Journal of School Psychology 69 (2018) 56–72

Contents lists available at ScienceDirect

Journal of School Psychology

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

Evaluation of the Olweus Bullying Prevention Program: A large scale study of U.S. students in grades 3–11

Susan P. Limber[a,*,1], Dan Olweus[b,1], Weijun Wang[c], Matthew Masiello[d], Kyrre Breivik[e]

[a] Department of Youth, Family & Community Studies, 2038 Barre Hall, Clemson University, Clemson, SC 29634, United States
[b] Psykologkonsult Dan Olweus, Vognstolbakken 16, NO-5096 Bergen, Norway
[c] The State University of New York at Buffalo, Research Institute on Addictions. University at Buffalo, State University of New York, 1021 Main Street, Buffalo, NY 14203, United States
[d] University of Massachusetts Memorial Children's Medical Center, Health Alliance Hospital, 60 Hospital Rd., Leominster, MA 01453, United States
[e] Regional Centre for Child and Youth Mental Health and Child Welfare, Uni Research Health, Nygårdsgaten 112-114, 5008 Bergen, Norway

ARTICLE INFO

Action Editor: Michelle K. Demaray

Keywords:
Bullying victimization
Bullying perpetration
Evaluation
Olweus Bullying Prevention Program
Anti-bullying programs

ABSTRACT

The purpose of this quasi-experimental study was to evaluate a large-scale implementation of the Olweus Bullying Prevention Program with children and youth in grades 3–11 in the U.S. Two major sets of analyses are presented, one following 210 schools over two years (Study 1; $n = 70{,}998$ at baseline) and the other following a subsample of 95 schools over three years (Study 2; $n = 31{,}675$ at baseline). Schools were located in 49 counties in central and western Pennsylvania. The Mplus 7.0 program was used to analyze the data which had a multilevel structure, with students nested in schools and program effects based on school-aggregated outcome variables. For almost all grades, there were clear reductions in the two key dimensions, being bullied and bullying other students. Average Absolute Change amounted to approximately 3%, implying that almost 2000 students had escaped being bullied in the two-year study. School-level Cohen's *d*'s were large or fairly large. The longitudinal analyses documented increases in students' expressions of empathy with bullied peers, marked decreases in their willingness to join in bullying, and perceptions that their primary teacher had increased his or her efforts to address bullying. Overall, effects were stronger the longer the program had been in place. The analyses provided strong support for the effectiveness of the OBPP with U.S. students in elementary, middle, and early high school grades in Pennsylvania schools. Future research is warranted to assess program effectiveness in different racial/ethnic and community settings and to examine the relation between fidelity of implementation and outcomes.

Bullying is an ancient phenomenon, yet systematic research on the nature and prevalence of bullying and efforts to prevent bullying are relatively recent. The earliest study on bullying was conducted in the 1970s in Scandinavia (Olweus, 1973, 1978), and the first systematic attempts to prevent bullying also began in Scandinavia. In 1983, the Norwegian Ministry of Education launched a nationwide campaign to address bullying in schools, in response to public concerns that were fueled by the suicides of three teenagers who allegedly had been severely bullied by their peers. An initial version of what later became known as the Olweus Bullying

* Corresponding author.
E-mail addresses: slimber@clemson.edu (S.P. Limber), olweus@uni.no (D. Olweus), wwang@ria.buffalo.edu (W. Wang), mmasiello@healthalliance.com (M. Masiello), Kyrre.breivik@uni.no (K. Breivik).
[1] The first two authors contributed equally to this article.

https://doi.org/10.1016/j.jsp.2018.04.004
Received 24 February 2017; Received in revised form 5 January 2018; Accepted 25 April 2018

Prevention Program (OBPP) was developed, implemented, and evaluated within this same context (Olweus, 1991, 1993; Olweus & Limber, 2010b). Although international research on bullying grew slowly in the 1980s and 1990s, it was not until the late 1990s and 2000s that attention to bullying exploded among researchers, policy makers, and the general public in the U.S. and in many other countries (Berger, 2007). Today, bullying is commonly recognized as a serious public health problem affecting children and adolescents in the U.S. and around the world (Masiello & Schroeder, 2014; National Academies of Sciences, Engineering, and Medicine, 2016; Office of the Special Representative of the Secretary-General on Violence against Children, 2016).

Bullying is a subset of aggressive behavior that is commonly repeated and involves a power imbalance between a target and his or her perpetrator(s) (Gladden, Vivolo-Kantor, Hamburger, & Lumpkin, 2014; Olweus, 1993). In the U.S., nearly one-quarter of students ages 12–18 report having been bullied during the school year (Zhang, Musu-Gillette, & Oudekerk, 2016) and 14% of 3rd-12th graders reported having been bullied 2–3 times per month or more (Luxenberg, Limber, & Olweus, 2015). The many short- and long-term negative psychosocial, mental health, physiological, and behavioral effects of being bullied have been thoroughly documented and will not be reiterated here (for overviews, see Cook, Williams, Guerra, Kim, & Sadek, 2010; National Academies of Sciences, Engineering, and Medicine, 2016; Olweus, 2013; Ttofi, Farrington, Lösel, & Loeber, 2011a). Similarly, bullying others has been associated with a number of short- and long-term negative characteristics, but of a different nature than those of individuals who have been bullied (for overviews, see Cook et al., 2010; National Academies of Sciences, Engineering, and Medicine, 2016; Olweus, 2013; Ttofi, Farrington, & Lösel, 2012). Common consequences of being bullied by peers are largely internalizing, and include depression, poor self-esteem, and suicidal ideation, whereas children and youth who bully peers are characterized by externalizing problems, such as violence, rule breaking behavior, and delinquency.

As Olweus (1993) emphasized at an early stage, bullying is not only a health problem but also a serious violation of fundamental human rights. In recognition of children's rights to be safe in school, and in response to concerns about the negative human and societal effects of bullying, a number of school-based bullying prevention efforts have been launched in recent years (National Academies of Sciences, Engineering, and Medicine, 2016). Several meta-analyses and/or systematic reviews of these prevention programs have been conducted, with somewhat mixed results (for a review, see Ttofi, Eisner, & Bradshaw, 2014). The most comprehensive meta-analysis to date was conducted by Ttofi and Farrington (2009, 2011). In this analysis, which included 44 evaluations of school-based intervention programs, the authors concluded that anti-bullying programs were effective in reducing bullying and/or victimization by an average of 17–23% but that the effect sizes were relatively small. There was great variation in program effects, however, and the authors noted that programs implemented in Europe were more effective than those implemented in the U.S. The authors also observed that “programs inspired by the work of Dan Olweus worked best” (Ttofi & Farrington, 2011, pp. 41–42). Indeed, while the average Odds Ratio for all included studies was 1.36 for bullying perpetration and 1.29 for bullying victimization, the (unweighted) average for evaluations of the Olweus Bullying Prevention Program was markedly higher, 1.83 and 1.80, respectively (Ttofi & Farrington, 2011, Table 3, pp. 36–37).

1. Brief description of the OBPP

The Olweus Bullying Prevention Program (OBPP; Olweus, 1991, 1993, Olweus & Limber, 2010a, 2010b) is the oldest and one of the most researched bullying prevention programs in the world. It represents a whole-school comprehensive approach that includes schoolwide, classroom, individual, and community components. The program is focused on both short-term and long-term change that will create a safe and positive school environment. The overarching goals of the OBPP are to reduce existing bullying problems among students, prevent new bullying problems, and achieve better peer relations (Olweus, 1993; Olweus & Limber, 2010b). These goals are pursued by restructuring the school environment to reduce opportunities and rewards for bullying behavior and to build a sense of community. The program was designed and initially evaluated for use in elementary, middle, and junior high schools.

The OBPP is built on four basic principles. Adults at school should: (a) show warmth and positive interest in students; (b) set firm limits to unacceptable behavior; (c) use consistent positive consequences to acknowledge and reinforce appropriate behavior and non-physical, non-hostile consequences when rules are broken; and (d) function as authorities and positive role models (Olweus, 1993; Olweus et al., 2007). These principles have been translated into interventions at the school, classroom, individual, and community levels (Olweus, 1993; Olweus & Limber, 2010a, 2010b). Parent involvement is encouraged at all levels (Olweus & Limber, 2010b). There are eight school-level components, which are implemented school-wide, including the development of a Bullying Prevention Coordinating Committee (BPCC, a building-level coordinating team of administrators, teachers, non-teaching staff, and parents that is responsible for ensuring that all components of the OBPP are implemented with fidelity); the yearly administration of the Olweus Bullying Questionnaire (OBQ, Olweus, 2007a); training and ongoing consultation for members of the coordinating team and all school staff; the adoption of clear rules and policies about bullying; and the review and refinement of the school's supervisory system to reduce bullying. Classroom-level components of the program include holding regular class meetings to build understanding of bullying and related issues through discussions and role play, and build class cohesion; posting and enforcing school-wide rules against bullying; and holding periodic class-level meetings with parents. Teachers are also encouraged to integrate bullying prevention messages and strategies throughout the curriculum. Individual-level components of the OBPP include supervision of students' activities, particularly in known hot-spots for bullying; training for all staff to help them intervene on-the-spot when bullying occurs or is suspected; and follow-up interventions with children and youth involved in bullying. Finally, BPCC members are encouraged to involve one or more community members on their team, look for ways that community members can support the school's program, and collaborate to spread bullying prevention strategies and messages into other community settings that involve children and youth (Olweus & Limber, 2010b).

School staff are supported in their implementation by manuals for teachers and for members of the BPCC, class meeting resource

guides, and videos. Training and ongoing consultation are provided by certified OBPP Trainer-Consultants, who help schools address challenges and maintain fidelity to the model. As described in detail elsewhere (Limber & Olweus, 2017; Olweus & Limber, 2010b), although the basic principles and core components of the OBPP have remained largely unchanged since its development and initial implementation in Norway, research and extensive experience have naturally led to some adaptations of the program to the U.S. educational context, including the current implementation and evaluation. Such adaptations include development of English language print and video resources for administrators, teachers, and parents; the organization of training of trainer-consultants and school teams; and the use of readiness checklists.

1.1. Evaluations of the OBPP in Norway

The first evaluation of the OBPP, which took place in Bergen, Norway, followed 2500 students in grades 5–8 in 42 schools over a period of two-and-a-half years (between 1983 and 1985) (Olweus, 1991, 1993, 1997). Using an "extended age cohort design" (Olweus, 2005, and below), in which same-aged students were compared across time, Olweus documented marked reductions in students' self-reports of being bullied and bullying others, and significant reductions in teachers' and students' ratings of bullying among students within the classroom. In addition, there were significant improvements in several measures of school/classroom climate, including improvements in students' satisfaction with school life, improved order and discipline, more positive social relationships, and a more positive attitude towards school (1991, 1993, 1997; see also summaries of findings in Olweus & Limber, 2010a, 2010b). An indicator of fidelity of program implementation was significantly related to program outcomes (Olweus & Kallestad, 2010).

Subsequent to this initial study, six additional large-scale evaluations of the OBPP have been carried out in Norway, involving > 30,000 students from > 300 schools (Olweus, 2005; Olweus & Limber, 2010b). Findings have revealed consistently positive program effects among students in grades 4–7, typically with reductions in bullying problems in the 35–50% range after eight months of intervention (Olweus & Limber, 2010b). Although positive findings have also been obtained with students in grades 8–10, results have been less consistent and it has taken longer to achieve as strong effects as with younger students (Olweus & Limber, 2010b). In likely the first study of long-term effects of a program, Olweus followed students from 14 schools in Oslo (with approximately 3000 students at each assessment) and observed reductions in self-reports of victimization of 40% and self-reported bullying of 51% over a period of five years (Olweus & Limber, 2010b).

1.2. Evaluations of the OBPP in the U.S.

The effectiveness of OBPP has been evaluated in two relatively large-scale studies in the U.S. The first took place in the mid-1990s and involved elementary and middle schools in six rural school districts in South Carolina with high percentages of African American students (districts ranged from 46% to 95% African American; Limber, Nation, Tracy, Melton, & Flerx, 2004; Olweus & Limber, 2010b). After seven months of program implementation, significant differences between intervention and comparison schools were observed with regard to students' reports of bullying other students, self-reported delinquency, vandalism, school misbehavior, and sanctions for school misbehavior. There were no significant program effects for students' reports of being bullied, however. Bauer, Lozano, and Rivara (2007) used a nonrandomized controlled design to evaluate the OBPP with seven intervention and three control schools in Washington State. Findings revealed significant program effects for relational and physical victimization among white students but not among students of other races/ethnic backgrounds (see Olweus & Limber, 2010b for more details). Although these U.S. findings are clearly encouraging, it is also clear that the results from these studies in the U.S. have not been uniformly positive.

The concern about the lack of positive program effects in the U.S. is echoed in a recent review of evaluations of bullying prevention programs (Evans, Fraser, & Cotter, 2014), which was completed after the Ttofi and Farrington meta-analysis. Unfortunately, the conclusions from this review are somewhat compromised by the inclusion of four exceptionally small samples ($n < 50$) and seven evaluations (out of 22) with < 200 participants. Still, it is worth noting that six of the eight studies examining bullying victimization with nonsignificant results were conducted in the U.S., as were six of the ten nonsignificant studies examining bullying perpetration. Several researchers have expressed doubt about the usefulness and likely success of bullying prevention efforts in general, and such programs implemented outside of the U.S., in particular (e.g., Cohen, Espelage, Twemlow, Berkowitz, & Comer, 2015; Hong & Espelage, 2012).

In view of these concerns, there is obviously a marked need for a new large-scale study of the OBPP in the U.S., in which the program has been systematically implemented over a period of at least two years and data from a sample of appropriate size have been adequately analyzed with multilevel techniques taking account of cluster effects. The goal of the present article is to report the results of such a study.

1.3. Research questions and expectations

The current study addressed several research questions pertaining to changes in students' behaviors and attitudes with respect to bullying. Of primary interest was whether there could be documented systematic changes in students' reports of being bullied and bullying other students after implementation and as a consequence of the OBPP.

We expected that there would be such reductions in being bullied and bullying others, that these effects would be observed for both males and females and, tentatively (due to smaller sample sizes and a relatively large percentage of missing data on the relevant variable, see Table 1 below) also for students of the races/ethnicities included in the study. In addition, and based on experiences

Table 1
Characteristics of the participants (at baseline) from the Two-year and Three-year studies.

	2-Year study	3-Year study
	N (%)	*N* (%)
Sex		
Female	34,820 (49.0)	15,560 (49.2)
Male	35,862 (50.5)	15,937 (50.4)
Missing (sex)	316 (0.4)	123 (0.4)
Grade		
Grade 3	8636 (12.2)	4447 (14.1)
Grade 4	8586 (12.1)	4402 (13.9)
Grade 5	9161 (12.9)	4446 (14.1)
Grade 6	11,397 (16.1)	4502 (14.2)
Grade 7	12,002 (16.9)	4156 (13.1)
Grade 8	11,913 (16.8)	4185 (13.2)
Grade 9	3404 (4.8)	1990 (6.3)
Grade 10	2844 (4.0)	1555 (4.9)
Grade 11	3055 (4.3)	1937 (6.1)
Race/ethnicity		
White	44,028 (62.0)	19,291 (61.0)
Black or African American	3575 (5.0)	1432 (4.5)
Hispanic or Latino	3177 (4.5)	1006 (3.2)
Other	8165 (11,5)	3539 (11.2)
Missing data/don't know	12,053 (17.0)	6352 (20.1)

with OBPP evaluations in Norway (Olweus & Kallestad, 2010; Olweus & Limber, 2010b), we assumed effects for being bullied would be somewhat stronger for elementary and middle school-aged students compared with high school students. Consistent with the findings of Olweus (1991, 2005) and Olweus and Limber (2010b), we expected that the effects would be stronger the longer the implementation of the OBPP.

Not only did we expect to observe changes in students' bullying behaviors over time, but we also predicted changes in students' attitudes towards bullying after being exposed to the OBPP. Specifically, we anticipated that negative attitudes towards bullying—in particular, expressing a disinclination to join in bullying—would increase over time. Not surprisingly, research has found that attitudes towards bullying are strongly related to the likelihood of bullying others (Salmivalli & Voeten, 2004; Van Goethem, Scholte, & Wiers, 2010). Moreover, a focus of the OBPP is on recognizing and altering the behavior of bystanders to bullying—those who do not initiate bullying but may assume a variety of roles upon observing bullying, ranging from joining in the bullying, to watching passively, to helping to stop bullying (Olweus et al., 2007; Salmivalli, Lagerspetz, Björkqvist, Österman, & Kaukiainen, 1996). Thus, we expected to see reductions in the extent to which students believed they could join in bullying over time, for boys and girls, and across all age groups.

The current study also examined whether students expressed more empathy for bullied peers over time. A recent meta-analysis has confirmed that empathy is negatively associated with bullying others and is positively associated with defending peers from bullying (Zych, Ttofi, & Farrington, 2017). In addition, short-term longitudinal research has found that higher empathy predicted less future involvement in bullying (Stavrindes, Georgiou, & Theofanous, 2010). Several components of the OBPP—namely class meeting discussions and role playing—are intended to build a sense of community among students and increase empathy for bullied peers (Olweus & Limber, 2010b). Thus, we expected that over time, students over time would express more empathy for bullied peers.

Finally, we also expected that students would perceive increased anti-bullying activities by their teachers in their schools over time. We anticipated that with the implementation of the OBPP, obvious changes in teachers' efforts to prevent and address bullying would occur and that these behaviors would be apparent to students at all grade levels.

These research questions were addressed in two studies, which will be described in detail below. Study 1 addresses possible program effects with regard to being bullied and bullying others for a large sample of students over two years, and assesses differences in outcomes by gender and race/ethnicity. Study 2 examines longer-term effects for a sub-group of students involved in the program. In this second study, we examined changes in students' involvement in bullying involvement (being bullied or bullying others), attitudes towards bullying, empathy for bullied peers, and students' perceptions of teachers' actions to address bullying over a period of three years.

2. Method

2.1. Participants

The original sample included students in grades 3–12 who were drawn from schools that were involved in a wide-scale effort to implement the OBPP in elementary, middle, and high schools in 49 counties in southern and central Pennsylvania (Limber & Olweus, 2017). Our target population in this study consisted of students from schools that had completed successful standard implementation

of the OBPP, as determined by the minimum criterion that the school had conducted the Olweus Bullying Questionnaire (Olweus, 2007a, 2007b) at baseline and two years later. Although completion of these two surveys does not in itself represent a detailed assessment of the degree of fidelity of program implementation, the surveys are important components of the program and thereby provide a rough indication that the program had been introduced in the school in a standardized way, as recommended in the guides for program leaders (Olweus et al., 2007) and teachers (Olweus & Limber, 2007). Accordingly, 20 of the original 230 schools were removed from the analyses because they had conducted only one survey. Twelfth grade students were also omitted from the analyses because their numbers were much lower than those in other high school grades (approximately 50% of the numbers of students in grades 9–11). The resulting sample (Study 1) thus included participants who were students in grades 3–11, drawn from 210 schools, and representing three different cohorts that began to implement the program in 2008 (83 schools), 2009 (103 schools), and 2010 (24 schools). Locales for the 210 schools were primarily suburban (59%), and rural (29%), with 9% of the schools located in urban locales, and 4% located in towns (National Center for Education Statistics [NCES], 2014). A total of 70,998 students completed baseline (T0) assessments prior to implementation of the OBPP, and 67,374 (94.9%) students completed the assessments at Time 2, two years after their first assessment.

A total of 116 schools had taken the survey up to four consecutive years. Twenty-one of these schools had one or more missing data points and were excluded from the analyses. The remaining 95 schools were used to assess (relative) long-term changes over a period of three years (Study 2). The average levels of being bullied or bullying others (2 or 3 times a month or more) at baseline were only marginally different for these schools than for the larger groups of schools they were part of in the full-sample analyses: 21.6% versus 20.8% for being bullied, and 9.6% versus 10.1% for bullying others. A total of 31,620 students completed baseline assessment (T0), and 29,814 (94.1%) students in grades 3–11 completed assessments at Time 3. Demographic information for participants in the 2-year and 3-year studies is presented in Table 1.

Even though the samples in the two sets of analyses were very large by common standards in the field, it should be noted that the number of participating students from grades 9–11 was much smaller than in the lower grades, constituting only one-third or one-fourth of the average numbers in grades 3–8. This will, of course, have a clear effect in terms of significance testing. It is therefore important to use other indicators of program effects in addition to significance tests (below). Also, the samples for the non-White races/ethnic groups were relatively small. In addition, it should be noted that > 12,000 students (17%) of students in the 2-year study did not respond to the question about race/ethnic background or answered “I do not know.” We therefore decided to conduct analyses on this variable only for subgroups consisting of at least 1000 students.

Although not being part of our target population, it is worth mentioning that the 20 schools that were excluded from Study 1 comprised a total of 5980 students (7.8% of the total number) and the ethnicity and gender distributions of these students did not differ significantly from those of the target population. In the longitudinal Study 2, which contains a subsample of the participants from Study 1, the number of students in the 21 excluded schools amounted to 8001 (20.2% of the total number). In both studies, the levels of the two key bullying variables were somewhat smaller in the excluded schools (by an average of 2.5 percentage points in Study 1 and 1.6 percentage points in Study 2). Overall, the students in the excluded schools were basically similar to the students in the target population, except that their levels of being bullied and bullying other students were somewhat lower.

2.2. Measures

In the following sections, we describe the variables used in the analyses for both Study 1 and Study 2 and provide selected psychometric information about some of them (also see section Approach to analysis about intraclass correlations as reliability estimates). Participants completed the Olweus Bullying Questionnaire (OBQ), a 40-item anonymous questionnaire that assesses students' self-reports of bullying others, being bullied, their own actions and reactions when they witness bullying, their attitudes about bullying, and their perceptions of the efforts of their teachers to counteract bullying (Olweus, 2007a). Most questions ask students about their experiences during the past couple of months (Olweus, 2007a; Olweus, 2013). The OBQ is recommended for use in grades 3 and higher (Olweus et al., 2007).

2.2.1. Being bullied

Students were presented with the following detailed definition of bullying:

> We say a student is being bullied when another student, or several other students:
>
> - say mean and hurtful things, or make fun of him or her, or call him or her mean and hurtful names
> - completely ignore or exclude him or her from their group of friends or leave him or her out of things on purpose
> - hit, kick, push, shove around, or lock him or her inside a room
> - tell lies or spread false rumors about him or her or send mean notes and try to make other students dislike him or her
> - and do other hurtful things like that.
>
> When we talk about bullying, these things happen more than just once, and it is difficult for the student being bullied to defend himself or herself. We also call it bullying when a student is teased more than just once in a mean and hurtful way. But we do not call it bullying when the teasing is done in a friendly and playful way. Also, it is not bullying when two students of about equal strength or power argue or fight.
>
> (Olweus, 2007a, p. 2)

Students were asked how often they had been bullied at school in the past couple of months. There were five response options: "*I have not been bullied at school in the past couple of months*" (coded 1); "*It has only happened once or twice*" (coded 2), "*2 or 3 times a month*" (coded 3), "*About once a week*" (coded 4), or "*Several times a week*" (coded 5). Following this global question, students were asked about the frequency with which they had experienced nine specific forms of bullying, capturing direct verbal, direct physical, indirect/relational, and electronic forms of bullying. Each of the nine questions had the same response alternatives as the global question. Two bullying victimization scores were computed. A dichotomized Being Bullied Global score was calculated based on students responding that they had been bullied "2 or 3 times a month" or more often (see e.g., Olweus, 2013; Solberg & Olweus, 2003). In addition, a composite Being Bullied scale score was created for each participant by taking the average of all nine items ($\alpha = 0.85$ and $\alpha = 0.87$ for the 2-year and 3-year analyses, respectively).

2.2.2. Bullying others

Students were also asked a global question about the frequency with which they had bullied other students at school in the past couple of months (Bullying Others Global score), coupled with questions about nine different forms of bullying others. As a parallel to the Being Bullied scale, a composite Bullying Others scale was created ($\alpha = 0.88$ and 0.89 for the 2-year and 3-year analyses, respectively).

Since the psychometric characteristics of the Being Bullied and Bullying Others variables have typically been assessed in the same studies, we report information on their reliability and validity below. At least seven empirical studies from independent researchers have reported reliabilities for the bullying victimization and perpetration scales in the range of 0.80–0.90 (see Breivik & Olweus, 2015; Olweus, 2013). With respect to the global questions of being bullied and bullying others, Solberg and Olweus (2003) found very high correlations at the school level between the global scores and scale scores ($r = 0.79$ for being bullied and $r = 0.77$ for bullying others). There also is good evidence of construct and convergent validity of the global questions of being bullied and bullying others (Solberg & Olweus, 2003). Both the two global questions and the corresponding scales have shown substantial correlations (in the 0.35–0.40 range) with independent peer ratings of corresponding dimensions in a large-scale project with students in grades 3–10 ($n = 19{,}780$; see Olweus, 2013).

2.2.3. Reactions to bullying

Students were asked about their own (re)actions when they witness bullying. One question assessed the propensity of students to join in bullying of another student ("Do you think you could join in bullying a student whom you don't like?"), to which students had the following response options: "*Definitely no*" (coded 0), "*No*" (coded 1), "*No, I don't think so*" (coded 2), "*I do not know*" (coded 3), "*Yes, maybe*" (coded 4) or "*Yes*" (coded 5). Responses to this item are likely to capture not only students who actually bully other students but also those who may have a propensity to bully if they observe it around them. In our large Study 2 sample (measured at baseline, before intervention), the individual-level (Pearson), correlation between this item and the Bullying Others variable was clearly positive ($r = 0.41$), as expected, and the form of the association was monotone-increasing (basically linear).

Students were also asked, "When you see a student your age being bullied at school, what do you feel or think?" Response options included: "*That is probably what he or she deserves*" (coded 0), "*I do not feel much"* (coded 1), "*I feel a bit sorry for him or her*" (coded 2), "*I feel sorry for him or her and want to help*" (coded 3). This question is intended to measure the extent to which a student feels empathy for bullied youth. In our analyses based on the same large sample, the correlation of this item with Bullying Others was negative ($r = -0.25$), as expected (students with bullying tendencies typically score low on empathy variables; e.g., Olweus, 1993; Olweus & Endresen, 1998), and the association was monotone-decreasing (basically linear).

2.2.4. Perceptions of class teacher's actions

Students were further asked, "Overall, how much do you think your class (homeroom) teacher has done to counteract bullying in the past couple of months?" The five response options ranged from "*little or nothing*" (coded 0) to "*much*" (coded 4). This question is designed to measure students' perceptions of the degree to which educators (and in particular the educator with whom the students interact the most) actively address bullying, and although there is no published evidence of the validity of this item, it has substantial face validity.

2.2.5. Demographic questions

Finally, students were asked questions about their sex, grade in school, and their race or ethnicity. For this latter question, students were asked, "How do you describe yourself?" and were asked to indicate as many of the following categories that applied to them: American Indian, Black or African American, Arab or Arab American, Hispanic or Latino, Asian American, White, other, or I don't know.

2.3. Procedure

Training of school staff and implementation of the OBPP followed standard practices. The OBPP was implemented with the support of local certified OBPP trainer-consultants, who provided a 2-day training and monthly in-person or telephone consultation to members of each school's BPCC throughout implementation of the OBPP. BPCCs, with assistance from certified OBPP trainers-consultants, provided a full day of training for all staff prior to implementing the program. Schools received all necessary OBPP materials during the first year of the program. (For more details about program implementation, see Olweus & Limber, 2010b; Masiello & Schroeder, 2014.)

To evaluate the OBPP, classroom teachers distributed the Olweus Bullying Questionnaire (OBQ; Olweus, 2007a) in a pencil/paper scannable, anonymous format to the students in the relevant grades approximately 3–4 months before the official start of the program. Schools either began the program in the fall (shortly after start of the school year) or winter (shortly after winter holidays). Thus, the month in which the OBQ was administered varied among schools. However, dates of survey administration were carefully recorded so that new measurements with the same questionnaire were made at the same time of the year one, two and three years after the first administration. School personnel received a detailed school-level report of the findings from the questionnaire (Olweus, 2007b) to assist in their planning and internal evaluation of progress/lack of progress. When school personnel submit their scannable forms to be processed, they are asked to include an information form, upon which they may indicate whether or not they agree to share their data with Dan Olweus and his fellow researchers. In this study, all agreed to do so. This study was conducted in compliance with the human subjects review board of the first author's institution.

2.4. Study design

The design for this quasi-experimental study was an "extended age cohort design," as developed by Olweus (Olweus, 2005; Olweus & Limber, 2010b; Shadish, Cook, & Campbell, 2002). In an extended age cohort design, same-aged students from the same schools are compared across periods in time. For example, when the effect of a program is evaluated after a one-year period, students in grade 7 at Time 0 (T0, before intervention) will be compared with students in grade 7 from the same school at Time 1 (T1) one year later. These students were in grade 6 at T0, and at T1 they have been exposed to the program for approximately eight or nine months. In such a comparison, possible maturational or age-related differences between the comparison groups are controlled.

When the groups to be compared belong to the same schools, there are good grounds for assuming that a grade cohort differs in only minor ways from its contiguous cohort. Usually, the majority of the members in the various grade cohorts have been recruited from the same relatively stable populations and are also likely to have been students in the same schools for several years. The schools thus serve as their own controls and in this way, the problem with initial differences between the groups to be compared can be largely reduced or avoided. In other words, the "pretest" (T0) values for the individual schools can be considered good answers to the critical counterfactual question in all evaluation research: How do we obtain reasonable estimates of what the result would have been if the intervention subjects had not been exposed to the intervention (e.g., Cook, Shadish, & Wong, 2008)? As has been repeatedly documented, attempts to statistically correct for preexisting initial differences in common quasi-experimental designs (with nonequivalent control and intervention groups) are fraught with great difficulties (see e.g., Judd & Kenny, 1981; Shadish et al., 2002; Weisberg, 1979).

Another strength of the extended age cohort design (with more than two consecutive grade cohorts) is that one or more of the cohorts may serve as a baseline group in one set of comparisons and as an intervention group in another (see Olweus, 2005, and Olweus & Limber, 2010b, for details.) It is worth noting that the extended age cohort design has later been used by several other researchers to evaluate bullying prevention programs (Ertesvåg & Vaaland, 2007; Kärnä, Little, Voeten, Poskiparta, Kaljonen, & Salmivalli, 2011; Kärnä, Little, Voeten, Poskiparta, Alanen, & Salmivalli, 2011; Salmivalli, Kaukiainen, & Voeten, 2005).

2.4.1. History effects

A possible threat to the internal validity of conclusions about program effects in this design is what is usually called "history effects." Such effects may occur due to general time trends or some irrelevant (subject or environmental) factor that may affect or have affected the intervention group(s) and not the baseline/comparison group(s). Such threats may be difficult to completely rule out in a single study if little is known about results from studies with the same outcome variables without intervention. The current study, however, which used consecutive cohorts of roughly similar schools, allowed us to examine such effects by comparing initial assessments on key measures of interest for adjacent cohorts. Clear differences in these initial assessments might indicate some form of history effect that need to be considered in the interpretation of the results. The absence of such differences, on the other hand, would make an explanation or partial explanation of registered changes over time as "history effects" considerably less likely.

2.4.2. Length of program exposure

In the current project, > 95% of the schools took the baseline measurement, T0, in May–June, and had their program start in the beginning of the next school year. With evaluation of program effects one year after baseline measurement, students in all grade cohorts had been exposed to the program for approximately eight-nine months. This is a common situation which has been described in some detail elsewhere (Olweus, 2005).

When the evaluation of program effects is made two years after the baseline measurement (as is the case for the full sample in the current project), the exposure for the youngest grade cohort is somewhat different than for the older cohorts. In this case, it is natural to make a distinction between exposure for the relevant grade cohort and exposure for the school to which the grade cohort belongs. To illustrate this distinction, we can use the example of a typical middle school with grades 6 through 8, and with a feeder school or feeder schools with lower grades that have not used the OBPP in previous years. In such a situation, students in the youngest (grade 6) cohort at T0 (C6) are compared with students who were in grade 4 at T0 (C4 at T0) and who were enrolled in grade 6 in the middle school in August in 2009. At T2 in May–June 2010, these (intervention) students had thus been directly exposed to (been potentially influenced by) the program for about 9 months, from August 2009, to May–June 2010. However, the students in the next higher comparison/intervention cohort (C5 at T0) were enrolled in grade 6 in the middle school in August already in 2008. These (intervention) students had thus been directly exposed to the OBPP for approximately 18 months at T2, in May–June 2010, when they were compared with the students in grade 7 at T0 (C7), before intervention. Also the comparison/intervention cohort (C6 at T0) for the

Table 2
Changes in Being Bullied and in Bullying Others (dichotomized global questions) between baseline (T0) and time 2 (T2, two years later): analyses by individual grade.

Grade	*N* N0/N2	*B*	T0%	T2%	AC%
Being bullied global score					
3	16,653 8590/8063	−0.179***	26.19	22.88	3.31
4	16,856 8546/8310	−0.168**	24.90	21.89	3.01
5	17,630 9089/8541	−0.205***	25.29	21.62	3.67
6	22,302 11,270/11032	−0.207***	21.90	18.57	3.33
7	23,257 11,859/11389	−0.220***	19.51	16.28	3.23
8	22,801 11,772/11029	−0.058	16.93	16.14	0.79
9	6411 3372/3039	−0.198*	17.51	14.83	2.68
10	5481 2823/2658	−0.160	14.05	12.23	1.82
11	5819 3041/2778	−0.134	12.83	11.40	1.43
Bullying others global score					
3	16,386 8467/7919	−0.434***	5.37	3.55	1.82
4	16,639 8437/8202	−0.338**	5.71	4.14	1.57
5	17,497 9023/8474	−0.492***	8.39	5.30	3.09
6	22,191 11,222/10969	−0.487***	9.92	6.34	3.58
7	23,163 11,826/11337	−0.368***	10.97	7.86	3.11
8	22,758 11,762/10996	−0.349***	13.44	9.87	3.57
9	6375 3363/3012	−0.454***	14.54	9.75	4.79
10	5455 2811/2644	−0.438**	13.43	9.10	4.33
11	5799 3022/2777	−0.382**	12.85	9.15	3.70

B = unstandardized regression coefficient.
AC = absolute change (computed as T0-T2).
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

oldest cohort in the school, the grade 8 cohort (C8 at T0) had been directly exposed to the program for approximately 18 months at T2.

Although this illustration shows that the length of direct exposure to the program is less for the youngest comparison/intervention group of students than for older grade cohorts, it is important to realize that the length of program exposure to the school also had been about 18 months for this grade cohort. When the youngest comparison/intervention students (C4 at T0) are enrolled in the middle school in August 2009, some nine months before evaluation measurement at T2 (in May–June 2010), they enter a school where the OBPP was implemented nine months earlier. This is likely to be a considerable advantage for the newcomers since the school has already worked actively with the program for nine months. Accordingly, it is not unreasonable to expect that program effects for the youngest cohort will be approximately the same as for older cohorts with a longer direct exposure time. We do not expect that the difference in length of direct program exposure will have measurable consequences for the youngest grade cohorts in the participating schools. Although a detailed test of this assumption would be quite challenging (since there are numerous different grade combinations represented in the sample of schools), one can get a good impression of its credibility by examining the 6th grade cohort (the youngest grade cohort in a large proportion of the grade 6–8 schools). When the results of this cohort were compared with the results of adjacent cohorts, their similarity was striking (see percentage change values and regression coefficients in Tables 2 and 3 below).

Table 3
Changes in Being Bullied Scale and Bullying Others scale between baseline (T0) and time 2 (T2): Analyses by individual grade.

	Being bullied scale			Bullying others scale		
	Total *N*	*B*	Cohen's *d* (S)	Total *N*	*B*	Cohen's *d* (S)
Grade 3	16,589	−0.099**	0.81	16,501	−0.039	0.31
Grade 4	16,960	−0.098**	0.86	16,758	−0.047***	0.61
Grade 5	17,747	−0.107***	1.02	17,576	−0.063***	1.00
Grade 6	22,541	−0.099***	1.18	22,296	−0.079***	0.94
Grade 7	23,488	−0.094***	1.21	23,274	−0.060***	0.85
Grade 8	23,001	−0.040*	0.40	22,852	−0.059***	0.76
Grade 9	6468	−0.120***	1.90	6390	−0.056**	1.25
Grade 10	5511	−0.084*	0.94	5473	−0.103**	1.33
Grade 11	5858	−0.028	0.44	5816	−0.066*	1.20

B = unstandardized regression coefficients.
Cohen's d (S) = school-level effects.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

2.5. Approach to analysis

The Mplus 7.0 program (Muthén & Muthén, 2012) was used to analyze the multilevel (two-level) data consisting of individuals nested within schools and program effects based on school-aggregated outcome variables. For the key global and scaled outcome variables, being bullied and bullying others, the within-school intraclass correlations (the "unit" reliabilities) varied between 0.021 and 0.041. The aggregate reliabilities for the school-aggregated versions of these variables (based on an average "school" size of 339 students) were very high, in the 0.87–0.94 range (Kallestad, Olweus, & Alsaker, 1998, Table 5; Snijders & Bosker, 2012, p. 26). Accordingly, the schools could be differentiated on these variables with excellent reliability also when based on single items (e.g., the global item measuring percent bullied students in a school at a certain time point). The aggregate reliability estimates are of particular importance because all program effects in the study are based on school-aggregated variables. For the other outcome variables, the within-school intraclass correlations varied between 0.086 and 0.133; the highest value was obtained for the variable measuring how much the primary teacher had done to counteract bullying. For all of these variables, the aggregate reliabilities were in the 0.90's. As evident from the low (but highly significant) intraclass correlations (percent between school variances), most of the variance was linked to individual differences, ranging between 87 and 98%.

Our general model can be described as a Multi-site Block Design (Spybrook, Bloom, Congdon, Hill, Martinez, & Raudenbush (2011, Chapter 5), where schools represent the blocks or sites. As evident from the description of the study design noted above, same-aged students (in the same grade) from the same schools are compared across periods in time. This blocking is likely to considerably increase the power of the analyses.

The general (combined) model is:

$$Y_{ij} = y_{00} + y_{10}X_{ij} + u_{0j} + u_{1j} + e_{ij}$$

where Y_{ij} is the outcome variable for an individual student in school j, y_{00} is the average school (grand) mean, X_{ij} is the treatment indicator (coded with sets of dummy variables reflecting program year), y_{10} is the main effect of the treatment (the average difference between the treatment conditions), u_{0j} the random error associated with the level-2 means, u_{1j} the random error associated with the treatment effects, and e_{ij} the random error associated with the students at school j. Program year (T0 = baseline, T1 = one year after program start, T2 = two years after program start, etc.) is treated as a within-school (grade or grade grouping) treatment indicator (X_{ij}). In this model, we have not specified treatment as a fixed factor, but we will not explore further the variance of the treatment effects (across schools) in this article.

The outcome variables were treated as continuous, with the exception of the two global questions dichotomized into being bullied/bullying others "2-3 times a month or more" (coded 1) versus "less than 2–3 times a month" (coded 0). For the latter variables, results were analyzed with multilevel logistic regression using a logit link (Heck & Thomas, 2015). The MLR estimator was used in all multilevel analyses except for analyses of possible interaction effects between program and sex with categorical outcome variables. In these cases, the Bayes estimator was used because it is less computationally demanding than MLR (Muthén & Asparouhov, 2012). When a significant program x sex interaction effect was found, analyses were rerun on boys and girls separately using the MLR estimator.

In order to facilitate identification of possible main developmental trends in the longitudinal analyses (Study 2), we combined the students into sets of three grades levels, which roughly correspond to elementary, middle, and high school grades (grades 3–5, 6–8, and 9–11. Doing so provides more stable and replicable estimates than analysis of nine individual grades.

With continuous variables such as the Being Bullied/Bullying Others scales, use of a common individual-level effect size such as Cohen's *d* gives results that are misleadingly low, since a majority of the participants have scores of zero and cannot improve. Also, because program effects are based on school-aggregated variables (Spybrook et al., 2011, Chapter 9), for a number of analyses, we

have calculated school-level Cohen's *d*'s with the between-school standard deviation in the denominator (Hedges, 2007, 2011) rather the individual-level variant.

A key indicator of program effects for the dichotomized global questions is Absolute Change, AC% (% bullied students at T0 - % bullied students at T2). This measure has the advantage of being largely independent of the levels of baseline values of the groups compared and is easy to interpret (Wilkinson, 1999). In our study, it can be directly translated into an estimate of the number of students who have escaped being bullied (or the number of new victims) after implementation of the program. To facilitate comparison with some other studies, we have also reported Relative Change defined as [(% bullied students at T0 - % bullied students at T2)/% bullied students at T0].

3. Results

Results are presented in two sections. The first section reports analyses of possible program effects with respect to being bullied and bullying others for the full sample of students over two years (Study 1). Detailed analyses by grade and grade-groupings are provided, in addition to analyses by sex and race/ethnicity. The second section provides more detailed examinations of longer-term effects (over three years) for the sub-group of 95 schools for whom such data were available (Study 2). In all these analyses, schools belonging to different school cohorts (2008, 2009, and 2010 cohorts) were aligned. To achieve more stable estimates, most analyses were based on grade groupings (grades 3–5, 6–8, and 9–11, which corresponds to elementary, middle, and high school grades in most schools) rather than individual grades.

3.1. Study 1: analyses for the full sample over two years

3.1.1. Grade-level analyses for being bullied

Table 2 and Fig. 1 present relevant data for the dichotomized Being Bullied Global variable. All grades showed reductions, and the changes over time were significant for all but three grades (8th, 10th, and 11th grades). Average Absolute Change (unweighted) amounted to 2.58% and varied between 0.79% (8th grade) and 3.67% (5thgrade). Corresponding analyses for the scale scores of being bullied are presented in Table 3. Significant changes from T0 to T2 were observed for all but one grade (11th). With the exception of grades 8 and 11, all school level effect sizes [Cohen's d(S)] were large to very large. Average change scores (Table 2) and the regression coefficient for grade 11 (Table 3) indicated that the program effects for being bullied seemed overall to be somewhat weaker in the high school grades than in elementary and middle school grades.

3.1.2. Grade-level analyses for bullying others

Similar analyses were conducted for the dichotomized Bullying Others Global variable (Table 2) and for the related scale scores (Table 3). Significant reductions in Bullying Others Global were observed for all grades. Average Absolute Change (unweighted) amounted to 3.28% and varied between 1.57% (4th grade) and 4.79% (9th grade). With regard to students' scaled scores of bullying others, significant changes from T0 to T2 were observed for all grades except 3rd grade. Again, school level effect sizes were large to very large with a few exceptions (3rd and 4th grades). In terms of relative strength of program effects at different grade levels,

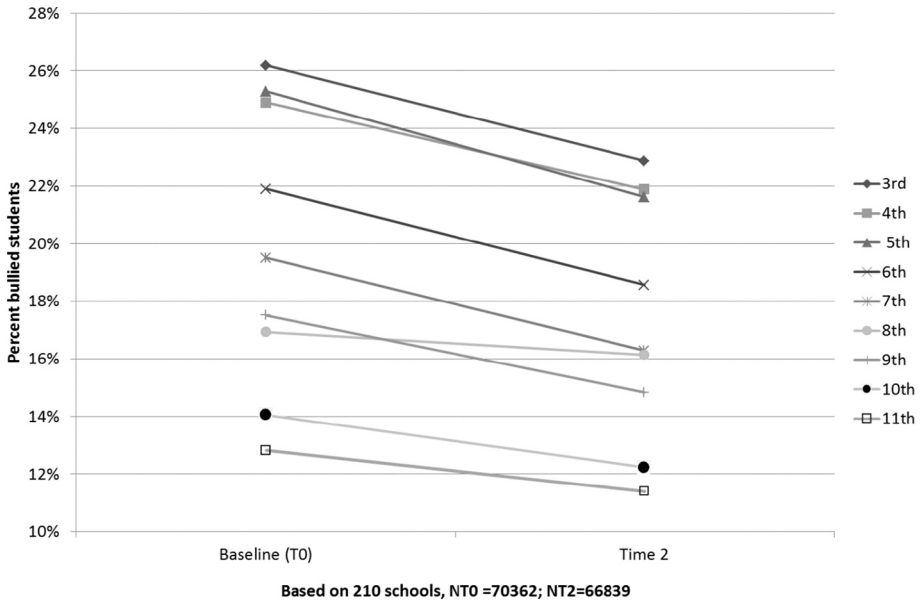

Fig. 1. Change in Being Bullied Global after implementation of the OBPP (2-year study, 210 schools). Analyses by grade.

Table 4
Changes in Being Bullied and Bullying Others (dichotomized global questions) between baseline (T0) and time 2 (T2): Analyses by sex and grade groupings.

Grades	Total *N*	*B*	T0%	T2%	AC (%)
Being bullied global score					
3–5	Girls: 25,188	Girls: −0.115**	Girls: 26.43	Girls: 24.26	Girls: 2.17
	Boys: 25,838	Boys: −0.274***	Boys: 24.75	Boys: 20.00	Boys: 4.75
6–8	68,360	−0.161***	19.93	17.48	2.45
9–11	17,711	−0.136*	14.91	13.27	1.64
Bullying others global score					
3–5	50,522	−0.449***	6.73	4.40	2.33
6–8	68,112	−0.389***	11.18	7.85	3.32
9–11	17,629	−0.414***	13.80	9.57	4.17

B = unstandardized regression coefficient.
AC = absolute change (computed as T0-T2).
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

average change scores were actually higher in high schools grades than other grades.

3.1.3. Examination of effects for sex and race

Analyses of the two dichotomized global variables are presented in Table 4. For five out of six analyses, the same regression coefficient could describe the significant positive program effects for boys and girls. Significant positive decreases over time were observed for both dichotomous variables for all three grade levels. In addition, a program by sex interaction was observed for students in grades 3–5 for the Being Bullied Global score (unstandardized beta = 0.099, $p < 0.05$, CI (90%) = 0.046–0.140), indicating somewhat greater reductions for boys than girls.

To examine the extent to which differential program effects were found for students of different races/ethnic groups, we conducted separate analyses on the two dichotomous global variables for students who identified themselves as Black, Hispanic, and White (Table 5). As mentioned, analyses were reported only if at least a total of 1000 students were represented within a group. As a

Table 5
Changes in Being Bullied and Bullying Others (dichotomized Global Questions) between baseline (T0) and time 2 (T2): Analyses by race/ethnicity and grade groupings.

Grades	*B* (program effects)	Baseline (T0)%	T2%	AC (%)	Total *N*
Black students: being bullied global score					
3–5	−0.165	23.90	21.03	2.87	2177
6–8	−0.041	14.70	14.19	0.51	3616
Hispanic students: being bullied global score					
3–5	−0.158	25.42	22.53	2.89	1462
6–8	−0.092	14.18	13.09	1.09	4595
White students: being bullied global score					
3–5	−0.191***	25.07	21.66	3.41	28,849
6–8	−0.190***	19.98	17.12	2.86	42,529
9–11	−0.204**	13.78	11.52	2.26	12,644
Black students: bullying others global score					
3–5	−0.129	12.94	11.55	1.39	2155
6–8	−0.347**	17.88	13.34	4.54	3579
Hispanic students: bullying others global score					
3–5	−0.356	8.88	6.39	2.49	1454
6–8	−0.217(*)	14.09	11.66	2.43	4592
White students: bullying others global score					
3–5	−0.444***	6.00	3.93	2.07	28,571
6–8	−0.482***	10.33	6.64	3.69	42,544
9–11	−0.507***	10.97	6.91	4.06	12,656

B = unstandardized regression coefficient.
AC = absolute change (computed as T0 − T2).
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.
(*) $p = 0.052$.

Table 6
Changes in Being Bullied Scale and Bullying Others scale across four time periods.

Grades	Total *N*	*B* T0 vs. T1	*B* T0 vs. T2	*B* T0 vs. T3	*d* (S) T0 vs. T3	Additional contrasts
Being bullied scale score						
3–5	51,669	−0.075***	−0.111***	−0.120***	*d* = 1.05	T1 vs. T2**
						T1 vs. T3**
6–8	49,964	−0.056***	−0.072***	−0.079***	*d* = 0.94	T1 vs. T3*
9–11	20,433	−0.032	−0.039	−0.076**	*d* = 1.20	T1 vs. T3*
						T2 vs. T3*
Bullying others scale score						
3–5	50,996	−0.030***	−0.048***	−0.058	*d* = 0.75	T1 vs. T2**
						T1 vs. T3***
						T2 vs. T3*
6–8	49,522	−0.037***	−0.067***	−0.092***	*d* = 1.30	T1 vs. T2***
						T1 vs. T3***
						T2 vs. T3***
9–11	20,249	−0.029	−0.052	−0.090**	*d* = 1.27	T1 vs. T3 **
						T2 vs. T3*

T0 = baseline, T1 = Time 1 (one year later), T2 = Time 2 (two years later), T3 = Time 3 (three years later); *B* = unstandardized regression coefficient; d (S) = school-level effects.

* $p < 0.05$.

** $p < 0.01$.

*** $p < 0.001$.

consequence, we did not analyze the responses from Hispanic and Black students in the 9–11 grade grouping. Program effects were typically somewhat larger for White students but also significant for Black middle school students (bullying others) and Hispanic middle school students (bullying others).

3.1.4. Analyses of possible history effects

To examine historical effects, we used multiple group analyses and compared initial assessments (at baseline) on the two key measures of interest for the three cohorts beginning with the program in 2008, 2009, and 2010, respectively. Satorra Bentler Chi-square difference test (*TRd*) analyses revealed no significant differences across the cohorts with regard to the percentage of students being bullied (global question; $TRd = 0.99$, $df = 2$, $p > 0.05$), or the percentage of students bullying others (global question; $TRd = 3.42$, $df = 2$, $p > 0.05$), suggesting that the changes over time were not due to historical effects.

3.2. Study 2: analyses for the longitudinal sample over three years

Detailed analyses were conducted to assess somewhat longer-term and year-by-year changes in students' responses on variables of interest. In these analyses, we used grade groupings (3–5, 6–8, 9–11).

3.2.1. Students' reports of being bullied and bullying others

Table 6 presents slope values (program effects) for students' reports of being bullied and bullying others, based on the scale scores (see also Fig. 2). Among students in grades 3–5, there were significant reductions in students' reports of both variables between baseline (T0) and T1, between baseline and T2, and between baseline and T3. Moreover, there were reductions between T1 and T2 and T1 and T3 (far right column). Similar, though somewhat less marked results were obtained for the 6–8 grouping. For students in the 9–11 grade grouping, significant change occurred only after three years. By and large, effects became gradually stronger over time and were greatest after three years.

3.2.2. Students' reactions to bullying

Changes over time were analyzed for students' reports of how they feel when they see a student their age being bullied (upper panel of Table 7). Generally, there were increases over time in expressions of empathy for a bullied peer in all grade groupings. For students in grades 9–12, however, clear and significant increases did not emerge until T3. School-level effect sizes varied between moderate and large. Table 7 also presents changes over time in students' willingness to join in bullying a peer whom they don't like. Steady decreases over time were observed in students' reports for all grade groupings. The changes from one assessment to the next were marked and the strongest effects were seen after three years (TO vs T3). The school-level effects sizes were large overall.

3.2.3. Perceptions of class teacher's actions to address bullying

Table 7 also presents changes over time in students' assessment of how much their class or homeroom teacher had done to counteract bullying. In all grade groupings, significant and marked increases were observed in students' perceptions that their

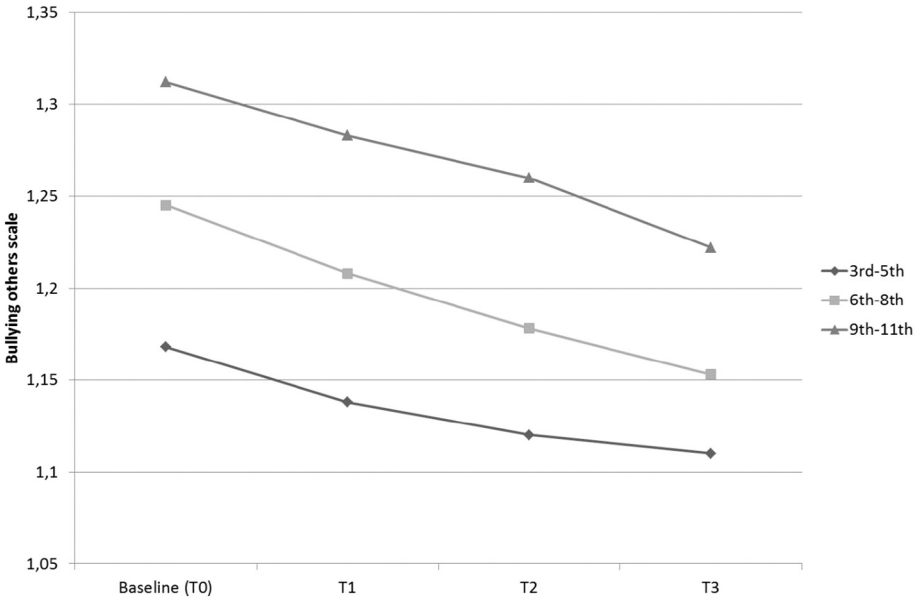

Fig. 2. Change in Bullying others scale scores after implementation of OBPP (3-year study, 95 schools). Analyses by grade groupings (grades 3–5, 6–8, and 9–11).

Table 7
Changes in students' perceptions and actions across four time periods.

Grades	Total *N*	*B* T0 vs. T1	*B* T0 vs. T2	*B* T0 vs. T3	*d* (S)	Additional contrasts
Students' perceptions of how they feel when a student is bullied						
3–5	51,056	0.007	0.030*	0.063***	*d* = 0.75	T1 vs. T3*** T2 vs. T3**
6–8	49,450	−0.054	0.030*	0.079***	*d* = 0.51	T1 vs. T2*** T1 vs. T3*** T2 vs. T3**
9–11	20,135	0.010	0.000	0.090*	*d* = 0.79	T1 vs. T3** T2 vs. T3***
Students' perceptions of whether they think they could join in bullying						
3–5	51,374	−0.104***	−0.118***	−0.234***	*d* = 1.02	T1 vs. T3*** T2 vs. T3***
6–8	49,575	−0.024	−0.197***	−0.300***	*d* = 0.99	T1 vs. T2*** T1 vs. T3*** T2 vs. T3***
9–11	20,211	−0.027	−0.151**	−0.265***	*d* = 1.25	T1 vs. T2* T1 vs. T3*** T2 vs. T3*
Students' perceptions that their teacher had addressed bullying						
3–5	51,046	0.265***	0.353***	0.323***	*d* = 1.38	T1 vs. T2** T1 vs. T3*
6–8	49,438	0.261***	0.302***	0.259***	*d* = 0.58	–
9–11	20,124	0.273***	0.314***	0.334***	*d* = 2.56	–

T0 = baseline, T1 = Time 1 (one year later), T2 = Time 2 (two years later), T3 = Time 3 (three years later); *B* = unstandardized regression coefficient; d (S) = school-level effects.

* $p < 0.05$.

** $p < 0.01$.

*** $p < 0.001$.

primary teacher had increased his or her efforts to address bullying. However, the general lack of significant changes between the various time periods after baseline (T1 vs. T2, T2 vs. T3), in particular for the two highest groupings, shows that there had occurred a change between baseline and T1 which was maintained thereafter but not increased further. School-level effects sizes were large, on average.

4. Discussion

The purpose of this study was to evaluate a large-scale implementation of the Olweus Bullying Prevention Program with children and youth in grades 3–11 in the United States. Given increasing concerns about the markedly negative short- and long-term effects of being bullied on individuals who are targeted (Olweus, 2013; Ttofi et al., 2011a) and evidence that regular engagement in bullying other students is related to later antisocial and criminal behavior (Olweus, 1993, 2011; Ttofi et al., 2012), the prevention of bullying is of utmost importance. In light of the proliferation of prevention and intervention programs and questions about the effectiveness of such efforts, there is a need for systematic evaluations of commonly-used prevention programs to reduce bullying.

The results of this large-scale study provide support for the effectiveness of the Olweus Bullying Prevention Program in the participating schools. Among students in all grades, there were clear reductions in the two key dimensions, being bullied and bullying other students, and results were generally consistent for both global questions and scaled scores. In most analyses, program effects were of similar magnitude for boys and girls but somewhat weaker and less comprehensive for students who identified themselves as Black or Hispanic, compared to the White majority students.

The average Absolute Change score over the two-year period amounted to somewhat < 3%. This change score can be translated into a rough estimate of nearly 2000 students who have escaped being bullied in the two-year period (and very likely, longer) and a similar number who have stopped bullying other students. Examination of year-to-year changes among the longitudinal sample of participants over three years confirmed and extended the reported findings. In particular, these analyses documented overall increases in students' expressions of empathy with bullied peers, marked decreases in their willingness to join in bullying, and clear perceptions that their primary teacher had increased his or her efforts to address bullying. The changes with regard to empathy and (un)willingness to join in bullying strongly suggest that the program had helped shift student attitudes to bullying and related behaviors to a more positive and inclusive school climate. In addition, the longitudinal analyses showed that program effects were generally larger, the longer the program had been in place.

Our findings are relevant for both educators and researchers. They provide strong empirical support for the position that a research based, whole-school bullying prevention program like the OBPP is likely to have systematic positive effects among U.S. students, in contrast to general concerns about the likely ineffectiveness of programs developed outside of the U.S (Cohen et al., 2015; Evans et al., 2014; Hong & Espelage, 2012). In addition, they provide educators, parents and policymakers guidance about effective strategies to reduce bullying in schools.

The positive program effects were generally consistent with our expectations and the findings of several evaluations of the OBPP in Norway (e.g., Olweus & Limber, 2010a, 2010b) but were more consistently positive than previous studies of the OBPP in the U.S. (Bauer et al., 2007; Limber et al., 2004). The reductions in bullying victimization and perpetration are also consistent with the general message from the meta-analysis of Ttofi and Farrington (2009, 2011), that research-based bullying prevention programs can be effective.

With regard to the issue of the relative strength of program effects at different age/grade levels, the effects on being bullied were weaker in higher grades. To illustrate, in Study 1, the (unweighted) average Absolute Change score for grades 3–7 was 3.31% and 1.68% for the higher grades, 8–11. We also found no significant program effects among 11th graders. Moreover, in the 3-year longitudinal analyses of Study 2 (Tables 6 and 7), effects were generally somewhat weaker and took longer time to achieve in the highest grade grouping comprising grades 9–11. This was in line with our predictions, and may be explained, in part, by contextual differences in the school environments of high schools vs. elementary or middle schools (e.g., size of the student body and staff, multiple teachers for each student and somewhat different definitions of the teacher's role, less flexible schedules) that may affect students' sense of community and ease of implementation of the program (Limber, 2010; Olweus & Kallestad, 2010). The fact that we observed less positive results with regard to being bullied for the 8th grade students was unexpected and warrants further investigation.

For bullying others, results were somewhat inconsistent. The average Absolute Change score for the lower grades (grades 3–7, Table 2) was 2.63% whereas the score for grades 8–11 was 4.10%, actually suggesting stronger program effects in higher grades. However, as mentioned in the section on being bullied, the 3-year longitudinal analyses (Tables 6 and 7) generally suggested somewhat weaker effects in the 9–11 grade grouping than in the lower grades at Time 3. These longitudinal results are rather consistent with results from Norwegian evaluations, where program effects at higher grade levels have generally been somewhat weaker and have taken longer to obtain (Olweus & Limber, 2010b). Since different indicators pointed in opposite directions, the empirical results do not permit a strong and clear conclusion on this issue for the bullying others variables in this sample. Future research is warranted to further clarify possible grade-level differences in program effects and explore possible rationales.

For the most part, our findings were consistent, whether based on dichotomous (global) variables or scale scores of being bullied or bullying others. Of the 36 analyses that examined changes in being bullied and bullying others across 9 grades, consistent findings using dichotomous and scale scores were observed in all but 5 cases. With respect to students' reports of being bullied, analyses using the dichotomous variable revealed no significant program effects among students in grades 8, 10, or 11, whereas analyses using the scale score revealed significant program effects for all grades, with the exception of 11th. However, an analysis of the effect sizes of the scale scores for the 8th and 10th grade students (see Table 3) revealed they were somewhat weaker than most other grades. There also was one discrepancy in findings between the dichotomous variable for bullying others and the corresponding scaled score. Although significant program effects were observed for all age groups using the dichotomous variable, no significant program effects were observed for 3rd graders when using the scale score.

Consistent with expectations (Olweus & Limber, 2010b) and evaluations of other prevention programs outside of the U.S. that address bullying (Kärnä, Little, Voeten, Poskiparta, Kaljonen, & Salmivalli, 2011; Kärnä, Little, Voeten, Poskiparta, Alanen, &

Salmivalli, 2011), we observed significant reductions in being bullied and bullying others for both boys and girls. Although the majority of analyses indicated similar program effects for boys and girls, there was also a significant program by sex interaction in the grade 3–5 grouping in favor of boys. Although these results might suggest that the program is somewhat more effective in elementary grades among boys than girls, the difference should be interpreted cautiously, as there were no other program by sex interactions. The issue of possible differential program effects for boys and girls will benefit from more detailed analyses of different forms of bullying (such as physical and relational bullying, for example), as research has documented gender differences in the frequency of some forms of bullying (Zhang et al., 2016)

We tentatively examined program effects for White, Black, and Hispanic students and observed significant reductions in involvement in bullying for 8 of 14 possible analyses (Table 5). Although some of the percentage reductions among Black and Hispanic students were of about the same size as the reductions among White students, results did not reach significance for the ethnic minorities due to considerably smaller numbers of responding students. Nonetheless, the overall weaker program effects for Black and Hispanic students deserves careful attention. Additional research is needed to understand how students of different races and ethnicities understand, experience, and engage in bullying (Wang, 2013) and the extent to which such students (and possibly adults within their schools) are receptive or non-receptive to particular prevention and intervention strategies.

Students of all ages were generally positive in their assessments of their own teachers' actions to address bullying. These positive changes appeared after one year for all grade levels and the higher levels of teacher engagement were maintained over time. The fact that the teachers (according to their students) maintained the increased levels of their efforts over three years must be viewed as a very good sign, and in general agreement with the OBPP goal of making key elements of the program a natural, permanent part of the schools' everyday life and culture. More generally, these and other reported results can also be seen as confirming evidence that the program had achieved positive changes with regard to bullying in these schools and that the changes had been marked enough for the students to perceive them.

4.1. Strengths and limitations

One clear strength of this study was the large sample size and the wide range of grade levels included. Although the numbers of schools involved in evaluations of bullying prevention programs seldom surpass 20 schools (Farrington & Ttofi, 2009), our research followed students from 210 schools over two years (Study 1) and students from the sub-sample of 95 schools over three years (Study 2). Moreover, whereas many evaluations have focused on more narrow age ranges (e.g., elementary grades [Waasdorp, Bradshaw, & Leaf, 2012] or middle school grades [Espelage, Low, Polanin, & Brown, 2013]), this study assessed students across nine grades. However, our findings can be primarily generalized to schools in Pennsylvania with a demographic predominance of White students from schools in primarily suburban and rural locales and, only by reasonable assumption, to similar schools in other states in the U.S.

Outcome variables were limited to student self-reports, assessing both bullying perpetration and bullying victimization through global questions and scaled scores. As mentioned, these questions and scales have been shown to have a number of good psychometric qualities. In addition, it is important to realize that all program effects in the current study were based on school-aggregated outcome variables which are much more reliable (typically, in the high 0.80's or 0.90's) than corresponding (single) items at the individual level. Moreover, we assessed changes in students' perceptions of their own attitudes and reactions as potential witnesses to bullying. The positive changes on these "supplementary" variables were as expected and lend additional support to the results on the key outcome variables, being bullied and bullying others.

Although the results of the current evaluations were generally consistent with the findings of a number of evaluations of the OBPP in Norway (Olweus, 2005; Olweus & Limber, 2010b), it should also be noted that program effects were somewhat weaker/took somewhat longer time to achieve in the U.S. Additional research is needed to examine what factors might explain such a difference if it is replicated. This study did not analyze fidelity of implementation of the OBPP. In future studies of this and other bullying prevention programs, it will be important to examine the extent to which program implementation is related to outcomes.

This study did not employ an experimental design using random assignment of schools to intervention or control conditions. Although randomized controlled trials (RCTs) are usually, and often somewhat uncritically, considered the "gold standard" in all evaluation research, there are often serious problems realizing such a design in practice, particularly with large, complex organizations such as schools (Olweus, 2005; Shadish et al., 2002; Weisberg, 1979). In many cases, a good quasi-experimental design, such as the extended age cohort design used in the current study, can be considered a strong alternative for the study of program effects. The extended age cohort design used in the present evaluation can actually be seen as an example of "intact group matching" which is one of the (few) conditions under which observational/quasi-experimental studies are likely to produce causal estimates that are comparable to those obtained with experimental (randomized controlled) studies (Cook et al., 2008).

Another challenge involves ruling out possible effects of other school-based programs that may have been implemented within intervention schools during the evaluation period. For several reasons, such effects are not likely to explain our results. First, as mentioned in the introduction, there are very few, if any, programs that have documented systematic bullying-reducing effects in the U.S. Even if there existed such a program, it is highly unlikely that it had used a temporal pattern of implementation that was largely aligned with our design. In addition, by virtue of participating in this study, school administrators made commitments not to implement other bullying or violence prevention programs during the intervention period.

In summary, this study provides strong support for the effectiveness of the OBPP among U.S. students in elementary, middle, and early high school grades. Although our results cannot be widely generalized beyond schools with similar demographics to those in the current study, the findings provide clear evidence that a program developed in Norway can (with some adaptations to cultural and educational conditions) also be effective in reducing bullying among students in the U.S. Although an absolute reduction by 2–4%

may superficially not look very impressive, it means in practice that a considerable number of children and youth in the project have escaped bullying, have experienced an improved school situation, and, in many cases, have been provided a platform for a positive turning point in their lives. Future research building on the experiences from the current large-scale project and using data on fidelity of program implementation will likely contribute to making the program even more effective in the United States.

Acknowledgements

This research was supported by generous funding from the Highmark Foundation. The authors wish to thank The Highmark Foundation, colleagues at the Center for Health Promotion and Disease Prevention, and colleagues at the Center for Schools and Communities for their remarkable support of the intervention program and evaluation.

References

Bauer, N. S., Lozano, P., & Rivara, F. P. (2007). The effectiveness of the Olweus Bullying Prevention Program in public middle schools: A controlled trial. *Journal of Adolescent Health, 40*, 266–274. http://dx.doi.org/10.1016/j.jadohealth.2006.10.005.

Berger, K. S. (2007). Update on bullying at school: Science forgotten? *Developmental Review, 27*, 90–126. http://dx.doi.org/10.1016/j.dr.2006.08.002.

Breivik, K., & Olweus, D. (2015). An item response theory analysis of the Olweus Bullying scale. *Aggressive Behavior, 41*, 1–13. http://dx.doi.org/10.1002/ab.21571.

Cohen, J., Espelage, D., Twemlow, S., Berkowitz, M. W., & Comer, J. P. (2015). Rethinking effective bully and violence prevention effects: Promoting healthy school climates, positive youth development, and preventing bully-victim-bystander behavior. *Review of Educational Research, 15*, 2–40.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management, 27*, 724–750. http://dx.doi.org/10.1002/pam.20375.

Cook, R. C., Williams, K. R., Guerra, N. G., Kim, T. E., & Sadek, S. (2010). Predictors of bullying and victimization in childhood and adolescence: A meta-analytic investigation. *School Psychology Quarterly, 25*, 65–83. http://dx.doi.org/10.1037/a0020149.

Ertesvåg, S. K., & Vaaland, G. S. (2007). Prevention and reduction of behavioural problems in school: An evaluation of the Respect program. *Educational Psychology, 27*, 713–736. http://dx.doi.org/10.1080/01443410701309258.

Espelage, D. L., Low, S., Polanin, J. R., & Brown, E. C. (2013). The impact of a middle school program to reduce aggression, victimization, and sexual violence. *Journal of Adolescent Health, 53*, 180–186. http://dx.doi.org/10.1016/j.jadohealth.2013.02.021.

Evans, C. B., Fraser, M. W., & Cotter, K. L. (2014). The effectiveness of school-based bullying prevention programs: A systematic review. *Aggression and Violent Behavior, 19*, 532–544. http://dx.doi.org/10.1016/j.avb.2014.07.004.

Farrington, D. P., & Ttofi, M. M. (2009). School-based programs to reduce bullying and victimization: A systematic review. *Campbell Systematic Reviews, 6*, 1–149.

Gladden, R. M., Vivolo-Kantor, A. M., Hamburger, M. E., & Lumpkin, C. D. (2014). Bullying surveillance among youths: Uniform definitions for public health and recommended data elements, version 1.0. Atlanta. *GA: National Center for Injury Prevention and Control*. Centers for Disease Control and Prevention and U.S. Department of Education.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modelling techniques: MLM and SEM approaches using Mplus.* New York: Routledge.

Hedges, L. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*, 341–370. http://dx.doi.org/10.3102/1076998606298.

Hedges, L. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics, 36*, 346–380. http://dx.doi.org/10.3102/10769986103766.

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*, 312–322. http://dx.doi.org/10.1016/j.avb.2012.03.003.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions.* New York: Cambridge University Press.

Kallestad, J. H., Olweus, D., & Alsaker, F. (1998). School climate reports from Norwegian teachers: A methodological and substantive study. *School Effectiveness and School Improvement, 9*, 70–94. http://dx.doi.org/10.1080/0924345980090104.

Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Alanen, E., & Salmivalli, C. (2011a). Going to scale: A nonrandomized nationwide trial of the KiVa antibullying program for grades 1–9. *Journal of Consulting and Clinical Psychology, 79*, 796–805. http://dx.doi.org/10.1037/a0025740.

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

Limber, S. P. (2010). Implementation of the Olweus Bullying Prevention Program: Lessons learned from the field. In D. Espelage, & S. Swearer (Eds.). *Bullying in North American schools: A social-ecological perspective on prevention and intervention* (pp. 291–306). (2nd ed.). New York: Routledge.

Limber, S. P., Nation, M., Tracy, A. J., Melton, G. B., & Flerx, V. (2004). Implementation of the Olweus Bullying Prevention programme in the southeastern United States. In P. K. Smith, D. Pepler, & K. Rigby (Eds.). *Bullying in schools: How successful can interventions be?* (pp. 55–79). Cambridge, UK: Cambridge Press.

Limber, S. P., & Olweus, D. (2017). Lessons learned from scaling-up the Olweus Bullying Prevention Program. In C. Bradshaw (Ed.). *Handbook on bullying prevention: A life course perspective* (pp. 189–199). Washington, DC: National Association of Social Workers Press.

Luxenberg, H., Limber, S. P., & Olweus, D. (2015). *Bullying in U.S. schools: 2014 status report.* Center City, MN: Hazelden Foundation.

Masiello, M. G., & Schroeder, D. (2014). *A public health approach to bullying prevention.* Washington, DC: American Public Health Association.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313–335. http://dx.doi.org/10.1037/a0026802.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus User's Guide* (7th Ed). Los Angeles, CA: Muthén and Muthén.

National Academies of Sciences, Engineering, and Medicine (2016). *Preventing bullying through science, policy, and practice.* Washington, DC: The National Academies Press.

National Center for Education Statistics (2014). What are locale codes? Retrieved from https://nces.ed.gov/ccd/rural_locales.asp.

Office of the Special Representative of the Secretary-General on Violence against Children (2016). *Ending the torment: Tackling bullying from the schoolyard to cyberspace.* New York: United Nations.

Olweus, D. (1973). *Hackkycklingar och översittare. Forskning om skolmobbing.* Stockholm: Almqvist & Wicksell.

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

Olweus, D. (1991). Bully/victim problems among schoolchildren: Basic facts and effects of a school based intervention program. In D. J. Pepler, & K. H. Rubin (Eds.). *The development and treatment of childhood aggression* (pp. 411–448). Hillsdale, NJ: Erlbaum.

Olweus, D. (1993). *Bullying at school: What we know and what we can do.* New York: Blackwell.

Olweus, D. (1997). Bully/victim problems in school: Facts and intervention. *European Journal of Psychology of Education, 12*, 495–510. http://dx.doi.org/10.1007/BF03172807.

Olweus, D. (2005). A useful evaluation design, and effects of the Olweus Bullying Prevention Program. *Psychology, Crime & Law, 11*, 389–402. http://dx.doi.org/10.1080/10683160500255471.

Olweus, D. (2007a). *Olweus Bullying Questionnaire.* Center City, MN: Hazelden.

Olweus, D. (2007b). *Olweus Bullying Questionnaire: Standard school report.* Center City, MN: Hazelden.

Olweus, D. (2011). Bullying at school and later criminality: Findings from three Swedish community samples of males. *Criminal Behaviour and Mental Health, 21*,

151–156.
Olweus, D. (2013). School bullying: Development and some important challenges. *Annual Review of Clinical Psychology, 9*, 751–780. http://dx.doi.org/10.1146/annurev-clinpsy-050212-185516.
Olweus, D., & Endresen, I. (1998). The importance of sex-of-stimulus object: Age trends and sex differences in empathic responsiveness. *Social Development, 7*, 370–388. http://dx.doi.org/10.1111/1467-9507.00073.
Olweus, D., & Kallestad, J. H. (2010). The Olweus Bullying Prevention Program: Effects of classroom components at different grade levels. In K. Osterman (Ed.). *Indirect and direct aggression* (pp. 113–131). New York: Peter Lang.
Olweus, D., & Limber, S. P. (2007). *Olweus Bullying Prevention Program: Teacher guide.* Center City, MN: Hazelden.
Olweus, D., & Limber, S. P. (2010a). Bullying in school: Evaluation and dissemination of the Olweus Bullying Prevention Program. *American Journal of Orthopsychiatry, 80*, 124–134. http://dx.doi.org/10.1111/j.1939-0025.2010.01015.x.
Olweus, D., & Limber, S. P. (2010b). The Olweus Bullying Prevention Program: Implementation and evaluation over two decades. In S. R. Jimerson, S. M. Swearer, & D. L. Espelage (Eds.). *Handbook of bullying in schools: An international perspective* (pp. 377–401). New York: Routledge.
Olweus, D., Limber, S. P., Flerx, V., Mullin, N., Riese, J., & Snyder, M. (2007). *Olweus Bullying Prevention Program: Schoolwide guide.* Center City, MN: Hazelden.
Salmivalli, C., Kaukiainen, A., & Voeten, M. (2005). Intervention: Implementation and outcome. *British Journal of Educational Psychology, 75*, 465–487. http://dx.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*, 1–15. http://dx.doi.org/10.1002/(SICI)1098-2337(1996)22:1<1::AID-AB1>3.0.CO;2-T.
Salmivalli, C., & Voeten, M. (2004). Connections between attitudes, group norms, and behaviour in bullying situations. *International Journal of Behavioral Development, 28*, 246–258. http://dx.doi.org/10.1080/01650250344000488.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference.* Boston: Houghton-Mifflin.
Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis* (2nd edition). London: Sage.
Solberg, M. E., & Olweus, D. (2003). Prevalence estimation of school bullying with the Olweus Bully/Victim Questionnaire. *Aggressive Behavior, 29*, 239–268. http://dx.doi.org/10.1002/ab.10047.
Spybrook, J., Bloom, H., Congdon, H., Hill, C., Martinez, A., & Raudenbush, S. (2011). Optimal design plus empirical evidence: Documentation for the "Optimal Design" software. Retrieved from: http://hlmsoft.net/od/od-manual-20111016-v300.pdfGoogle Scholar.
Stavrindes, P., Georgiou, S., & Theofanous, V. (2010). Bullying and empathy: A short-term longitudinal investigation. *Educational Psychology, 30*, 793–802. http://dx.doi.org/10.1080/01443410.2010.506004.
Ttofi, M. M., Eisner, M., & Bradshaw, C. P. (2014). Bullying prevention: Assessing existing meta-evaluations. *Encyclopedia of Criminology and Criminal Justice* (pp. 231–242). Springer New York.
Ttofi, M., & Farrington, D. (2009). What works in preventing bullying: Effective elements of programmes. *Journal of Aggression, Conflict and Peace Research, 1*, 13–24. http://dx.doi.org/10.1108/17596599200900003.
Ttofi, M. M., & Farrington, D. P. (2011). Effectiveness of school-based programs to reduce bullying: A systematic and meta-analytic review. *Journal of Experimental Criminology, 7*, 27–56. http://dx.doi.org/10.1007/s11292-010-9109-1.
Ttofi, M. M., Farrington, D. P., & Lösel, F. (2012). School bullying as a predictor of violence later in life: A systematic review and meta-analysis of prospective longitudinal studies. *Aggression and Violent Behavior, 17*, 405–418. http://dx.doi.org/10.1016/j.avb.2012.05.002.
Ttofi, M. M., Farrington, D. P., Lösel, F., & Loeber, R. (2011). Do the victims of school bullies tend to become depressed later in life? A systematic review and meta-analysis of longitudinal studies. *Journal of Aggression, Conflict and Peace Research, 3*, 63–73. http://dx.doi.org/10.1108/17596591111132873.
Van Goethem, A. A., Scholte, R. H., & Wiers, R. W. (2010). Explicit-and implicit bullying attitudes in relation to bullying behavior. *Journal of Abnormal Child Psychology, 38*(6), 829–842. http://dx.doi.org/10.1007/s10802-010-9405-2.
Waasdorp, T. E., Bradshaw, C. P., & Leaf, P. J. (2012). The impact of schoolwide positive behavioral interventions and supports on bullying and peer rejection: A randomized controlled effectiveness trial. *Archives of Pediatrics & Adolescent Medicine, 166*, 149–156. http://dx.doi.org/10.1001/archpediatrics.2011.755.
Wang, W. (2013). Bullying among U.S. school children: An examination of race/ethnicity and school-level variables on bullying. *All Dissertations* (Paper 1204).
Weisberg, H. I. (1979). Statistical adjustments and uncontrolled studies. *Psychological Bulletin, 86*, 1149–1164. http://dx.doi.org/10.1037/0033-2909.86.5.1149.
Wilkinson, L., & The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604. http://dx.doi.org/10.1037/0003-006X.54.8.594.
Zhang, A., Musu-Gillette, L., & Oudekerk, B. A. (2016). *Indicators of school crime and safety: 2015.* Washington, DC: National Center for Education Statistics, U.S. Department of Education, and Bureau of Justice Statistics, Office of Justice Programs, U.S. Department of Justice.
Zych, I., Ttofi, M. M., & Farrington, D. P. (2017). Empathy and callous–unemotional traits in different bullying roles: A systematic review and meta-analysis. *Trauma, Violence, & Abuse* (doi: 1524838016683456).