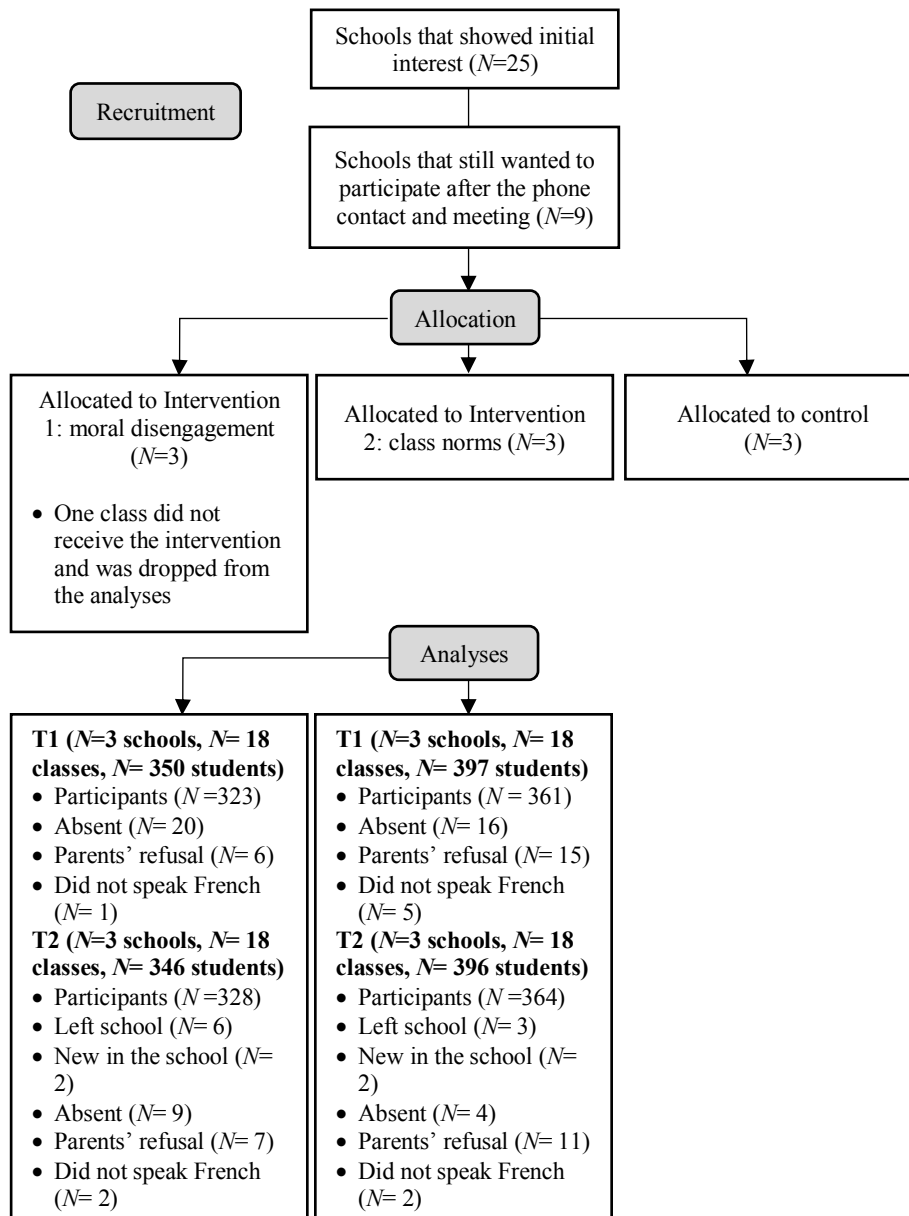


Figure S1. Flowchart of recruitment and retention of intervention and control schools

Chapter 5

General discussion

General discussion

1. Introduction

Many anti-bullying interventions have emerged in the last few decades, driven by the prevalence, pervasiveness and detrimental consequences of school bullying for the students involved. Meta-analyses and systematic reviews have revealed an overall mixed picture in terms of results (Bradshaw, 2015; Gaffney et al., 2019). While the most favorable meta-analyses suggest a 20% reduction in bullying on average (Farrington & Ttofi, 2009; Gaffney et al., 2019), it should not be forgotten that some programs have not proven effective, and many others have never been evaluated (Menesini & Salmivalli, 2017). Three methodological gaps might explain these mixed and/or limited results, and form the “black box” of anti-bullying program evaluations. First, most evaluation studies do not assess the hypothesized mediators of the intervention (Saarento et al., 2015). Second, program components are usually assessed as a whole rather than within separate interventions to determine their unique contribution to the outcomes (Menesini & Salmivalli, 2017). Third, very few evaluation studies associate the outcomes with how the program was implemented (Haataja et al., 2014). These gaps limit our understanding of how anti-bullying programs work and thus our knowledge of the effective components.

These mixed and/or limited results of anti-bullying interventions also call for innovation in program development and the study of new directions to prevent bullying, based on prior research and theory (Eisner & Malti, 2012). While most students consider bullying immoral, many display some approval of bullying publicly, and few of them actually stand up for the victims (Hymel et al., 2010). Two relevant frameworks aim to explain this paradox between students’ actual behavior and private attitudes toward bullying. Both moral disengagement and class norms have been suggested as promising directions for prevention but have rarely been targeted as components in anti-bullying programs yet (e.g., Almeida et al., 2009; Menesini et al., 2015).

The research described in this dissertation aimed to take a first step in opening the black box of anti-bullying programs, and to test these two innovative and promising anti-bullying components: intervening on moral disengagement and class norms. A necessary preliminary step, however, to determining the effects of intervening on moral disengagement as an anti-bullying component was to validate a measurement tool that matched the theoretical model. As the moral disengagement mechanisms identified by Bandura are conceptually distinct, yet are supposed to operate together (Thornberg & Jungert, 2014), a second-order factorial structure encompassing the mechanisms as first-order factors is most suitable.

In this closing chapter, we sum up the main findings from the three papers that this dissertation comprises and discuss them in light of the current relevant literature. We discuss the strengths and limits of this thesis, as well as its practical implications. Finally, we suggest directions for future research.

2. Overview of the main findings

To answer our research questions and address the aforementioned gaps, we designed a three-armed randomized controlled trial (RCT) with a moral disengagement intervention condition, a class norm intervention condition, and a control group in a sample of about 1,200 elementary school students (Grades 4-6). Data were collected in the fall of 2018 and the spring of 2019. Each group of teachers attended a training day delivered in January or February 2019. The five intervention sessions were then delivered by the class teachers to their students for one hour a week within a month after the training. Implementation data was collected by two independent trained observers who attended in parallel one specific lesson in each intervention classroom.

2.1 Researchers can validly assess either the mechanisms or the global dimension of moral disengagement using the Fr-MDBS

In the second chapter of this dissertation, we presented the validation of the French version of the Moral Disengagement in Bullying Scale. Valid measurement tools are critical to health, social, and behavioral sciences (Boateng et al., 2018). When assessing anti-bullying components within separate interventions, it is even more important to have a valid measure of the variable targeted by the specific component (i.e., the mediator). Our results indicated that the factorial structure that matched the conceptual model of moral disengagement was most suitable. The Fr-MDBS thus captures the multidimensional concept of moral disengagement accurately, which allowed us to validly examine the overarching dimension of moral disengagement in this thesis. Researchers may validly examine either the mechanisms or the global dimension of moral disengagement using this instrument.

2.2 It is worthwhile to include a moral disengagement component in anti-bullying interventions

This dissertation indicated that intervening on moral disengagement decreased students' use of moral disengagement mechanisms, which, in turn, impacted student outcomes in the expected direction. Through moral disengagement change, the intervention significantly decreased indirect and direct bullying, as well as (marginally) cyberbullying. The intervention also impacted bystander behaviors through moral disengagement change, increasing defending and decreasing outsider behaviors. Our research design also allowed an alternative mediating path flowing through another relevant mediator (perceived class norm). As this path was not significant, we can be reasonably confident that our intervention successfully targeted moral disengagement and did not incidentally elicit another anti-bullying component.

Even if most of the effect of an intervention is explained by the mediating path, the size of an indirect effect cannot exceed the size of the total effect. Thus where program effects are small, which is usually the case in bullying prevention, the size of the indirect effect is inevitably small as well (Saarento et al., 2015). Indices of effect size are not available yet for

multiple-mediator models (Preacher & Kelley, 2011; Saarento et al., 2015). In terms of practical significance, however, it is noticeable that the standardized effect of moral disengagement on bullying is larger than that of mediators such as anti-bullying attitudes, perceived proportion of defenders, and teacher's perceived attitude toward bullying at the student level in KiVa (Saarento et al., 2015).

Theories of change that postulate the trajectory from program components to relevant predictors and then to desired outcomes are rarely specified or tested (Gottfredson et al., 2015). Our research design specifically addressed this issue and allowed us to test the theory underlying our interventions accurately, which is critical for improving bullying prevention (Menesini & Salmivalli, 2017; Saarento et al., 2015). Our findings support the hypothesis that when students cannot rely on moral disengagement mechanisms, they are no longer able to deactivate internal controls when there is a moral transgression, and thus experience the unpleasant self-sanctions that they were trying to avoid. These self-sanctions, in turn, tend to dissuade them from behaving in an immoral way.

In addition, testing the conceptual theories behind interventions not only enhances our understanding of behavioral change but also increases our knowledge of the psychosocial factors involved in bullying (Eisner & Malti, 2012). Because our intervention targeted only moral disengagement, our research design, in particular, provided sound evidence supporting the involvement of moral disengagement in bullying and bystander behaviors. Overall, our results provide strong support for the integration of a moral disengagement component into anti-bullying programs.

2.3 Intervening on the class norms decreased bullying despite no significant effect on the expected mediator

In this dissertation, we found that intervening on the class norms did not have the expected effect on students' perception of their classmates' attitudes toward bullying. Social norms theory postulates that providing people with correct information about the actual group norm rectifies their misperception of the norm. Although this has been proven effective in various norm-based interventions (Miller & Prentice, 2016), including anti-bullying ones (Perkins et al., 2011; Dillon & Lochman, 2019), this was not the case in our study. However, our intervention did have a significant direct effect on direct and indirect (but not cyber) bullying. That is, bullying significantly decreased in students who received the class norm intervention, but this change did not flow through the expected mediator. Several possible reasons may explain this unexpected pattern of results.

2.3.1 *Weakening the power of the norm versus mobilizing peer influence*

Sometimes, behavior change occurs without change in norm perception after a norm-based intervention (Miller & Prentice, 2016). A possible explanation is that despite the absence of perceived norm change, the intervention nevertheless weakens the power of the norm (Prentice, 2008). Norm-based interventions can impact behaviors in two different ways:

they can either mobilize or weaken peer influence processes. That is, norm-based interventions may aim to encourage people to conform to the (actual) norm or reduce the perceived peer pressure¹ to conform to the (misperceived) norm (Prentice, 2008). In a drinking norm-based intervention, discussing the pluralistic ignorance phenomenon with students yielded a 40% drop in drinking behavior without any change in perception of the average attitudinal norm (relative to a control group). Overall, the findings of that study suggested a reduction in perceived social pressure rather than a cognitive reconstruction of the perceived norm (Prentice, 2008). In another study (Schroeder & Prentice, 1998), despite an actual drinking behavior change a few months after the intervention, students did not report less pluralistic ignorance about the injunctive norm than the control group. Moreover, they did not remember participating in the intervention in the first place (Prentice, 2008). It is thus possible that students in our intervention took on board the fact that not everyone approved of bullying within their class, even though they did not reconstruct their estimate of the attitudinal norm – or perhaps forgot about the information provided later on. This weakening of perceived peer support toward bullying might nevertheless have been enough to initiate a pattern of behavior that became self-sustaining (Prentice, 2008).

A direct effect of our class norm intervention was not found for outsider and defender behaviors, however. The possible weakening of the norm's power was therefore apparently not sufficient to encourage bystanders to stand up for the victims. Bullies, on the other hand, may have been more responsive to the information that there was not much approval for bullying among their classmates, thus adapting their behavior accordingly. This explanation is in line with the idea that the power of the group norm is stronger for students who have more fear of negative evaluations from peers (e.g., Schroeder & Prentice, 1998; Miller & Prentice, 2016). As bullies are particularly driven by dominance goals (Olthof et al., 2011; Pellegrini & Long, 2002; Sijtsema et al. 2009), it is likely that they would not want to risk their social status and may be more prone to adjust their behavior to align with the peer norm. Standing up to the bullies, by contrast, might represent too high a risk for bystanders, who may fear for their own safety and status (e.g., becoming the next victim) (Juvonen & Galván, 2008). Some studies have highlighted that standing up to bullies requires both a strong sense of self-efficacy for defending and a high status in the group (Gini et al., 2008; Pöyhönen et al., 2010). A high social status, in particular, may protect defenders from possible negative consequences of their laudable behavior (Pöyhönen et al., 2010).

2.3.2 Different types of social norm intervention may operate differently

In line with this first explanation, we can also imagine that the type of social norm intervention matters. Our intervention comprised elements from social norms marketing

¹ It should be noted that we use the term “peer pressure” in its broadest sense. That is, we do not restrict its meaning to peers’ *active* attempts to make a student change his or her behaviors, attitudes, or other characteristics. Instead, we also include *passive* and implicit forms of peer influence in which students tend to conform to their peers’ behaviors without being urged to do so. This feeling of internal pressure to conform is sometimes referred to as “imitation” as opposed to (active) “peer pressure” (Veenstra & Laninga-Wijnen, 2022).

(SNM), personalized normative feedback (PNF), and focus group discussion (FGD) interventions. As in the original SNM-type intervention (Perkins et al., 2011), we revealed correct information about the prevalence of desirable attitudes to the students through the intervention. However, the written norm-based communication posters displayed general messages about the actual class norm (e.g., *“In our class, we like to include everyone.”*) instead of more precise numerical information (e.g., *“94% of [name of the school] middle school students ...”*, *“Most [name of the school] middle school students (3 out of 4) ...”*; Perkins et al., 2011) – which was judged less easy to understand for elementary students. We also provided students with information about the gap between the perceived and actual injunctive norm, as in a PNF intervention, but at the class rather than at the individual level. Finally, this feedback was given by the class teachers, who led (short) interactive class discussions about the misperception finding, as in a FGD intervention (Miller & Prentice, 2016).

In a recent study using a PNF intervention (Dillon & Lochman, 2019), a significant difference in perceived injunctive class norm was found at post-test between the experimental (PNF) condition and both the ‘no normative feedback’ and the ‘personalized normative feedback unrelated to bullying’ control conditions. By contrast, there was no significant difference between the SNM-type general normative feedback and the two aforementioned control conditions. Although there was no significant difference between the PNF and SNM conditions either, taken together these results suggest that PNF may be more effective than SNM in changing students’ estimate of the injunctive class norm. Perhaps this personalized component elicits the cognitive process of reconstructing the perceived norm in students more than other types of interventions, notably because they feel more involved.

Similarly, in the two examples of drinking norm intervention presented in the previous section, the behavioral change without perceived norm change followed a FGD intervention, which involves facilitators leading interactive group discussions about the misperception between the perceived and actual norm (Miller & Prentice, 2016). It might be that FGD-type norm interventions are more effective in weakening the power of the norm because of the collective component and the greater involvement of the group. By contrast, FGD interventions may be less effective in eliciting a reconstruction of the perceived norm in students, as the focus is on the misperception phenomenon rather than the correct estimate of the norm.

Our intervention comprised mainly characteristics of PNF- and FGD-type interventions. The characteristics of SNM-type intervention were the least well represented, as revealing correct numerical information about the norm was a one-shot action. Students were repeatedly exposed to the norm-based messages displayed in the corridors, which, however, were quite general and did not contain precise estimates of the norm. This could explain why our intervention led to behavior change without any change in students’ estimate of the injunctive norm.

2.3.3 A measurement issue?

An alternative explanation might be that our measure of class norms failed to capture the intended mediator. Asking students to report to what extent they think that their classmates would (dis)agree with a list of statements about bullying attitudes is a classic way of assessing the perceived injunctive norm (e.g., Prentice & Miller, 1993; Shelton & Richeson, 2005), and has successfully been used with students as young as sixth and seventh graders (Dillon & Lochman, 2019; Perkins et al., 2011). However, it may have been too difficult a cognitive exercise for students of the age of our sample (Grades 4-6). Our intervention may thus have actually had the expected effect on students' perception of the injunctive class norm. The increase in perceived anti-bullying class norms may, in turn, have decreased bullying behaviors, but our measurement tool failed to capture this mediation path.

2.3.4 Popularity norm versus class norm

Another explanatory hypothesis is related to the type of norm assessed as the mediator. It has been argued that popular peers rather than the whole peer group set the norm (Dijkstra et al., 2008; Henry et al., 2000). Popular students' behavior is very visible and others may be tempted to copy it, whether in an attempt to gain social status, belong to the group, or conform out of fear of social rejection (Veenstra & Lodder, 2022). The popularity norm is usually operationalized as the within-class correlation between perceived popularity and a given behavior or as the average behavior of popular students within the class (Veenstra & Lodder, 2022). Evidence that students conform to the popularity norm has been found in diverse domains (e.g., health-risk behavior, alcohol, pro-sociality) (Veenstra & Lodder, 2022). The bullying popularity norm has notably been found to influence friendship dynamics (Laninga-Wijnen et al., 2017), as well as defending behaviors (Garandeau et al., 2019; Peets et al., 2015).

If we transpose this logic to attitudes, it might be that popular students' perception of the class injunctive norm toward bullying mediated bullying change. This is consistent with our previous assumption that our intervention had more impact on bullies, presumably because of a greater fear of losing their dominant position (Olthof et al., 2011; Pellegrini & Long, 2002; Sijtsema et al. 2009). Our intervention may thus have primarily influenced bullies' perceptions, which, in turn, led them to adjust their bullying behaviors. A remaining question is whether this relatively short-term effect would have translated into a change in bystanders' perception of the injunctive norm – and subsequently, in bystanders' behaviors in the long term.

2.3.5 Friend norm versus class norm

The subgroup of friends may also be seen as relevant norm-setters for students. Individuals with whom students preferably spend their free time (e.g., recess or after-school time) might be more influential than the whole class group (Paluck & Shepherd, 2012). Moreover, people are more influenced by individuals with whom they have an affinity or feel connection (Hogg, 2010). Individuals tend to adopt the behaviors of others in their network through influence or socialization processes, thus becoming more similar in behavior over

time (Huitsing et al., 2019). The influence of deviant peer affiliation has been observed regarding diverse health-related or problem behaviors such as substance abuse (e.g., Walden et al., 2004), problematic internet use (e.g., Li et al., 2013) or delinquency (e.g., Keijsers et al., 2012). Deviant peer affiliation has been found to be both concurrently associated with and predictive of bullying perpetration over time (e.g., Cho, 2018; Hong et al., 2017; Wang et al., 2020). Taken together, these results highlight that social affiliations may affect bullying-related behaviors, and this influence is likely to operate through perceived norms.

When examining the influence of the perceived injunctive norm on defending behavior, some authors have restricted the reference group to “classmates who are important to [the student]” (e.g., Kollerová et al., 2018). Such a focus on significant peers rather than peers or classmates in general as norm-setters was also present in an intervention study that used social referents as ‘seeds’ for norm-changing (Paluck & Shepherd, 2012). These authors specifically argued that personally initiated social interactions, as measured by spending time together, should be more important to changing perceptions of collective norms than interactions created by relatively arbitrary institutional channels [i.e., classrooms] (Paluck et Shepherd, 2012, p. 910). However, some norm-based studies in domains other than bullying have suggested that intervention-induced normative beliefs spilled over from the targeted referent group to more proximal groups such as friends (e.g., Nolan, 2011). Overall, it is still unclear whether the specific group of friends is more influential than the class group in anti-bullying norm-based interventions.

2.3.6 Teacher norm versus class norm

Teachers’ influence on student behavior and peer relationships has been widely recognized and referred to as the ‘invisible hand’ of the teacher (Bierman, 2011; Farmer et al., 2011). Students are therefore not the only ones to set the class norms – teachers contribute to shaping social norms as well (Veenstra et al., 2014). Teachers’ attitudes regarding bullying (e.g., normative beliefs) have been found to influence bullying prevalence and students’ willingness to defend victims, notably through the way teachers address bullying incidents (e.g., Colpin et al., 2021; Demol et al., 2020; Hektner & Swenson, 2012; Kochenderfer-Ladd & Pelletier, 2008; Troop-Gordon & Ladd, 2010). Students infer teachers’ attitudes from the way they respond to bullying (Demol et al., 2020; 2021), which serves as a socialization experience for students (Yoon & Bauman, 2014). Students’ perception of their teacher’s attitude toward bullying (i.e., perceived teacher injunctive norm) was found to predict victimization in third- to fifth-graders in one study (Saarento et al., 2013). It may thus be that students’ perception of their class teacher’s attitudes (rather than classmates’ attitudes) toward bullying mediated change in students’ bullying behaviors in our intervention. However, another study found no effect of perceived teacher injunctive norm beyond the effects of perceived peer injunctive norms in seventh-graders (toward defending; Kollerová et al., 2018). It has been suggested that teachers may not be the most relevant injunctive norm setters in early adolescence, as students typically strive for autonomy and independence from adult authorities. It is also possible that teachers’ influence on students’ lives varies, for instance depending on teachers’ personal characteristics or the quality of the student-teacher relationship (Kollerová et al., 2018). Overall, as they model students’ social behaviors

through both explicit and implicit messages (Wentzel, 2003), it is clear that teachers have a key role in bullying prevention (Kallestad & Olweus, 2003). However, whether students' perception of their teacher's attitudes mediated the effect of the class norm intervention still needs to be examined.

2.3.7 Norm-based messages may lack credibility in the participants' eyes

Finally, to be effective, information provided in norm-based interventions needs to be credible in the participants' eyes. This is one of the reasons why perceived injunctive rather than descriptive norm interventions are better suited in the case of pluralistic ignorance (Miller & Prentice, 2016). Students may not believe a norm-based message about the prevalence of a behavior because it contradicts their own observations of peers' behaviors. Private attitudes, on the other hand, cannot be easily accessed by external observers (Prentice, 2008). The claim that notwithstanding public behavior, not everyone approves of bullying in private may seem much more plausible.

Although our intervention fitted these recommendations in that it targeted injunctive rather than descriptive norms, we observed that not all students were fully convinced by the norm-based message. Although students' reports of their private and perceived classmates' attitudes were anonymous, students sometimes expressed their disbelief in the results. It is possible that the greater involvement of teachers in our intervention (compared to the original study; Perkins et al., 2011) elicited a bias of social desirability in students when reporting their attitudes toward bullying. The resulting actual norm may have seemed 'too good to be true' and thus lost its credibility in the eyes of some students, who then expressed their skepticism. This public communication of their doubts, in turn, may also have partially jeopardized the impact of the message conveyed by the intervention for the other students.

2.4 For intervening on moral disengagement a high implementation quality is always beneficial – but when quality is low, then fidelity matters.

As implementation of an anti-bullying intervention by teachers in natural contexts inevitably fluctuates (Goncy et al., 2015), it is crucial to determine the intervention's effects at different levels of implementation. When both implementation fidelity and quality are taken into account, intervening on moral disengagement significantly decreased moral disengagement, and, indirectly, bullying, for all the combinations of fidelity and quality levels. However, an increase in quality was significantly associated with a greater reduction in moral disengagement when fidelity was low. In the case of a low quality, on the other hand, a higher level of fidelity significantly increased the reduction in moral disengagement, and seemed to compensate the lack of quality in a way. Finally, a high quality seemed to ensure a greater reduction in moral disengagement in any case.

That a given intervention does not produce any iatrogenic effect when poorly implemented is a critical precondition to be further disseminated. On the other hand, to be effective even under suboptimal implementation conditions represents a strong asset.

Besides, the fact that student outcomes were nevertheless significantly impacted by the level of implementation to some extent provides further support for the effects being the result of the intervention rather than other factors (e.g., developmental changes) (Haataja et al., 2014).

It is important to note that different anti-bullying interventions and components are likely to react differentially to implementation conditions. That is, the fact that our intervention was effective even at a low level of fidelity does not mean that fidelity does not matter for anti-bullying interventions in general. The specific implementation conditions under which a given component is effective should be determined for each anti-bullying program component. This unique information about the outcomes of an intervention under various implementation conditions should be taken into account when considering the added value of including a component into an anti-bullying program, and its potential for bullying prevention in general. The fact that students who received the moral disengagement intervention systematically decreased their use of moral disengagement mechanisms regardless of the fidelity and quality levels provides strong support for the integration of a moral disengagement component into anti-bullying programs.

3. Strengths

The main strength of this thesis lies in its research design, which addressed three common methodological gaps of anti-bullying intervention studies that limit our understanding of how programs work and knowledge of the mechanisms involved in bullying. By assessing promising new anti-bullying components within separate interventions, this dissertation yielded valuable insights into their unique effects.

The assessment of the hypothesized mediators added significant value to the design. In that respect, this thesis provides an interesting illustration of the importance of measuring the mediating paths. First, the significant indirect effects found for the moral disengagement intervention would likely not have been detected without mediation, because they would have been concealed in the total effect at the distal level (i.e., on the behavioral outcomes) (Rucker et al., 2011), as confirmed by subsequent analyses not reported in the third chapter of this dissertation (same model, but without the mediation part; bullying $\beta = 0.07$, $p = .43$, outsider behaviors $\beta = 0.10$, $p = .36$, and defending $\beta = -0.05$, $p = .60$). We thus could have wrongly concluded that the moral disengagement component had no effect on the relevant student outcomes, and was not worth including in anti-bullying programs. Second, the class norm intervention had a direct but no indirect effect on bullying, thus leaving open the question of the mechanism(s) involved. Should this effect have been detectable at the distal level without mediation — which was also confirmed by subsequent analyses not reported in the third chapter ($\beta = -0.18$, $p < .01$), one could have wrongly assumed that the hypothesized (but not assessed) mediator did mediate change in bullying, and should therefore be targeted in future anti-bullying interventions.

Examination of the influence of implementation fidelity and quality on the effects of both interventions yielded additional information about their potential effectiveness and feasibility. The use of observational data for assessing the implementation is also a strength.

A particular asset of our design is the use of an active control group rather than a waiting list condition. Previous research into evidence-based practice has highlighted that spurious intervention effects may occur because of non-specific treatment factors. 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). By implementing an intervention of similar duration and intensity on a topic unrelated to bullying (climate and environmental issues), our design enabled us to control for these changes to some extent.

Finally, the relatively large sample with high retention rates, the use of latent change models, and the control for the nested structure of the data also strengthen this thesis.

4. Limitations

Despite its strengths, this thesis also has some limitations. First, some small yet significant between-groups differences were present at baseline despite the randomization process. To avoid this, we could have taken account of schools' initial average levels of bullying or other relevant variables in the allocation (following a stratification randomization procedure; e.g., Huitsing et al., 2020). However, a relatively small number of schools per condition (like in this thesis) may not give enough latitude in making a proper allocation.

Second, because our measure of the perceived injunctive class norm has not been previously validated with elementary students, we cannot be sure of its validity among students in our sample (Grades 4-6). Because of this uncertainty, we cannot definitely exclude the possibility that change in bullying was actually mediated by the perceived class injunctive norm. However, as a similar measure has been successfully used in seventh graders (Dillon & Lochman, 2019), we incline toward alternative interpretations of our findings, such as the popularity norm hypothesis previously presented.

Third, a limitation specific to the fourth chapter of this thesis was the absence of implementation data for the control group. We could therefore not include the control group in our analysis model because data were missing for all cases for the two moderators. As we contrasted our two intervention conditions to each other, all estimates and *p*-values provided by our analysis were conditional on the other intervention group. Such a study design was possible because each of our interventions specifically targeted one and only one component – and, accordingly, one mediator. Although each intervention could thus reasonably qualify as a comparison group for the other, a more suitable design would include implementation data from an active control group for proper comparison. As far as we know, however, such an 'ideal' research design has never been used before in anti-bullying intervention studies linking implementation to student outcomes, presumably because of the high cost in time and human resources.

Fourth, the absence of follow-up data collection limits our examination of the long-term effects of our interventions. A third wave of measurement would have enabled us to look at the sustainability and evolution of the effects. In the class norm intervention, in particular, we could have examined whether the decrease in bullying behaviors translated

into a change in bystanders' perceptions and behaviors at a later stage. Such outcome would have been consistent with our hypothesis that the change in perceived injunctive class norm translated into behavioral change in bullies first, which, in turn, could set a power balance more favorable to bystanders' intervention (Garandeau et al., 2019; Veenstra & Lodder, 2022).

5. Directions for future studies

While a significant number of anti-bullying programs have been developed over the past few decades, what the effective ingredients of anti-bullying programs are is an urgent question (Menesini & Salmivalli, 2017). The research described in this dissertation aimed to take a first step in opening the black box of anti-bullying programs, and to test two novel and promising anti-bullying components: intervening on moral disengagement and class norms. Below we outline some future directions that may help to expand our understanding of the effects of these two anti-bullying components for different students and age groups, clarify what type of norms and moral disengagement mechanism matter for intervention, and investigate conceptual and empirical articulation between these two concepts. Finally, we make suggestions regarding the next steps for anti-bullying program evaluation studies and implementation research.

5.1 Moral disengagement and class norms as anti-bullying components

5.1.1 Differential effects

It is likely that different subgroups of students are differentially impacted by anti-bullying components (e.g., De Mooij et al., 2020; Garandeau et al., 2014; Spilt et al., 2013). Our findings suggest that both moral disengagement and class norm may differentially affect the different participant roles in bullying (Salmivalli et al., 1996). Because the class norm intervention had a direct effect on bullying behaviors but not on bystander (defender and outsider) behaviors, we made the assumption that this component primarily had an effect on bullies. Whether this change, in turn, can influence bystanders still needs to be determined. Future studies should further investigate the effects of the class norm intervention on the different participant roles involved in bullying to clarify for whom it is effective.

Sensitivity analyses indicated a small direct effect increasing indirect (but not direct or cyber) bullying behaviors. 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. (This touches the notion of collective moral disengagement that we develop below.) Future studies should investigate whether some specific subgroups among the different participant roles are less or more receptive to this intervention. For instance, as popular bullies were found to be less responsive to KiVa than other students (Garandeau et al., 2014), they may be less responsive to the moral disengagement component as well. By

contrast, popular bystanders may be more receptive to the class norm intervention, and thus more likely to defend victims, as studies have suggested that a high social status was key in standing up to the bullies (e.g., Pöyhönen et al., 2010).

Taken together, these assumptions suggest the need to go beyond the variable-centered approach and to take a person-centered approach to investigate the differential effects of anti-bullying components. Sophisticated methods such as latent class and latent class growth analyses or (general) growth mixture modeling integrate variable- and person-centered analyses, making these methods particularly suitable for exploring these research questions (Muthén & Muthén, 2000).

5.1.2 Developmental effects

Differential effects of anti-bullying interventions are also likely to occur for different age groups (Yeager et al., 2015). Both moral judgement and peer influence evolve over the course of development. As children grow, they become increasingly capable of abstract reasoning and expanding their view from individual to institutional rules and laws regarding human well-being. Overall, they progress from external to increasing self-regulation. 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 middle- and high-school students, as it has been suggested that the tendency to conform and peer influence increase in preadolescence and adolescence (Sandstrom & Bartini, 2010), while the importance of teachers as injunctive norm setters seems to decrease (for instance, toward defending; Kollerová et al., 2018). Different types of norm-based interventions may also be more or less effective in different age groups. For instance, because younger children are more self-centered (Astington, 1993), they may be more receptive to PNF-type interventions. FGD-type interventions, on the other hand, may be more effective in preadolescents and adolescents, as children become increasingly susceptible to peer influence and conformity pressure during these developmental periods (Sandstrom & Bartini, 2010). Researchers have recently called for studies on how peer norm influence varies developmentally (Veenstra & Lodder, 2022). Future studies should thus examine the effects of moral disengagement and class norm anti-bullying components not only over the course of the school year to examine short- to mid-term effects of the interventions on different student profiles (e.g., on bystanders in the class norm intervention), but also across different age groups to investigate possible developmental effects.

5.1.3 What type of norms matters?

Our findings suggested that perceived injunctive class norm was not affected by the class norm intervention or related to change in bullying. We made the hypothesis that other types of norms mediated change in bullying. Perception of the injunctive class norm by high-status

students (rather than all students) may particularly matter. Alternatively, students' perception of the injunctive norm among their friends (rather than all their classmates) or conveyed by their class teacher may be more influential. Future studies should include and test these different types of norm as possible mediators within the same model to determine which of them plays a role in mediating change in bullying. Overall, it is clear that we still poorly understand the underlying processes of social norms (Veenstra & Laninga-Wijnen, 2022) and that research is needed that determines the differential effects of different types of norms (Veenstra & Lodder, 2022). Clarifying the underlying mechanisms involved in social norms would lead to different prevention avenues.

5.1.4 What mechanisms of moral disengagement matter?

Our validation of the Fr-MDBS enables us to validly examine either the global dimension or the distinct mechanisms of moral disengagement. The second chapter of this dissertation provided greater insight into the moral disengagement mechanisms related to bullying, victimization, and bystanders' roles. For instance, victim attribution was suggested as particularly relevant for bullying prevention, as it appeared to be related to all these variables. Different mechanisms have also been found to be involved in direct versus indirect bullying (e.g., Bjärehed et al., 2020) as well as in traditional versus cyber aggression (Pornari & Wood, 2010). Future investigations should examine further which of the mechanisms were involved in the mediation of direct, indirect, and cyber bullying change following the moral disengagement intervention in order to design more effective anti-bullying interventions.

5.1.5 Collective moral disengagement

In this dissertation, we examined the separate effects of moral disengagement and class norms as anti-bullying components. Even though these components are based on two different theoretical perspectives on the discrepancy between students' attitudes and actual behaviors regarding bullying, they are not mutually exclusive. Rather, in accordance with social cognitive theory, moral disengagement and social norms are likely to interact in influencing bullying-related behaviors (Bjärehed et al., 2019). Socio-structural and psychological views are sometimes deemed to be rival conceptions, but (im)moral actions result from bidirectional interactions between individual and social determinants (Bandura, 2002). The recent concept of collective moral disengagement refers to moral disengagement beliefs that are shared within a given social group, and was found to influence behaviors at both the individual and class levels (Gini et al., 2015). Importantly, collective moral disengagement is fundamentally a group phenomenon, rather than the sum of each individual's moral disengagement in the group (Bandura, 2016). It is about individuals' perceptions of the extent to which moral disengagement mechanisms are shared within the group, which are produced by the group dynamics (Bjärehed et al., 2019). As such, we argue that collective moral disengagement may be seen as a perceived norm regarding common or approved moral disengagement mechanisms regarding bullying within the class. Students' perception of collective moral disengagement in the class has been found to be associated with bullying in a few recent studies (e.g., Thornberg et al., 2021), even after controlling for individual moral disengagement (Bjärehed et al., 2019; Thornberg et al., 2019). More

research, including longitudinal studies, is needed to examine the links between collective moral disengagement and individual bullying-related behaviors (Thornberg et al., 2021). Besides, the theoretical articulation between individual moral disengagement, collective moral disengagement, and class norms should be further developed.

5.2 What next regarding the black box of anti-bullying program evaluations?

5.2.1 One step at a time: Combining components

Most evaluation studies assess anti-bullying program components as a whole, therefore limiting our knowledge of the respective contribution of each component (Menesini & Salmivalli, 2017). There has been a call for further innovative research rather than evaluating whole packages of interventions (Eisner & Malti, 2012). A few studies have recently made some efforts in this direction, for instance by comparing two experimental conditions of the same program, with and without a specific component (Gaete et al., 2017; Huitsing et al., 2020; van Verseveld et al., 2021). A recent meta-analysis that has examined the link between effect size and some specific characteristics of programs (e.g., whole-school approach, anti-bullying policies, classroom rules, information for parents) is also worthy of mention (Gaffney et al., 2021). In this dissertation, however, we have advocated a more drastic ‘return to the roots’ for anti-bullying programs by isolating promising components and investigating their unique effects within separate interventions (Eisner & Malti, 2012). However, intervening on a single component can hardly be seen as an anti-bullying program by itself. First, effect sizes of complete anti-bullying interventions are mostly small, especially when evaluated with more rigorous methodological designs such as randomized controlled trials (Gaffney et al., 2019). It is thus unlikely that the effect of a single component could exceed a small size. Second, the number of components included in an anti-bullying program has been shown to be related to its effectiveness (Fox et al., 2012), although no significant relation was found in another recent meta-analysis (Gaffney et al., 2021). To maximize the effectiveness of anti-bullying interventions, relevant components should therefore likely be combined. It is possible that when two components are combined, the effectiveness of a program increases. However, adding two or more components that are effective in isolation will not necessarily result in an addition of their respective effects. It is also possible that some components will no longer make a significant contribution to the overall program effects once they are combined with one or more other components (Menesini & Salmivalli, 2017). The inclusion of each additional component needs to be carefully weighted in terms of its cost-effectiveness from a public health perspective (Menesini & Salmivalli, 2017). The next step for opening the black box of anti-bullying programs will thus be to test, step by step, the joint effects of combining components that have proven to be effective in isolation.

5.2.2 Teacher-related factors promoting implementation

Proper implementation of an intervention is required for any effects to occur (Kallestad & Olweus, 2003; Salmivalli et al., 2013). In the fourth chapter of this dissertation, we highlighted that while a high quality was always beneficial, fidelity had an impact on the

effectiveness of the moral disengagement intervention when quality was low. Different components, however, are likely to require different conditions of implementation to be effective. Implementation fidelity, notably, is likely to matter in other interventions. More studies are needed that determine under what implementation conditions specific anti-bullying components are effective, and to what extent. Some studies have examined school-related factors that favor sustainable program implementation at the school level (e.g., Sainio et al., 2020). However, in order to foster a high level of implementation by teachers, future studies should examine what teacher-related factors predict different implementation features.

To date, literature on this question has been almost nonexistent. In the rare studies that have looked into this issue so far, teachers' beliefs regarding intervention effectiveness, principal support (Haataja et al., 2015), colleague support (Cunningham et al., 2016), and burn-out (Swift et al., 2017) have been found to influence implementation. We argue that both school- and individual-level factors are likely to influence how teachers implement anti-bullying interventions. For instance, as teachers' personal experience with bullying has been found to influence how they respond to bullying (Yoon et al., 2016), it would be worth examining whether this factor also affects how they implement anti-bullying interventions. Similarly, as teachers' self-efficacy toward bullying situations is related to their probability of intervention (Fisher & Bilz, 2018), future studies should examine whether it influences their implementation of anti-bullying interventions.

6. Practical implications

Determining the potential of anti-bullying components 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).

Overall, our findings provided strong support for the integration of a moral disengagement component into anti-bullying programs, as it appeared to be an effective way of targeting both bystanders and bullies, and reduced moral disengagement and (indirectly) bullying even when poorly implemented. Moreover, this dissertation showed that students' use of moral disengagement mechanisms could be easily and quickly reduced through a relatively light (five hours) intervention delivered by class teachers. Time constraints always represent a practical issue when it comes to implementing well-being or health-related prevention programs in schools. Finding a balance between regular academic instruction and initiatives for improving the social climate is usually challenging for teachers, and they tend to treat the two as separate tasks (Swearer et al., 2009). The integration of bullying prevention into daily instruction has been suggested as facilitating skill generalization, but is rarely done in practice (Domitrovich et al., 2010; Swearer et al., 2009).

Bibliotherapy, or the use of reading materials to assist personal problem-solving, is already well known to school psychologists and education professionals in general with regard to conduct problems (Wang & Goldberg, 2017). This approach has the double advantage of being familiar to teachers and easily integrated into the regular curriculum. Given their limited time resources, this makes it a particularly suitable approach for schools (Wang & Goldberg, 2017) and has been used in other bullying prevention programs (STORIES; Telgasi & Rothman, 2001; WITS Program; Leadbeater & Sukhawathanakul, 2011). Using storybooks displaying bullying situations to illustrate the use of moral disengagement mechanisms seems to be particularly suitable for targeting moral disengagement (Wang & Goldberg, 2017), and might also prove to be useful for targeting other anti-bullying components.

In sum, moral disengagement presents a good cost-effectiveness ratio, which makes it an appealing anti-bullying program component. Beyond the context of a structured anti-bullying intervention, sound 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. As such, raising teachers' awareness of this socio-cognitive process and its mechanisms should be part of their initial training curriculum.

Even though the causal path of our class norm intervention remains unknown, it did have an impact on students' bullying behaviors. As such, intervening on the class norms can be seen as a promising anti-bullying component to further explore, provided that future studies clarify the underlying process conveying change in bullying. However, our intervention involved the identification of students' social status and bullying role to form work groups with an optimal composition (e.g., prosocial and well-liked students rather than bullies or victims holding key roles in the activities). This was done in collaboration with the class teachers based on both their individual perceptions and knowledge of students' personalities and the pre-test data. As teachers generally have difficulty in interpreting social dynamics among their students correctly (Kaufman et al., 2020), leaving this step of the intervention to their discretion seems unwise. Notably, teachers widely underestimate bullying prevalence (e.g., Bradshaw et al., 2007) and struggle to identify victims (e.g., Haataja et al., 2015).

Providing teachers with student relationship information would thus be necessary to enable them to implement the class norm intervention in autonomy. In that respect, network diagnostics have been recently suggested as promising tools to support teachers in tackling bullying in everyday school life (Kaufman et al., 2020). Network diagnostics have been proven effective in diverse domains, including health-related interventions (Valente and Pitts 2017) and could be useful to combine with the class norm intervention. Moreover, teachers also convey norms to their students (Veenstra et al., 2014). Raising teachers' awareness of social dynamics, class norms, and group processes is therefore important (Veenstra & Lodder, 2022), notably during their initial training as teachers and first work experiences during traineeships.

This dissertation has highlighted that it mattered how well teachers implemented the moral disengagement intervention. Implementation quality encompasses not only technical

aspects such as the clarity and fluency of the delivery, but also pedagogical and relational skills such as classroom management and the ability to get students interested and involved in the lesson, and to adapt the lesson to their needs. From this conceptualization, it seems quite obvious that implementation quality is likely to reflect transversal skills that are highly desirable in teachers in general. Although it is likely impossible to define what is “a good teacher”, the role of teachers’ personal qualities (e.g., empathy, sensitivity, creativity, understanding, flexibility, enthusiasm; Tickle, 1999) has been emphasized by many researchers, challenging the more traditional competency-based view (Korthagen, 2004). This is of course closely linked to the question of what are the very missions of the school, which is far beyond the scope of this thesis. However, from a bullying prevention perspective, our findings suggest the need for a greater emphasis on these qualities in teachers’ initial training programs. It may be important to not only raise student teachers’ cognitive awareness of these core qualities, but also to ensure that they are emotionally in touch with them (Korthagen, 2004).

7. General conclusion

Most evaluation studies assess program components as a whole (Menesini & Salmivalli, 2017), omit to test the mediating paths (Saarento et al., 2015), and do not assess implementation (Salmivalli et al., 2013). We argue that these three gaps form the “black box” of program evaluations, which limits our understanding of how programs work and knowledge of the factors involved in bullying, and may explain the mixed and limited results of anti-bullying programs (Bradshaw, 2015; Gaffney et al., 2019). This dissertation thus advocates a return to the roots of anti-bullying programs. To determine the potential of anti-bullying components and design interventions with the best cost-effectiveness ratio, we need first to assess components in isolation, test their hypothesized mediating path, and associate their implementation with student outcomes. By addressing these methodological gaps, this dissertation provides clear guidelines for the assessment of anti-bullying components and aims to set the course for future evaluation studies.

The limited effectiveness of anti-bullying programs should also prompt researchers to keep studying new approaches to preventing bullying by testing interventions that target under-explored relevant predictors of bullying (Eisner & Malti, 2012). In this respect, this dissertation provides valuable insights about the effects of intervening on moral disengagement and class norms as anti-bullying components, and suggests that both theoretical frameworks be taught in student teachers’ training programs. There is a need for future studies to further explore the underlying mechanism of the class norm intervention, with a specific focus on injunctive norms perceived by popular students, and several measurement waves over the school year. The presence of differential effects for different age groups, bullying roles, as well as subgroups based on popularity should be investigated in both interventions.

Finally, this dissertation underlines the importance of quality as an implementation feature in anti-bullying interventions, highlights its close link with teachers’ more personal

qualities, and suggests that these qualities should be covered in teachers' initial training. In addition, future studies should examine what teacher-related factors predict implementation and which ones might be directly enhanced during the anti-bullying intervention training. In conclusion, no step should be skipped in the fight against school bullying.

References

References

- Allison, K. R., & Bussey, K. (2017). Individual and collective moral influences on intervention in cyberbullying. *Computers in Human Behavior*, 74, 7-15. <https://doi.org/10.1016/j.chb.2017.04.019>
- Almeida, A., Correia, I., & Marinho, S. (2009). Moral disengagement, normative beliefs of peer group, and attitudes regarding roles in bullying. *Journal of School Violence*, 9(1), 23-36. <https://doi.org/10.1080/15388220903185639>
- 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. <http://dx.doi.org/10.1016/j.avb.2015.05.007>
- Andrews, P. W. (2001). The psychology of social chess and the evolution of attribution mechanisms: Explaining the fundamental attribution error. *Evolution and Human Behavior*, 22(1), 11-29. [https://doi.org/10.1016/S1090-5138\(00\)00059-3](https://doi.org/10.1016/S1090-5138(00)00059-3)
- Astington, J. W. (1993). *The child's discovery of the mind* (Vol. 31). Harvard University Press.
- Axford, N., Bjornstad, G., Clarkson, S., Ukoumunne, O. C., Wrigley, Z., Matthews, J., ... & Hutchings, J. (2020). The effectiveness of the KiVa bullying prevention program in Wales, UK: Results from a pragmatic cluster randomized controlled trial. *Prevention Science*, 21(5), 615-626. <https://doi.org/10.1007/s11121-020-01103-9>
- Bagès, C., Hoareau, N., & Guerrien, A. (2020). Play to Reduce Bullying! Role-Playing Games Are a Useful Tool for Therapists and Teachers. *Journal of Research in Childhood Education*. Advance online publication. <https://doi.org/10.1080/02568543.2020.1810834>
- Baldry, A. C. (2004). The impact of direct and indirect bullying on the mental and physical health of Italian youngsters. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, 30(5), 343-355. <https://doi.org/10.1002/ab.20043>
- Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ, 23-28.
- Bandura, A. (1990). Selective activation and disengagement of moral control. *Journal of Social Issues*, 46(1), 27-46.
- Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50(2), 248-287.
- Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, 71(2), 364-374. <https://doi.org/10.1037/0022-3514.71.2.364>
- Bandura, A., Caprara, G. V., Barbaranelli, C., Pastorelli, C., & Regalia, C. (2001). Sociocognitive self-regulatory mechanisms governing transgressive behavior. *Journal of Personality and Social Psychology*, 80(1), 125-135. <https://doi.org/10.1037/0022-3514.80.1.125>

- Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of Moral Education*, 31(2), 101-119. <https://doi.org/10.1080/0305724022014322>
- Bandura, A. (2016). *Moral disengagement: How people do harm and live with themselves*. Worth publishers.
- Barchia, K., & Bussey, K. (2011). Individual and collective social cognitive influences on peer aggression: Exploring the contribution of aggression efficacy, moral disengagement, and collective efficacy. *Aggressive Behavior*, 37(2), 107-120. <https://doi.org/10.1002/ab.20375>
- Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the pieces together: An integrated model of program implementation. *Prevention Science*, 12(1), 23-33. <https://doi.org/10.1007/s11121-010-0186-1>
- Berkowitz, A. D. (2004). *The social norms approach: Theory, research, and annotated bibliography*.
- Bierman, K. L. (2011). The promise and potential of studying the “invisible hand” of teacher influence on peer relations and student outcomes: A commentary. *Journal of Applied Developmental Psychology*, 32(5), 297-303. <https://doi.org/10.1016/j.appdev.2011.04.004>
- *Biggs, B. K., Vernberg, E. M., Twemlow, S. W., Fonagy, P., & Dill, E. J. (2008). Teacher adherence and its relation to teacher attitudes and student outcomes in an elementary school-based violence prevention program. *School Psychology Review*, 37(4), 533-549. <https://doi.org/10.1080/02796015.2008.12087866>
- Bjärehed, M., Thornberg, R., Wänström, L., & Gini, G. (2019). Individual moral disengagement and bullying among Swedish fifth graders: The role of collective moral disengagement and pro-bullying behavior within classrooms. *Journal of Interpersonal Violence*. Advanced online publication. <https://doi.org/10.1177/088620519860889>
- Bjärehed, M., Thornberg, R., Wänström, L., & Gini, G. (2020). Mechanisms of moral disengagement and their associations with indirect bullying, direct bullying, and pro-aggressive bystander behavior. *The Journal of Early Adolescence*, 40(1), 28-55. <https://doi.org/10.1177/0272431618824745>
- *Black, S. A., & Jackson, E. (2007). Using bullying incident density to evaluate the Olweus Bullying Prevention Programme. *School Psychology International*, 28(5), 623-638. <https://doi.org/10.1177/0143034307085662>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in Public Health*, 6, 149. <https://doi.org/10.3389/fpubh.2018.00149>
- Bradshaw, C. P. (2015). Translating research to practice in bullying prevention. *American psychologist*, 70(4), 322-332. <http://dx.doi.org/10.1037/a0039114>
- Bradshaw, C. P., Sawyer, A. L., & O'Brennan, L. M. (2007). Bullying and peer victimization at school: Perceptual differences between students and school staff. *School Psychology Review*, 36(3), 361-382. <https://doi.org/10.1080/02796015.2007.12087929>

- Brighi, A., Ortega, R., Scheitauer, H., Smith, P. K., Tsormpatzoudis, C., Barkoukis, V., Del Rey, R., & Thompson, J. (2012). European bullying intervention project questionnaire (EBIPQ). University of bologna.
- Burchell, K., Rettie, R., & Patel, K. (2013). Marketing social norms: social marketing and the 'social norm approach'. *Journal of Consumer behaviour*, 12(1), 1-9. <https://doi.org/10.1002/cb.1395>
- Byrne, B. M. (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment*, 85(1), 17-32. https://doi.org/10.1207/s15327752jpa8501_02
- Caprara, G. V., Pastorelli, C., & Bandura, A. (1995). La misura del disimpegno morale in età evolutiva. *Età Evolutiva*, 51, 18-29.
- Carney, A. G., & Merrell, K. W. (2001). Bullying in schools: Perspectives on understanding and preventing an international problem. *School Psychology International*, 22(3), 364-382. <https://doi.org/10.1177/0143034301223011>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 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(1), 219-251. <https://doi.org/10.1111/j.1467-6494.2011.00739.x>
- Cho, S. (2018). The impact of low self-control and delinquent peer associations on bullying perpetration and victimization among South Korean adolescents: Time-concurrent, time-lagged, and latent growth curve modeling. *Journal of school violence*, 17(4), 500-520. <https://doi.org/10.1080/15388220.2018.1453821>
- Cillessen, A. H., & Marks, P. E. (2017). Methodological choices in peer nomination research. *New Directions for Child and Adolescent Development*, 157, 21-44. <https://doi.org/10.1002/cad.20206>
- Clarkson, S., Charles, J. M., Saville, C. W., Bjornstad, G. J., & Hutchings, J. (2019). Introducing KiVa school-based antibullying programme to the UK: A preliminary examination of effectiveness and programme cost. *School Psychology International*, 40(4), 347-365. <https://doi.org/10.1177/0143034319841099>
- Colpin, H., Bauman, S., & Menesini, E. (2021). Teachers' responses to bullying: Unravelling their consequences and antecedents. Introduction to the special issue. *European Journal of Developmental Psychology*, 1-17. <https://doi.org/10.1080/17405629.2021.1954903>
- Crapanzano, A. M., Frick, P. J., Childs, K., & Terranova, A. M. (2011). Gender differences in the assessment, stability, and correlates to bullying roles in middle school children. *Behavioral Sciences & the Law*, 29(5), 677-694. <https://doi.org/10.1002/bsl.1000>

- Cross, D., & Barnes, A. (2014). One size doesn't fit all: Re-thinking implementation research for bullying prevention. In R. Schott & D. Søndergaard (Eds.), *School Bullying: New Theories in Context* (pp. 405-418). Cambridge University Press. <https://doi.org/10.1017/CBO9781139226707.022>
- Cunningham, C. E., Rimas, H., Mielko, S., Mapp, C., Cunningham, L., Buchanan, D., ... & Marcus, M. (2016). What limits the effectiveness of antibullying programs? A thematic analysis of the perspective of teachers. *Journal of School Violence*, 15(4), 460-482. <https://doi.org/10.1080/15388220.2015.1095100>
- Currie, C., Zanotti, C., Morgan, A., Cuurie, D., de Looze, M., Roberts, C., ... Barnekow, V. (2012). Social determinants of health and well-being among young people. *Health Behaviour in School-aged Children (HBSC) study: international report from the 2009/2010 survey*. Copenhagen, WHO Regional Office for Europe (Health Policy for Children and Adolescents, No. 6). http://www.euro.who.int/_data/assets/pdf_file/0007/167281/E96444_part1.pdf
- Demol, K., Verschueren, K., Jame, M., Lazard, C., & Colpin, H. (2021). Student attitudes and perceptions of teacher responses to bullying: An experimental vignette study. *European Journal of Developmental Psychology*, Advanced online publication. <https://doi.org/10.1080/17405629.2021.1896492>
- Demol, K., Verschueren, K., Samlivalli, C., & Colpin, H. (2021). Perceived teacher responses to bullying influence students' social cognitions. *Frontiers in Psychology*, 11. Advanced online publication. <https://doi.org/10.3389/fpsyg.2020.592582>
- de Mooij, B., Fekkes, M., Scholte, R. H., & Overbeek, G. (2020). Effective components of social skills training programs for children and adolescents in nonclinical samples: A multilevel meta-analysis. *Clinical child and family psychology review*, Advanced online publication. <https://doi.org/10.1007/s10567-019-00308-x>
- Dijkstra, J. K., Lindenberg, S., & Veenstra, R. (2008). Beyond the class norm: Bullying behavior of popular adolescents and its relation to peer acceptance and rejection. *Journal of Abnormal Child Psychology*, 36, 1289-1299. <https://doi.org/10.1007/s10802-008-9251-7>
- Dillon, C. E., & Lochman, J. E. (2019). Correcting for norm misperception of anti-bullying attitudes. *International Journal of Behavioral Development*, Advance online publication. <https://doi.org/10.1177/0165025419860598>
- Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, 43(2), 121-149. <https://doi.org/10.1177/0748175610373459>
- *Doğan, A., Keser, E., Şen, Z., Yanagida, T., Gradingier, P., & Strohmeier, D. (2017). Evidence based bullying prevention in Turkey: Implementation of the ViSC social competence program. *International Journal of Developmental Science*, 11(3-4), 93-108. <https://doi.org/10.3233/DEV-170223>
- Domitrovich, C. E., Moore, J. E., & Greenberg, M. T. (2012). Maximizing the effectiveness of social-emotional interventions for young children through high-quality implementation of evidence-based interventions. In B. Kelly & D. F. Perkins (Eds.), *Handbook of implementation science for psychology in education* (pp. 207-229). Cambridge University Press. <https://doi.org/10.1017/CBO9781139013949.017>

- Domitrovich, C. E., Bradshaw, C. P., Greenberg, M. T., Embry, D., Poduska, J. M., & Jalongo, N. S. (2010). Integrated models of school-based prevention: Logic and theory. *Psychology in the Schools*, 47(1), 71-88. <https://doi.org/10.1002/pits.20452>
- 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(1), 144-172. <https://doi.org/10.13110/merrpalmquar1982.61.1.0144>
- Durlak, J. A. (2010). The importance of doing well in whatever you do: A commentary on the special section, "Implementation research in early childhood education". *Early Childhood Research Quarterly*, 25(3), 348-357. <https://doi.org/10.1016/j.ecresq.2010.03.003>
- 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(3-4), 327-350. <https://doi.org/10.1007/s10464-008-9165-0>
- Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health education research*, 18(2), 237-256. <https://doi.org/10.1093/her/18.2.237>
- Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, 21(2), 553-565. <https://doi.org/10.1093/bib/bbz016>
- 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(2), 166-175. <https://doi.org/10.4119/ijcv-2909>
- Enders, C. K. (2010). *Applied missing data analysis*. Guilford.
- Espelage, D. L., Holt, M. K., & Henkel, R. R. (2003). Examination of peer-group contextual effects on aggression during early adolescence. *Child Development*, 74(1), 205-220. <https://doi.org/10.1111/1467-8624.00531>
- Farmer, T. W., Lines, M. M., & Hamm, J. V. (2011). Revealing the invisible hand: The role of teachers in children's peer experiences. *Journal of Applied Developmental Psychology*, 32(5), 247-256. <https://doi.org/10.1016/j.appdev.2011.04.006>
- Farrington, D. P., & Ttofi, M. M. (2009). School-based programs to reduce bullying and victimization. *Campbell Systematic Reviews*, 6(1), i-148. <https://doi.org/10.4073/csr.2009.6>
- Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Stanford university press.
- Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* 5th ed.
- Fischer, S. M., & Bilz, L. (2019). Teachers' self-efficacy in bullying interventions and their probability of intervention. *Psychology in the Schools*, 56(5), 751-764. <https://doi.org/10.1002/pits.22229>

- Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., Wallace, F., Burns, B., ... & Shern, D. (2005). *Implementation research: A synthesis of the literature*. University of South Florida.
- Fox, B. H., Farrington, D. P., & Ttofi, M. M. (2012). Successful bullying prevention programs: Influence of research design, implementation features, and program components. *International Journal of Conflict and Violence (IJCIV)*, 6(2), 273-282.
- 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(1), 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), 1-9. <https://doi.org/10.1186/s13063-017-1810-1>
- Gaffney, H., Ttofi, M. M., & Farrington, D. P. (2019). 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. (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>
- Gage, N. A., Rose, C. A., & Kramer, D. A. (2019). When prevention is not enough: students' perception of bullying and school-wide positive behavior interventions and supports. *Behavioral Disorders*, 45(1), 29-40. <https://doi.org/10.1177/019874291810761>
- Galand, B., Hospel, V. & Baudoin, N. (2014). Prévenir le harcèlement via les pratiques de classe ? Une étude multiniveaux. [Preventing bullying through classroom practices? A multilevel study.] *Revue Québécoise de Psychologie*, 35(3), p. 137-157. <http://hdl.handle.net/2078.1/152752>
- Garandeau, C. F., Lee, I. A., & Salmivalli, C. (2014). Differential effects of the KiVa anti-bullying program on popular and unpopular bullies. *Journal of Applied Developmental Psychology*, 35(1), 44-50. <https://doi.org/10.1016/j.appdev.2013.10.004>
- Garandeau, C. F., Vermande, M. M., Reijntjes, A. H. A., & Aarts, E. (2019). Classroom bullying norms and peer status: Effects on victim-oriented and bully-oriented defending. *International Journal of Behavioral Development*. Advanced online publication. <https://doi.org/10.1177/0165025419894722>
- Geiser, C. (2013). *Data analysis with Mplus*. Guilford.
- Gibbs, J. C., Potter, G. B., Barriga, A. Q., & Liao, A. K. (1996). Developing the helping skills and prosocial motivation of aggressive adolescents in peer group programs. *Aggression and Violent Behavior*, 1(3), 283-305. [https://doi.org/10.1016/1359-1789\(95\)00018-6](https://doi.org/10.1016/1359-1789(95)00018-6)
- Gini, G. (2006). Social cognition and moral cognition in bullying: What's wrong?. *Aggressive Behavior*, 32(6), 528-539. <https://doi.org/10.1002/ab.20153>

- Gini, G., Albiero, P., Benelli, B., & Altoe, G. (2008). Determinants of adolescents' active defending and passive bystanding behavior in bullying. *Journal of Adolescence*, 31(1), 93-105. <https://doi.org/10.1016/j.adolescence.2007.05.002>
- Gini, G., Pozzoli, T., & Bussey, K. (2015). The role of individual and collective moral disengagement in peer aggression and bystanding: A multilevel analysis. *Journal of Abnormal Child Psychology*, 43(3), 441-452. <https://doi.org/10.1007/s10802-014-9920-7>
- Gini, G., Pozzoli, T., & Hauser, M. (2011). Bullies have enhanced moral competence to judge relative to victims, but lack moral compassion. *Personality and Individual Differences*, 50(5), 603-608. <https://doi.org/10.1016/j.paid.2010.12.002>
- Gini, G., Pozzoli, T., & Hymel, S. (2014). Moral disengagement among children and youth: A meta-analytic review of links to aggressive behavior. *Aggressive Behavior*, 40(1), 56-68. <https://doi.org/10.1002/ab.21502>
- Goncy, E. A., Sutherland, K. S., Farrell, A. D., Sullivan, T. N., & Doyle, S. T. (2015). Measuring teacher implementation in delivery of a bullying prevention program: The impact of instructional and procedural adherence and competence on student responsiveness. *Prevention Science*, 16(3), 440-450. <https://doi.org/10.1007/s11121-014-0508-9>
- Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention science*, 16(7), 893-926. <https://doi.org/10.1007/s11121-015-0555-x>
- Graham, S., & Juvonen, J. (1998). Self-blame and peer victimization in middle school: an attributional analysis. *Developmental Psychology*, 34(3), 587. <https://doi.org/10.1037/0012-1649.34.3.587>
- *Greif Green, J., Holt, M. K., Oblath, R., Robinson, E., Storey, K., & Merrin, G. J. (2020). Engaging professional sports to reduce bullying: an evaluation of the Boston vs. Bullies Program. *Journal of School Violence*, 19(3), 389-405. <https://doi.org/10.1080/15388220.2019.1709849>
- Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3), 271-299.
- Guo, Q., Guo, Y., Qiao, X., Leng, J., & Lv, Y. (2021). Chance locus of control predicts moral disengagement which decreases well-being. *Personality and Individual Differences*, 171 (online). <https://doi.org/10.1016/j.paid.2020.110489>
- *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(5), 479-493. <http://dx.doi.org/10.1016/j.jsp.2014.07.001>
- 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(4), 564. <http://dx.doi.org/10.1037/spq0000121>

- Hawley, P. H. (1999). The ontogenesis of social dominance: A strategy-based evolutionary perspective. *Developmental Review*, 19(1), 97-132. <https://doi.org/10.1006/drev.1998.0470>
- Hayes, A. F. (2018) *Introduction to mediation, moderation, and conditional process analysis*. (2nd ed.). Guilford.
- Hayes, A. F., & Preacher, K. J. (2014). Statistical mediation analysis with a multicategorical independent variable. *British Journal of Mathematical and Statistical Psychology*, 67(3), 451-470. <https://doi.org/10.1111/bmsp.12028>
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60-87. <https://doi.org/10.3102/0162373707299706>
- Hektner, J. M., & Swenson, C. A. (2012). Links from teacher beliefs to peer victimization and bystander intervention: Tests of mediating processes. *The Journal of Early Adolescence*, 32(4), 516-536. <https://doi.org/10.1177/0272431611402502>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29. <https://doi.org/10.1038/466029a>
- Henry, D. B., Dymnicki, A. B., Schoeny, M. E., Meyer, A. L., Martin, N. C., & Multisite Violence Prevention Project. (2013). Middle school students overestimate normative support for aggression and underestimate normative support for nonviolent problem-solving strategies. *Journal of applied social psychology*, 43(2), 433-445. <https://doi.org/10.1111/j.1559-1816.2013.01027.x>
- Henry, D., Guerra, N., Huesmann, R., Tolan, P., VanAcker, R., & Eron, L. (2000). Normative influences on aggression in urban elementary school classrooms. *American Journal of Community Psychology*, 28(1), 59-81. <https://doi.org/10.1023/A:1005142429725>
- *Hirschstein, M. K., van Schoiack Edstrom, L., Frey, K. S., Snell, J. L., & MacKenzie, E. P. (2007). Walking the talk in bullying prevention: Teacher implementation variables related to initial impact of the Steps to Respect program. *School Psychology Review*, 36(1), 3-21. <https://doi.org/10.1080/02796015.2007.12087949>
- Hoareau, N., Bagès, C., Allaire, M., & Guerrien, A. (2019). The role of psychopathic traits and moral disengagement in cyberbullying among adolescents. *Criminal Behaviour and Mental Health*, 29(5-6), 321-331. <https://doi.org/10.1002/cbm.2135>
- Hong, J. S., & Espelage, D. L. (2012). A review of research on bullying and peer victimization in school: An ecological system analysis. *Aggression and Violent Behavior*, 17(4), 311-322. <https://doi.org/10.1016/j.avb.2012.03.003>
- Hogg, M. A. (2010). Influence and leadership. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.). *Handbook of social psychology* (Vol. 2, pp. 1166-1207). John Wiley & Sons.
- Hong, J. S., Kim, D. H., & Piquero, A. R. (2017). Assessing the links between punitive parenting, peer deviance, social isolation and bullying perpetration and victimization in South Korean adolescents. *Child abuse & neglect*, 73, 63-70. <https://doi.org/10.1016/j.chiabu.2017.09.017>

- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Huitsing, G., Dijkstra, J. K., & Veenstra, R. (2019). Social development and group processes: A social network application to bullying and network interventions. In E. S. Kunnen, N. M. P. De Ruiter, B. F. Jerominus, & M. A. E. Van Der Gaag (Eds.) *Psychosocial development in adolescence* (pp. 193-208). Routledge.
- Huitsing, G., Lodder, G. M., 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>
- Hymel, S., & Bonanno, R. A. (2014). Moral disengagement processes in bullying. *Theory into Practice*, 53(4), 278-285. <https://doi.org/10.1080/00405841.2014.947219>
- Hymel, S., Rocke-Henderson, N., & Bonanno, R. A. (2005). Moral disengagement: A framework for understanding bullying among adolescents. *Journal of Social Sciences*, 8(1), 1-11.
- 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(4), 293. <https://doi.org/10.1037/a0038928>
- Jacobs, D. & Rea, A. (2005). Construction et importation des classements ethniques. Allochtones et immigrés aux Pays-Bas et en Belgique. *Revue Européenne des Migrations Internationales*, 21(2) 35-59. <https://doi.org/10.4000/remi.2487>
- Jenkins, L. N., & Troop-Gordon, W. (2020). Bystanders in bullying : Introduction to the special issue. *Journal of School Psychology*, 81, 47-50. <https://doi.org/10.1016/j.jsp.2020.07.002>
- *Johander, E., Turunen, T., Garandeau, C. F., & Salmivalli, C. (2021). Different approaches to address bullying in KiVa schools: adherence to guidelines, strategies implemented, and outcomes obtained. *Prevention Science*, 22(3), 299-310. <https://doi.org/10.1007/s11121-020-01178-4>
- Jolliffe, D., & Farrington, D. P. (2011). Is low empathy related to bullying after controlling for individual and social background variables? *Journal of Adolescence*, 34(1), 59-71. <https://doi.org/10.1016/j.adolescence.2010.02.001>
- *Joronen, K., Konu, A., Rankin, H. S., & Åstedt-Kurki, P. (2012). An evaluation of a drama program to enhance social relationships and anti-bullying at elementary school: a controlled study. *Health Promotion International*, 27(1), 5-14. <https://doi.org/10.1093/heapro/dar012>
- Juvonen, J., & Galvan, A. (2008). Peer influence in involuntary social groups: Lessons from research on bullying. In M. J. Prinstein & K. A. Dodge (Eds.), *Understanding peer influence in children and adolescents* (pp. 225-244). The Guilford Press.

- Kallestad, J. H., & Olweus, D. (2003). Predicting teachers' and schools' implementation of the Olweus bullying prevention program: A multilevel study. *Prevention & Treatment*, 6(1).
- *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(1), 311-330. <https://doi.org/10.1111/j.1467-8624.2010.01557.x>
- Kaufman, T. M., Huising, G., Bloemberg, R., & Veenstra, R. (2021). The systematic application of network diagnostics to monitor and tackle bullying and victimization in schools. *International Journal of Bullying Prevention*, 3(1), 75-87. <https://doi.org/10.1007/s42380-020-00064-5>
- Keijsers, L., Branje, S., Hawk, S. T., Schwartz, S. J., Frijns, T., Koot, H. M., ... & Meeus, W. (2012). Forbidden friends as forbidden fruit: Parental supervision of friendships, contact with deviant peers, and adolescent delinquency. *Child Development*, 83(2), 651-666. <https://doi.org/10.1111/j.1467-8624.2011.01701.x>
- *Kelm, J. L., McIntosh, K., & Cooley, S. (2014). Effects of implementing school-wide positive behavioural interventions and supports on problem behaviour and academic achievement in a Canadian elementary school. *Canadian Journal of School Psychology*, 29(3), 195-212. <https://doi.org/10.1177/0829573514540266>
- Kert, A. S., Coddington, R. S., Tryon, G. S., & Shiyko, M. (2010). Impact of the word “bully” on the reported rate of bullying behavior. *Psychology in the Schools*, 47(2), 193-204. <https://doi.org/10.1002/pits.20464>
- 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(4), 450-462. <https://doi.org/10.1002/ab.21833>
- Kochenderfer-Ladd, B., & Pelletier, M. E. (2008). Teachers' views and beliefs about bullying: Influences on classroom management strategies and students' coping with peer victimization. *Journal of School Psychology*, 46(4), 431-453. <https://doi.org/10.1016/j.jsp.2007.07.005>
- Kochenderfer-Ladd, B., & Troop-Gordon, W. (2010). Introduction to the special issue contexts, causes, and consequences: New directions in peer victimization research. *Merrill-Palmer Quarterly*, 56(3), 221-231.
- Kollerová, L., Yanagida, T., Mazzone, A., Soukup, P., & Strohmeier, D. (2018). “They think that I should defend”: Effects of peer and teacher injunctive norms on defending victimized classmates in early adolescents. *Journal of youth and adolescence*, 47(11), 2424-2439. <https://doi.org/10.1007/s10964-018-0918-2>
- Korthagen, F. A. (2004). In search of the essence of a good teacher: Towards a more holistic approach in teacher education. *Teaching and teacher education*, 20(1), 77-97. <https://doi.org/10.1016/j.tate.2003.10.002>
- Koski, J. E., Xie, H., & Olson, I. R. (2015). Understanding social hierarchies: The neural and psychological foundations of status perception. *Social Neuroscience*, 10(5), 527-550. <https://doi.org/10.1080/17470919.2015.1013223>

- Kretschmer, T., Veenstra, R., Deković, M., & Oldehinkel, A. J. (2017). Bullying development across adolescence, its antecedents, outcomes, and gender-specific patterns. *Development and Psychopathology*, 29(3), 941-955.
- 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(9), 1174-1193. <https://doi.org/10.1080/01443410.2018.1530735>
- Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374. <https://www.jstor.org/stable/2529786>
- Laninga-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(4), 1265–1283. <https://doi.org/10.1111/cdev.12650>
- Lansu, T. A., & Troop-Gordon, W. (2017). Affective associations with negativity: Why popular peers attract youths' visual attention. *Journal of Experimental Child Psychology*, 162, 282-291. <https://doi.org/10.1016/j.jecp.2017.05.010>
- *Leadbeater, B. J., Thompson, K., & Sukhawathanakul, P. (2016). Enhancing social responsibility and prosocial leadership to prevent aggression, peer victimization, and emotional problems in elementary school children. *American Journal of Community Psychology*, 58(3-4), 365-376. <https://doi.org/10.1002/ajcp.12092>
- *Leadbeater, B.J., Thompson, K., Sukhawathanakul, P., & Merrin, G. J. (2018). How program users enhance fidelity: implementing the WITS programs in rural Canadian elementary schools. *Prevention Science*, 19(8), 1066-1078. <https://doi.org/10.1007/s11121-018-0948-8>
- Leadbeater, B., & Sukhawathanakul, P. (2011). Multicomponent programs for reducing peer victimization in early elementary school: A longitudinal evaluation of the WITS primary program. *Journal of Community Psychology*, 39(5), 606-620. <https://doi.org/10.1002/jcop.20447>
- Lereya, S., Copeland, W., Costello, E., Wolke, D. (2015). Adult mental health consequences of peer bullying and maltreatment in childhood: two cohorts in two countries. *Lancet Psychiatry*, 2, 524-531. [http://dx.doi.org/10.1016/S2215-0366\(15\)00165-0](http://dx.doi.org/10.1016/S2215-0366(15)00165-0)
- Li, D., Li, X., Wang, Y., Zhao, L., Bao, Z., & Wen, F. (2013). School connectedness and problematic Internet use in adolescents: A moderated mediation model of deviant peer affiliation and self-control. *Journal of abnormal child psychology*, 41(8), 1231-1242. <https://doi.org/10.1007/s10802-013-9761-9>
- 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(4), 355-387. <https://doi.org/10.1177/1745691614535216>
- Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford press.
- Lösel, F. & Bender, D. (2011). Emotional and antisocial outcomes of bullying and victimization at school: A follow-up from childhood to adolescence. *Journal of*

- Aggression, Conflict and Peace Research*, 3(2), 89–96. <https://doi.org/10.1108/17596591111132909>
- *Low, S., Van Ryzin, M. J., Brown, E. C., Smith, B. H., & Haggerty, K. P. (2014). Engagement matters: lessons from assessing classroom implementation of steps to respect: a bullying prevention program over a one-year period. *Prevention Science*, 15(2), 165–176. <https://doi.org/10.1007/s11121-012-0359-1>
- *Matischek-Jauk, M., Krammer, G., & Reicher, H. (2018). The life-skills program Lions Quest in Austrian schools: implementation and outcomes. *Health Promotion International*, 33(6), 1022–1032. <https://doi.org/10.1093/heapro/dax050>
- Mazzone, A., Yanagida, T., CS Caravita, S., & Strohmeier, D. (2019). Moral emotions and moral disengagement: Concurrent and longitudinal associations with aggressive behavior among early adolescents. *Journal of Early Adolescence*, 39(6), 839–863. <https://doi.org/10.1177/0272431618791276>
- McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, 28(2), 295–314. <https://doi.org/10.1007/s10648-014-9287-x>
- 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(1), 124–143. <https://doi.org/10.13110/merrpalmquar1982.61.1.0124>
- Menesini, E., & Salmivalli, C. (2017). Bullying in schools: the state of knowledge and effective interventions. *Psychology, Health & Medicine*, 22(sup1), 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>
- Modecki, K.L., Minchin, J., Harbaugh, A.G., Guerra, N.G., & Runions, K.C. (2014). Bullying prevalence across contexts: A meta-analysis measuring cyber and traditional bullying. *Journal of Adolescent Health*, 55, 602–611. <http://dx.doi.org/10.1016/j.jadohealth.2014.06.007>
- Monks, C. P., Robinson, S., & Worlidge, P. (2012). The emergence of cyberbullying: A survey of primary school pupils' perceptions and experiences. *School Psychology International*, 33(5), 477–491. <https://doi.org/10.1177/0143034312445242>
- Nickerson, A. B., Tullidge, J., Manges, M., Kesselring, S., Parks, T., Livingston, J. A., & Dudley, M. (2019). Randomized controlled trial of the Child Protection Unit: Grade and gender as moderators of CSA prevention concepts in elementary students. *Child Abuse & Neglect*, 96, 104101. <https://doi.org/10.1016/j.chiabu.2019.104101>
- Nishina, A., & Juvonen, J. (2005). Daily reports of witnessing and experiencing peer harassment in middle school. *Child Development*, 76(2), 435–450. <https://doi.org/10.1111/j.1467-8624.2005.00855.x>
- Nolan, J. M. (2011). The cognitive ripple of social norms communications. *Group Processes & Intergroup Relations*, 14(5), 689–702. <https://doi.org/10.1177/1368430210392398>
- Obermann, M. L. (2011). Moral disengagement in self-reported and peer-nominated school bullying. *Aggressive Behavior*, 37(2), 133–144. <https://doi.org/10.1002/ab.20378>

- Obermann, M. L. (2013). Temporal aspects of moral disengagement in school bullying: Crystallization or escalation?. *Journal of School Violence*, 12(2), 193-210. <https://doi.org/10.1080/15388220.2013.766133>
- OECD (2013), *Education at a Glance 2013*, Chart D2.1, available at <http://dx.doi.org/10.1787/888932851668>.
- Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology*, 80(3), 381–396. <https://doi.org/10.1037//0022-3514.80.3.381>
- Olthof, T., Goossens, F. A., Vermande, M. M., Aleva, E. A., & van der Meulen, M. (2011). Bullying as strategic behavior: Relations with desired and acquired dominance in the peer group. *Journal of School Psychology*, 49(3), 339-359. <https://doi.org/10.1016/j.jsp.2011.03.003>
- Olweus, D. (1994). Bullying at school. In *Aggressive behavior*(pp. 97-130). Springer, Boston, MA.
- *Ossa, F. C., Jantzer, V., Eppelmann, L., Parzer, P., Resch, F., & Kaess, M. (2020). Effects and moderators of the Olweus bullying prevention program (OBPP) in Germany. *European Child & Adolescent Psychiatry*, 1-10. <https://doi.org/10.1007/s00787-020-01647-9>
- Paciello, M., Fida, R., Tramontano, C., Lupinetti, C., & Caprara, G. V. (2008). Stability and change of moral disengagement and its impact on aggression and violence in late adolescence. *Child Development*, 79(5), 1288-1309. <https://doi.org/10.1111/j.1467-8624.2008.01189.x>
- *Pack, C., White, A., Raczynski, K., & Wang, A. (2011). Evaluation of the safe school ambassadors program: A student-led approach to reducing mistreatment and bullying in schools. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 84(4), 127-133. <https://doi.org/10.1080/00098655.2011.564974>
- 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(6), 899-915. <https://doi.org/10.1037/a0030015>
- Peeters, M., Cillessen, A. H., & Scholte, R. H. (2010). Clueless or powerful? Identifying subtypes of bullies in adolescence. *Journal of Youth and Adolescence*, 39(9), 1041-1052. <https://doi.org/10.1007/s10964-009-9478-9>
- 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(7), 913–920. <https://doi.org/10.1037/a0039287>
- Pellegrini, A. D., & Long, J. D. (2002). A longitudinal study of bullying, dominance, and victimization during the transition from primary school through secondary school. *British journal of developmental psychology*, 20(2), 259-280. <https://doi.org/10.1348/026151002166442>
- Pelton, J., Gound, M., Forehand, R., & Brody, G. (2004). The moral disengagement scale: Extension with an American minority sample. *Journal of Psychopathology and Behavioral Assessment*, 26(1), 31-39. <https://doi.org/10.1023/B:JOBA.0000007454.34707.a5>

- 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, Cham.
- 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(5), 703-722. <https://doi.org/10.1177/1368430210398004>
- Perren, S., Gutzwiller-Helfenfinger, E., Malti, T., & Hymel, S. (2012). Moral reasoning and emotion attributions of adolescent bullies, victims, and bully-victims. *British Journal of Developmental Psychology*, 30(4), 511-530. <https://doi.org/10.1111/j.2044-835X.2011.02059.x>
- *Polanin, J. R., & Espelage, D. L. (2015). Using a meta-analytic technique to assess the relationship between treatment intensity and program effects in a cluster-randomized trial. *Journal of Behavioral Education*, 24(1), 133-151. <https://doi.org/10.1007/s10864-014-9205-9>
- Pornari, C. D., & Wood, J. (2010). Peer and cyber aggression in secondary school students: The role of moral disengagement, hostile attribution bias, and outcome expectancies. *Aggressive Behavior*, 36(2), 81-94. <https://doi.org/10.1002/ab.20336>
- Pöyhönen, V., Juvonen, J., & Salmivalli, C. (2010). What does it take to stand up for the victim of bullying? The interplay between personal and social factors. *Merrill-Palmer Quarterly* (1982-), 143-163.
- Pozzoli, T., Gini, G., & Vieno, A. (2012). The role of individual correlates and class norms in defending and passive bystanding behavior in bullying: A multilevel analysis. *Child Development*, 83(6), 1917-1931. <https://doi.org/10.1111/j.1467-8624.2012.01831.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., & Kelley, K. (2011). Effect size measures for mediation models: quantitative strategies for communicating indirect effects. *Psychological Methods*, 16(2), 93. <https://doi.org/10.1037/a0022658>
- 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.
- Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology*, 64(2), 243.
- Prentice, D. A., & Miller, D. T. (1996). Pluralistic ignorance and the perpetuation of social norms by unwitting actors. In *Advances in Experimental Social Psychology* (Vol. 28, pp. 161-209). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60238-5](https://doi.org/10.1016/S0065-2601(08)60238-5)

- Reilly, D. W., & Wood, M. D. (2008). A randomized test of a small-group interactive social norms intervention. *Journal of American College Health*, 57(1), 53-60.
- *Renshaw, T. L., & Jimerson, S. R. (2012). Enhancing student attitudes via a brief, universal-level bullying prevention curriculum. *School Mental Health*, 4(2), 115-128. <https://doi.org/10.1007/s12310-011-9069-2>
- Ribeaud, D., & Eisner, M. (2010). Are moral disengagement, neutralization techniques, and self-serving cognitive distortions the same? Developing a unified scale of moral neutralization of aggression. *International Journal of Conflict and Violence (IJCV)*, 4(2), 298-315. <https://doi.org/10.4119/ijcv-2833>
- 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>
- Rivers, I., Poteat, V. P., Noret, N., & Ashurst, N. (2009). Observing bullying at school: The mental health implications of witness status. *School Psychology Quarterly*, 24(4), 211-223. <https://doi.org/10.1037/a0018164>
- 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(6), 359-371. <https://doi.org/10.1111/j.1751-9004.2011.00355.x>
- Ruggieri, S., Friemel, T., Sticca, F., Perren, S., & Alsaker, F. (2013). Selection and influence effects in defending a victim of bullying: The moderating effects of school context. *Procedia-Social and Behavioral Sciences*, 79, 117-126. <https://dx.doi.org/10.1016/j.sbspro.2013.05.060>
- 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>
- Saarento, S., Kärnä, A., Hodges, E. V., & Salmivalli, C. (2013). Student-, classroom-, and school-level risk factors for victimization. *Journal of School Psychology*, 51(3), 421-434. <https://doi.org/10.1016/j.jsp.2013.02.002>
- 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(1), 61-76. <https://doi.org/10.1007/s10802-013-9841-x>
- Sacco, M., Smits, W., Kavadias, D., Spruyt, B., & d'Andrimont, C. (2016). Jeunes bruxelloises : entre diversité et précarité. *Brussels Studies (BSI synopsis #98)* <https://doi.org/10.4000/brussels.1339>
- Sainio, M., Herkama, S., Turunen, T., Rönkkö, M., Kontio, M., Poskiparta, E., & Salmivalli, C. (2020). Sustainable antibullying program implementation: School profiles and predictors. *Scandinavian Journal of Psychology*, 61(1), 132-142. <https://doi.org/10.1111/sjop.12487>
- Salmivalli, C. (2001). Group view on victimization. *Peer Harassment in School*, 398-417.
- Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and Violent Behavior*, 15(2), 112-120. <https://doi.org/10.1016/j.avb.2009.08.007>

- Salmivalli, C. & Hymel, S. (2019, June 4-6) *Bullying through the lens of morality and peer dynamics*. [Keynote address] World Anti-Bullying Forum 2019, Dublin, Ireland. <https://worldantibullyingforum.com/videos/christina-salmivalli-and-shelley-hymel/>
- Salmivalli, C., & Kaukiainen, A. (2004). "Female aggression" revisited: Variable-and person-centered approaches to studying gender differences in different types of aggression. *Aggressive Behavior*, 30(2), 158-163. <https://doi.org/10.1002/ab.20012>
- *Salmivalli, C., Kaukiainen, A., & Voeten, M. (2005). Anti-bullying intervention: Implementation and outcome. *British Journal of Educational Psychology*, 75(3), 465-487. <https://doi.org/10.1348/000709905X26011>
- 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), 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. & Peets, K. (2018). Bullying and victimization. In K. Rubin, W. Bukowski, & B. Laursen (Eds.), *Handbook of peer interactions, relationships, and groups*. (2nd ed., pp. 322-340) Guilford.
- Salmivalli, C., Poskiparta, E., Ahtola, A., & Haataja, A. (2013). The Implementation and Effectiveness of the KiVa Antibullying Program in Finland. *European Psychologist*, 18(2), 79-88. <https://doi.org/10.1027/1016-9040/a000140>
- 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(3), 217-225. <https://doi.org/10.1080/01973533.2010.495645>
- 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(2-3), 196-215. <https://doi.org/10.1080/15534510.2011.651302>
- Sandstrom, M. J., & Cillessen, A. H. (2006). Likeable versus popular: Distinct implications for adolescent adjustment. *International Journal of Behavioral Development*, 30(4), 305-314. <https://doi.org/10.1177/0165025406072789>
- Savahl, S., Montserrat, C., Casas, F., Adams, S., Tiliouine, H., Benninger, E., & Jackson, K. (2019). Children's experiences of bullying victimization and the influence on their subjective well-being: A multinational comparison. *Child Development*, 90(2), 414-431. <https://doi.org/10.1111/cdev.13135>
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74. <http://www.mpr-online.de>
- Schroeder, C. M., & Prentice, D. A. (1998). Exposing Pluralistic Ignorance to Reduce Alcohol Use Among College Students 1. *Journal of Applied Social Psychology*, 28(23), 2150-2180.
- *Schultes, M. T., Stefanek, E., van de Schoot, R., Strohmeier, D., & Spiel, C. (2014). Measuring implementation of a school-based violence prevention program. *Zeitschrift für Psychologie*. <https://doi.org/10.1027/2151-2604/a000165>

- *Sharma, D., Mehari, K. R., Kishore, J., Sharma, N., & Duggal, M. (2020). Pilot evaluation of Setu, a school-based violence prevention program among Indian adolescents. *The Journal of Early Adolescence*, 40(8), 1142-1166. <https://doi.org/10.1177/0272431619899480>
- Shelton, J. N., & Richeson, J. A. (2005). Intergroup contact and pluralistic ignorance. *Journal of Personality and Social Psychology*, 88(1), 91. <https://doi.org/10.1037/0022-3514.88.1.91>
- 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(1), 57-67. <https://doi.org/10.1002/ab.20282>
- Sjögren, B., Thornberg, R., Wänström, L., & Gini, G. (2020). Bystander behaviour in peer victimisation: moral disengagement, defender self-efficacy and student-teacher relationship quality. *Research Papers in Education*, 1-23. <https://doi.org/10.1080/02671522.2020.1723679>
- Spilt, J. L., Koot, J. M., & van Lier, P. A. (2013). For whom does it work? Subgroup differences in the effects of a school-based universal prevention program. *Prevention Science*, 14(5), 479-488. <https://doi.org/10.1007/s11121-012-0329-7>
- Sticca, F., & Perren, S. (2015). The chicken and the egg: Longitudinal associations between moral deficiencies and bullying: A parallel process latent growth model. *Merrill-Palmer Quarterly*, 61(1), 85-100. <https://doi.org/10.13110/merrpalmquar1982.61.1.0085>
- Sutton, J., & Smith, P. K. (1999). Bullying as a group process: An adaptation of the participant role approach. *Aggressive Behavior*, 25(2), 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)
- Swearer, S. M., Espelage, D. L., & Napolitano, S. A. (2009). *Bullying prevention and intervention: Realistic strategies for schools*. Guilford press.
- Swift, L. E., Hubbard, J. A., Bookhout, M. K., Grasseti, S. N., Smith, M. A., & Morrow, M. T. (2017). Teacher factors contributing to dosage of the KiVa anti-bullying program. *Journal of School Psychology*, 65, 102-115. <https://doi.org/10.1016/j.jsp.2017.07.005>
- Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social Issues and Policy Review*, 10(1), 181-211. <https://doi.org/10.1111/sipr.12022>
- Teglasi, H., & Rothman, L. (2001). Stories a classroom-based program to reduce aggressive behavior. *Journal of School Psychology*, 39(1), 71-94. [https://doi.org/10.1016/S0022-4405\(00\)00060-1](https://doi.org/10.1016/S0022-4405(00)00060-1)
- Thornberg, R., & Jungert, T. (2013). Bystander behavior in bullying situations: Basic moral sensitivity, moral disengagement and defender self-efficacy. *Journal of Adolescence*, 36(3), 475-483. <https://doi.org/10.1016/j.adolescence.2013.02.003>
- Thornberg, R., & Jungert, T. (2014). School bullying and the mechanisms of moral disengagement. *Aggressive Behavior*, 40(2), 99-108. <https://doi.org/10.1002/ab.21509>
- 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>
- Thornberg, R., Wänström, L., Gini, G., Varjas, K., Meyers, J., Elmelid, R., ... & Mellander, E. (2021). Collective moral disengagement and its associations with bullying perpetration and victimization in students. *Educational Psychology*, 1-15. <https://doi.org/10.1080/01443410.2020.1843005>
- Thornberg, R., Wänström, L., & Hymel, S. (2019). Individual and classroom social-cognitive processes in bullying: A short-term longitudinal multilevel study. *Frontiers in Psychology*, 10, article 1752. <https://doi.org/10.3389/fpsyg.2019.01752>
- Tickle, L. (1999). Teacher self-appraisal and appraisal of self. In R. P. Lipka & T. M. Brinthaupt (Eds.) *The role of self in teacher development*, pp. 121-141. State University of New York Press.
- 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 ? [Can redesigning the school playground and setting up classroom forums help to reduce bullying?] *Revue française de Pédagogie*, 209, 129-147. <https://doi.org/10.4000/rfp.9846>
- Tolmatcheff, C., Hénoumont, F., Klée, E., & Galand, B. (2019). Stratégies et réactions des victimes et de leur entourage face au harcèlement scolaire : une étude rétrospective [Strategies and reactions of school bullying victims and their entourage : A retrospective study] *Psychologie française*, 64(4), 391-407. <https://doi.org/10.1016/j.psfr.2018.09.002>
- Troop-Gordon, W., Frosch, C. A., Totura, C. M. W., Bailey, A. N., Jackson, J. D., & Dvorak, R. D. (2019). Predicting the development of pro-bullying bystander behavior : A short-term longitudinal analysis. *Journal of School Psychology*, 77, 77-89. <https://doi.org/10.1016/j.jsp.2019.10.004>
- Troop-Gordon et al. (2019). Predicting the development of pro-bullying bystander behavior : A short-term longitudinal analysis. *Journal of School Psychology*, 77, 77-89. <https://doi.org/10.1016/j.jsp.2019.10.004>
- Troy, A. S., Shallcross, A. J., & Mauss, I. B. (2013). A person-by-situation approach to emotion regulation: Cognitive reappraisal can either help or hurt, depending on the context. *Psychological Science*, 24(12), 2505-2514. <https://doi.org/10.1177/0956797613496434>
- *Twemlow, S. W., Biggs, B. K., Nelson, T. D., Vernberg, E. M., Fonagy, P., & Twemlow, S. W. (2008). Effects of participation in a martial arts-based antibullying program in elementary schools. *Psychology in the Schools*, 45(10), 947-959. <https://doi.org/10.1002/pits.20344>
- Vaillancourt, T., Hymel, S., & McDougall, P. (2003). Bullying is power: Implications for school-based intervention strategies. *Journal of Applied School Psychology*, 19(2), 157-176. https://doi.org/10.1300/J008v19n02_10
- Valente, T. W., & Pitts, S. R. (2017). An appraisal of social network theory and analysis as applied to public health: challenges and opportunities. *Annual Review of Public Health*, 38, 103-118. <https://doi.org/10.1146/annurev-publhealth-031816-044528>

- Van der Velden, F., Brugman, D., Boom, J., & Koops, W. (2010). Effects of EQUIP for educators on students' self-serving cognitive distortions, moral judgment, and antisocial behavior. *Journal of Research in Character Education*, 8(1), 77-95.
- van Verseveld, 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*.
- Van der Wal, M. F., De Wit, C. A., & Hirasing, R. A. (2003). Psychosocial health among young victims and offenders of direct and indirect bullying. *Pediatrics*, 111(6), 1312-1317. <http://pediatrics.aappublications.org/content/111/6/1312.full.html>
- Veenstra, R., Dijkstra, J. K., & Kreager, D. A. (2018). Pathways, networks, and norms: A sociological perspective on peer research. In W. M. Bukowski, B. Laursen, & K. H. Rubin (Eds.), *Handbook of peer interactions, relationships, and groups (2nd edition)* (pp. 45–63). Guilford.
- Veenstra, R., Lindenberg, S., Huitsing, G., Sainio, M., & Salmivalli, C. (2014). The role of teachers in bullying: The relation between antibullying attitudes, efficacy, and efforts to reduce bullying. *Journal of Educational Psychology*, 106, 1135–1143. <https://doi.org/10.1037/a0036110>
- Veenstra, R. (2021). The need for a 'we-culture': The importance of the larger network and social norms for tackling bullying. In: B. Johansson & R. Thornberg (eds.) *Always Take Action: Researchers on Their Results and Children's Voices on the Journey from Bullied to Acknowledged (pp. 150-163)*. Stockholm, SWE: Friends.
- 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 Classroom Level Link between Social Norms and Behavioral Development. *International Journal of Behavioral Development*.
- 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>
- Wachs, S. (2012). Moral disengagement and emotional and social difficulties in bullying and cyberbullying: Differences by participant role. *Emotional and Behavioural Difficulties*, 17(3-4), 347-360. <http://dx.doi.org/10.1080/13632752.2012.704318>
- Walden, B., McGue, M., Burt, S. A., & Elkins, I. (2004). Identifying shared environmental contributions to early substance use: the respective roles of peers and parents. *Journal of abnormal psychology*, 113(3), 440. <https://doi.org/10.1037/0021-843X.113.3.440>
- Walters, G. D. (2020). School-age bullying victimization and perpetration: A meta-analysis of prospective studies and research. *Trauma, Violence, & Abuse*. <https://doi.org/10.1177/1524838020906513>
- 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(9), 918-931. <https://doi.org/10.1002/pits.22042>

- Wang, X., Yang, J., Wang, P., Zhang, Y., Li, B., Xie, X., & Lei, L. (2020). Deviant peer affiliation and bullying perpetration in adolescents: The mediating role of moral disengagement and the moderating role of moral identity. *The Journal of psychology*, 154(3), 199-213. <https://doi.org/10.1080/00223980.2019.1696733>
- Weare, K., & Nind, M. (2011). Mental health promotion and problem prevention in schools: what does the evidence say?. *Health promotion international*, 26(suppl_1), i29-i69. <https://doi.org/10.1093/heapro/dar075>
- Wentzel, K. R. (2003). Motivating students to behave in socially competent ways. *Theory Into Practice*, 42(4), 319-326. https://doi.org/10.1207/s15430421tip4204_9
- Wilson, S. J., & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior: Update of a meta-analysis. *American journal of preventive medicine*, 33(2), S130-S143. <https://doi.org/10.1016/j.amepre.2007.04.011>
- Wolke, D., & Lereya, S. T. (2015). Long-term effects of bullying. *Archives of Disease in Childhood*, 100(9), 879-885. <http://dx.doi.org/10.1136/archdischild-2014-306667>
- 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(S1), 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>
- Yoon, J., & Bauman, S. (2014). Teachers: A critical but overlooked component of bullying prevention and intervention. *Theory Into Practice*, 53(4), 308-314. <https://doi.org/10.1080/00405841.2014.947226>
- Yoon, J. S., & Kerber, K. (2003). Bullying: Elementary teachers' attitudes and intervention strategies. *Research in Education*, 69(1), 27-35.
- Yoon, J., Sulkowski, M. L., & Bauman, S. A. (2016). Teachers' responses to bullying incidents: Effects of teacher characteristics and contexts. *Journal of school violence*, 15(1), 91-113. <https://doi.org/10.1080/15388220.2014.963592>
- Yun, H. Y. (2020). New approaches to defender and outsider roles in school bullying. *Child Development*, 91(4), e814-e832. <https://doi.org/10.1111/cdev.13312>