Journal of School Psychology 52 (2014) 479–493

Contents lists available at ScienceDirect

Journal of School Psychology

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

The KiVa antibullying curriculum and outcome: Does fidelity matter?☆

Anne Haataja [a,*], Marinus Voeten [b], Aaron J. Boulton [c], Annarilla Ahtola [a], Elisa Poskiparta [a], Christina Salmivalli [a]

[a] Department of Psychology, University of Turku, Finland
[b] Behavioral Science Institute, Radboud University Nijmegen, The Netherlands
[c] Department of Psychology, University of Kansas, USA

ARTICLE INFO

Article history:
Received 19 June 2013
Received in revised form 5 July 2014
Accepted 5 July 2014
Available online 22 August 2014

Keywords:
Antibullying program
Implementation
Fidelity
Victimization
Bullying
Evaluation

ABSTRACT

Research on school-based prevention suggests that the success of prevention programs depends on whether they are implemented as intended. In antibullying program evaluations, however, limited attention has been paid to implementation fidelity. The present study fills in this gap by examining the link between the implementation of the KiVa antibullying program and outcome. With a large sample of 7413 students (7–12 years) from 417 classrooms within 76 elementary schools, we tested whether the degree of implementation of the student lessons in the KiVa curriculum was related to the effectiveness of the program in reducing bullying problems in classrooms. Results of multilevel structural equation modeling revealed that after nine months of implementation, lesson adherence as well as lesson preparation time (but not duration of lessons) were associated with reductions in victimization at the classroom level. No statistically significant effects, however, were found for classroom-level bullying. The different outcomes for victimization and bullying as well as the importance of documenting program fidelity are discussed.

1. Introduction

In order to achieve desired aims, school-based prevention and intervention programs should be implemented as intended. In fact, a growing body of empirical evidence suggests that the closer the implementation of an intervention adheres to its original design, the more likely the desired outcomes take place (Durlak & DuPre, 2008; Dusenbury, Brannigan, Falco, & Hansen, 2003; Weare & Nind, 2011; Wilson & Lipsey, 2007). However, about half of prevention studies do not monitor implementation in any way (for a meta-analysis, see Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011). With respect to antibullying program evaluations in particular, it has been pointed out repeatedly that more attention needs to be paid to implementation fidelity (Ryan & Smith, 2009; Smith, Schneider, Smith, & Ananiadou, 2004; Ttofi & Farrington, 2010; Vreeman & Carroll, 2007).

Collecting implementation data is important for several reasons. The overall degree of program delivery informs program developers of whether the program is feasible and thus likely to be implemented with fidelity in the future. Second, when implementation is monitored, support can be adjusted for schools that are likely to fall short of implementation. Finally, a significant association

☆ The development of the KiVa program was financed by the Finnish Ministry of Education. The current study is supported by Academy of Finland grant 134843 to the last author. The authors thank the whole KiVa research team for support. Portions of this study were presented at the biennial Meeting of the Society for Research on Adolescence, Vancouver, March, 2012.
* Corresponding author at: Department of Psychology, University of Turku, FI-20014, Finland.
E-mail address: anmahaa@utu.fi (A. Haataja).
ACTION EDITOR: John Carlson.

http://dx.doi.org/10.1016/j.jsp.2014.07.001
0022-4405/

between the level of implementation and outcome (e.g., reduction of victimization) provides further support for the effects obtained being actually caused by the program rather than by other factors such as developmental changes or selection effects (i.e., initial differences between intervention and control schools).

The KiVa antibullying program (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011) has been identified as one of the most promising whole-school programs to prevent bullying (Ttofi & Farrington, 2010). Evidence of the effectiveness of KiVa has been acquired both in a randomized controlled trial (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; 2013) and during broad roll-out of the program in Finnish schools (Kärnä, Voeten, Little, Poskiparta, Alanen, et al., 2011). In both studies, KiVa was found to reduce bullying and victimization significantly already after 9 months (one school year) of implementation, especially in elementary school level (Grades 1–6). However, we know relatively little about teachers' fidelity to the program across classrooms. Therefore, the purpose of the present study is (a) to test whether stronger effects of the KiVa program can be obtained with a higher degree of implementation of the curriculum and (b) to provide insight into the preconditions of success of evidence-based antibullying programs more generally—for instance, which aspects of implementation are most strongly associated with positive outcomes.

1.1. The concept of implementation fidelity

Implementation fidelity refers to the degree to which individuals delivering a prevention program implement it as intended by program developers (Dane & Schneider, 1998; Durlak & DuPre, 2008). It can be conceptualized and operationalized in terms of quantity (how much was done) or quality (how well it was done) of program implementation.

In many studies, the analysis of implementation data focused on the *quantity of the program content delivered* to targeted individuals (e.g., children) or groups (e.g., classrooms). This aspect of implementation has been interchangeably referred to as *lesson adherence*, or *dosage*, and it has been operationalized, for instance, as the number of sessions completed, percentage of tasks covered, or time spent on program delivery (Cross, Hall, Hamilton, Pintabona, & Erceg, 2004; Dane & Schneider, 1998; Durlak & DuPre, 2008; Dusenbury et al., 2003; Elliott & Mihalic, 2004; Ennett et al., 2011; Jones, Brown, & Lawrence Aber, 2011).

Evidently, there is natural variation in the amount of program content covered in school-based prevention programs (Dusenbury, Brannigan, Hansen, Walsh, & Falco, 2005; Lillehoj, Griffin, & Spoth, 2004), and a level of 100% adherence is rarely reached (Dane & Schneider, 1998; Durlak & DuPre, 2008; Dusenbury et al., 2003) regardless of assessment methods, targeted students, or types of programs. This phenomenon has been referred to as *adaptation* (Durlak & DuPre, 2008). For instance, teachers may intentionally eliminate some program aspects if they find them too challenging (Fagan & Mihalic, 2003; Hahn, Noland, Rayens, & Christie, 2002) or if students are not interested or responsive (Martens, van Assema, Paulussen, Schaalma, & Brug, 2006). Some degree of program adaptation (e.g., choosing between program elements) is usually taken for granted, and successful outcomes have been reported at levels such as the delivery of 60% of program content (Fagan & Mihalic, 2003; Ferrer-Wreder et al., 2010; Hahn et al., 2002).

Findings from recent reviews have demonstrated that a higher degree of program elements delivered (Durlak & DuPre, 2008; Wilson & Lipsey, 2007) has a positive association with the outcomes obtained (Durlak et al., 2011). However, non-significant dose-outcome associations have also been reported: the explanations provided for such findings include minimal variability in the degree of implementation (Domitrovich, Gest, Jones, Gill, & DeRousie, 2010; Lillehoj et al., 2004; Low, Ryzin, Brown, Smith, & Haggerty, 2014; Resnicow et al., 1998), teachers' lack of prior experience with prevention programs (Lillehoj et al., 2004), or the use of implementation ratings provided by teachers rather than trained observers (Dane & Schneider, 1998; Lillehoj et al., 2004).

Besides adherence, or the dosage of program content covered, another aspect of fidelity is the *quality of implementation*. Observational measures are often used to assess implementation quality, such as the interaction between students and the teacher delivering the program (Hahn et al., 2002; Hirschstein, Edstrom, Frey, Snell, & MacKenzie, 2007; Melde, Esbensen, & Tusinski, 2006; Mihalic, Fagan, & Argamaso, 2008; Resnicow et al., 1998), the degree of program content taught clearly and correctly (Fagan & Mihalic, 2003; Kam, Greenberg, & Walls, 2003; Lillehoj et al., 2004; Melde et al., 2006; Spoth, Guyll, Trudeau, & Goldberg-Lillehoj, 2002), or the extent to which teachers encouraged and coached students to apply intervention concepts beyond formal lessons (Domitrovich et al., 2010; Tortu & Botvin, 1989).

The quality of implementation may be equally or even more important than the quantity. However, the use of quality as a measure of implementation is rare. In the review by Durlak and DuPre (2008), only 6 out of 59 studies paid attention to the quality of delivery whereas quantity (dosage) was assessed in 29 studies. This lack of attention to quality might be due to the fact that collecting observational data is time-consuming and expensive, and yet the studies utilizing observations have produced mixed findings—with some reporting a positive association between the quality of delivery and student outcomes (Domitrovich et al., 2010; Kam et al., 2003; Resnicow et al., 1998) and some finding the opposite (Hirschstein et al., 2007).

Besides observational data, teachers' self-reports of their preparedness and knowledge of intervention model can be used as an indicator of the quality in which a program was delivered (Dane & Schneider, 1998). Better prepared teachers, who are more comfortable with a program's methods and who more strongly support its purpose, are more likely to implement the program in a competent manner (Rohrbach, Graham, & Hansen, 1993). For instance, Kallestad and Olweus (2003) found that reading the program manual was one of the strongest predictors of antibullying lesson implementation.

1.2. Implementation and outcomes in antibullying program evaluations

School bullying has been defined as repeated aggression towards a relatively powerless peer (Olweus, 1993; Smith & Sharp, 1994). A substantial number of antibullying programs has been evaluated (Merrell, Gueldner, Ross, & Isava, 2008; Ryan & Smith, 2009; Smith et al., 2004; Ttofi & Farrington, 2010; Vreeman & Carroll, 2007). The findings have been somewhat inconsistent, some showing small

to moderate positive effects of the programs, and some unexpectedly negative effects indicating increases rather than decreases in bullying problems. In a review of Vreeman and Carroll (2007), only 4 out of 10 curriculum studies showed decreases in bullying problems. Although a recent meta-analysis based on 44 studies (Ttofi & Farrington, 2010) provided more encouraging findings, with average reductions of 17–23% for bullying others and 17–20% for being bullied, even these effects are relatively modest in size. One reason might be less-than-optimal implementation of the programs. It can be assumed that a higher degree of program delivery generates larger effects in antibullying trials, as evaluations of other types of school-based prevention programs have indicated (Dane & Schneider, 1998; Durlak & DuPre, 2008; Kam et al., 2003; Lillehoj et al., 2004; Rohrbach et al., 1993).

Implementation fidelity of antibullying programs has rarely been documented and hardly ever related to program outcomes. In general, the content of such programs has varied from standardized curricula to rather flexible guidelines. Concrete manuals describing program content and learning objectives promote systematic implementation (Dane & Schneider, 1998; Ryan & Smith, 2009), but they also provide a good starting point for the measurement of the implementation fidelity (Brown, Low, Smith, & Haggerty, 2011; Cross et al., 2004; Hirschstein et al., 2007; Kallestad & Olweus, 2003). Cross et al. (2004), for instance, found that during the Friendly Schools intervention, the average of classroom activities implemented was 67%. Brown et al. (2011) examined the implementation of the Steps to Respect program, reporting an exceptionally high level of classroom implementation; on average, more than 90% of activities were covered.

When the intervention content is loosely defined, the evaluation of the degree of implementation (at least in quantitative terms) is difficult. In some trials, schools have been encouraged to develop their own written antibullying policies, choose from among alternative intervention activities the ones they wish to implement, or engage in both strategies (Eslea & Smith, 1998; Fekkes, Pijpers, & Verloove-Vanhorick, 2006). In such cases, the average degree of implementation delivery across schools (or classrooms) cannot be defined, and more importantly, between-school or between-classroom differences in implementation can hardly be related to the outcomes obtained. The identification of more vs. less effective activities, on the other hand, would demand a large sample size and preferably random assignment of schools to different conditions implementing different activities.

In some studies, the documentation of implementation has been based on interviews with a single respondent per school (Eslea & Smith, 1998; Stevens, Van Oost, & De Bourdeaudhuij, 2001), or implementation data provided by several teachers has been aggregated at the school level (Salmivalli, Kaukiainen, & Voeten, 2005). It is thus assumed that each classroom in a school was targeted with the same degree of adherence to intervention content. However, the same content coverage is not likely, especially if the intervention includes activities delivered by teachers in classrooms.

Another threat to the reliability of implementation data is the assessment of implementation by teacher reports collected only after the intervention has ended (Fekkes et al., 2006; Kallestad & Olweus, 2003). Teachers' recall of their implementation actions during the intervention phase lasting for several months may not be accurate. In order to get comprehensive information regarding implementation and to increase the validity of self-reports, ratings should be collected at several time points during the delivery of intervention.

Keeping in mind the above limitations in measuring implementation it may not be surprising that only few studies reported findings regarding the association between the implementation of antibullying programs and the outcomes obtained. A study by Salmivalli et al. (2005) found a positive association between the overall degree of implementation and reductions in bullying and victimization at the school level. Kallestad and Olweus (2003) reported positive effects of some program aspects but not others. Still other studies have revealed very complicated patterns of findings leaving much room for interpretation. For instance, Hirschstein et al. (2007) assessed two aspects of classroom implementation of Steps to Respect bullying prevention program: lesson adherence and the quality of instruction. Although greater lesson adherence was related to better teacher-reported interpersonal skills among students, it was unrelated to observed or student-reported outcomes, including bullying and victimization. Unexpectedly, the quality of instruction (e.g., teachers' emotional tone and classroom management as rated by observers) was associated with increasing levels of student-reported victimization and difficulties in responding assertively to bullying. It should be noted that two implementation measures that were not directly related to the implementation of classroom curriculum (supporting skill generalization and coaching of students involved in bullying) were related to positive observed outcomes—but only for older students in the sample. Another study (Low et al., 2014) examining the implementation of the same program found no effects of lesson adherence on student outcomes either. However, students' engagement (as perceived by teachers) averaged across several lessons predicted classroom-level reductions in student-reported victimization but not in bullying perpetration.

In summary, studies examining the association between the implementation of antibullying programs and outcomes obtained are few, and they have produced mixed findings. There is clearly a need for more studies assessing several aspects of implementation fidelity at multiple time points in large enough samples. Also, modeling the effects of implementation at the level in which the intervention takes place (e.g., classroom) has not been adequately employed.

1.3. KiVa antibullying program

KiVa is a research-based antibullying program that has been developed in the University of Turku, Finland, with funding from the Ministry of Education and Culture. The KiVa program involves both universal actions targeted at all students, and indicated actions targeted at students who have been identified as targets or perpetrators of bullying (Salmivalli, Poskiparta, Ahtola, & Haataja, 2013). The name of the program, “KiVa” is a Finnish word meaning kind or nice, but it is also an acronym for “Kiusaamista Vastaan” (against bullying).

Although KiVa shares several features of other existing antibullying programs, such as student lessons delivered in classrooms, it has some unique aspects differentiating it from other programs. Most importantly, KiVa is based on the *participant role approach* to

bullying (Salmivalli, Lagerspetz, Björkqvist, Österman, & Kaukiainen, 1996), and thus it focuses on influencing the responses of peer bystanders witnessing bullying. It is assumed that once bystanders do not provide social rewards (such as laughing or cheering) to perpetrators, but instead support and defend victimized peers and express disapproval of bullying, an important motivation to bully others is lost. It should be noted that once the victims perceive the social environment as more supportive, their experience of the situation might be altered even if the bullies do not immediately stop their mean behaviors (Sainio, Veenstra, Huitsing, & Salmivalli, 2011; Salmivalli, Poskiparta, Ahtola, & Haataja, 2013). Thus, an important aim of the KiVa curriculum delivered in classrooms is to change the bullying-related norms and consequently reduce both bullying perpetration and experienced victimization.

As for the classroom curricula, at the elementary school level there are two different developmentally appropriate sets of student lessons (10 student lessons × 90 min each)—one for Grades 1–3 and the other for Grades 4–6. The main aims of the student lessons are to raise awareness of the role bystanders play in the bullying process, to increase empathic understanding of the victim's plight, and to provide students with safe strategies to support and defend their victimized peers. The topics of the lessons proceed from more general ones, such as the importance of respect in relationships, group communication, and group pressure, to bullying and its mechanisms and consequences. Each lesson includes various types of activities, such as teacher-led discussion, small group discussion, learning-by-doing exercises, adopting class rules, as well as individual tasks such as playing an antibullying computer game. Classroom teachers deliver the lessons during the school year (in Finland, from August to May, i.e., one lesson per month) according to the guidelines provided in the teacher's manuals.

The effects of KiVa antibullying program have been evaluated in several studies based on a randomized controlled trial that took place from 2007 to 2009 (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; Kärnä et al., 2013) and nationwide roll-out of the program (Kärnä, Voeten, Little, Poskiparta, Alanen, et al., 2011). There is some initial evidence of the degree of implementation of KiVa (at the school level) being associated with program outcomes as reported by Kärnä, Voeten, Little, Poskiparta, Alanen, et al. (2011). Despite the promising findings, there is so far a lack of evidence about the effects obtained at varying levels of implementation and the relative importance of different implementation aspects for outcomes (i.e., classroom level reductions in bullying and victimization).

1.4. The present study

Although authors of prevention studies have argued that the assessment of fidelity is an essential feature of program evaluations (Durlak & DuPre, 2008; Durlak et al., 2011; Dusenbury et al., 2003; Ennett et al., 2011; Gingiss, Roberts-Gray, & Boerm, 2006), it has been uncommon in antibullying program trials (for meta-analyses, see Ryan & Smith, 2009; Ttofi & Farrington, 2010). When data on implementation has been gathered, it has been assessed at a very general level (e.g., Eslea & Smith, 1998; Fekkes et al., 2006; Kallestad & Olweus, 2003; Salmivalli et al., 2005), in small samples (Eslea & Smith, 1998; Stevens et al., 2001), and typically reported as descriptive information rather than associated with the outcomes obtained.

Although the KiVa program has been found to reduce bullying and victimization (Kärnä, Voeten, Little, Poskiparta, Alanen, et al., 2011; Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; 2013), much remains to be known about program implementation and whether it explains variation in program outcomes. Therefore, we focused on the extent to which teachers delivered the classroom curriculum (quantity of the efforts) and how well they did it (quality of the efforts) throughout a school year. Thus, the current study extends the effectiveness studies of the KiVa program by examining the implementation fidelity of the program curriculum and relating it to reductions in bullying perpetration as well as victimization. From a broader perspective, it is one of the first studies in the field (a) assessing several aspects of lesson implementation, (b) utilizing longitudinal multilevel modeling of hierarchical data, and (c) linking implementation aspects with outcomes at the classroom level.

Implementation fidelity was systemically measured by collecting monthly teacher reports on the quantity as well as the quality of implementation of the KiVa program during the nine months of the trial. The measures of lesson adherence and duration of lessons were chosen to reflect the quantity of implementation; they represent the core features of the lesson plans provided in the structured teacher manuals of KiVa. Both measures have also been used in the previous research in the field (Dane & Schneider, 1998; Durlak & DuPre, 2008). The third measure, lesson preparation, was used as an indicator of the quality of implementation. All fidelity aspects were measured on continuous scales, providing more statistical power to detect effects on the outcome variables.

Besides exploring the overall degree of implementation of the KiVa curriculum, we tested the hypotheses that classroom-level reductions of victimization would be larger in classrooms where teachers reported more lesson adherence (Hypothesis 1a), longer duration of lessons (Hypothesis 1b), and more lesson preparation (Hypothesis 1c). Similarly, we hypothesized larger classroom-level reductions of bullying in classrooms with higher lesson adherence (Hypothesis 2a), longer lesson duration (Hypothesis 2b), and more lesson preparation (Hypothesis 2c). We had no hypotheses regarding more or less effective aspects of implementation. However, based on previous findings (Merrell et al., 2008; Vreeman & Carroll, 2007) we expected that effects on victimization might be larger than effects on bullying perpetration.

2. Method

2.1. Participants

The sample comprises the intervention schools that took part in the first evaluation studies of the KiVa antibullying program (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; 2013). In the present study, we included only the elementary schools (Grades 1 thru 6) that were in the intervention condition. In total 77 elementary schools took part, 39 intervention schools from the first phase of

evaluation in 2007–2008, and 38 intervention schools from the second phase in 2008–2009. Some intervention schools in the second phase had been involved as control schools in the first phase (Kärnä et al., 2013).

There were in total 439 teachers in the 77 participating schools, but the analyses in the present study were based on the data from 417 teachers from 76 schools, as a result of deleting students with missing values on self-reported victimization and bullying (discussed in the Appendix A). The majority of the teachers (85%) worked in general education, and the remaining 15% as special education teachers. The total number of students in the 77 intervention schools was 8452. There were missing value patterns in the outcome measures at the student level, which have already been discussed in the previous evaluation studies of the KiVa program (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; 2013). In the present analyses, we used data from 7413 students (49% girls and 51% boys). The deleted 1039 students (450 from Grades 4–6 and 589 from Grades 1–3) did not participate either in the pretest or in the posttest or participated only in the pretest but were not in the schools during the intervention. Another reason for not participating was the lack of parental consent, which was given for 94% of the students.

In addition to the four sources of missing values mentioned, there was a fifth source—namely, nonresponse by teachers in the intervention schools who were asked to provide information about implementation of the KiVa program. Altogether, 332 out of 417 teachers (80%) returned the lesson booklets. This produced missing values in the three implementation variables derived from the booklets for 85 teachers, associated with 20% of the students. On average, 80% ($SD = 24$) of the booklets for a school were returned, ranging from 0% to 100%. To deal with missing values, we estimated the models using full information maximum likelihood (FIML) estimation (e.g., Enders, 2010). Details regarding the five missing value patterns are further discussed in the Appendix A.

2.2. Measures of dependent variables

2.2.1. Victimization

At pretest and posttest assessments (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; 2013), the term bullying was defined for students by classroom teachers in the way it is formulated in the Olweus' Bully/Victim Questionnaire (OBVQ; Olweus, 1986); this definition emphasizes the repetitive nature of bullying and the power imbalance between the victim and the bully. Four questions assessing direct and indirect forms of victimization (verbal, exclusion, physical, and manipulative) were used as indicators of latent variables for victimization. Specific questions of experienced victimization were presented on separate pages and were seen one by one, prompted by the following: "How often have you been bullied at school in the last two months in this way?" The two direct forms were verbal and physical: "I was called mean names, was made fun of or teased in a hurtful way" and "I was hit, kicked or shoved by someone" respectively. The two indirect forms were exclusion and manipulative behavior: "Other students ignored me completely or excluded me from things or from their group of friends" and "Other students tried to make others dislike me by spreading lies about me." All four items were responded to on a 5-point scale (0 = *not at all*, 1 = *once or twice*, 2 = *two or three times a month*, 3 = *every week once or twice*, and 4 = *several times a week*). Psychometric studies of the OBVQ have demonstrated evidence that both the victimization and bullying scales are internally consistent (e.g., greater than .80 reliability across several grades; Olweus, 2007) In addition, studies have demonstrated evidence of construct validity and concurrent validity, such as acceptable item fit statistics and item characteristics (Kyriakides, Kaloyirou, & Lindsay, 2006) as well as significant correlations (r's = .40–.60) with related constructs (Olweus, 2007). In the present study, the ordinal coefficient alpha (Gadermann, Guhn, & Zumbo, 2012) for the four victimization items was .87 at pretest and .88 at posttest.

2.2.2. Bullying

The term bullying was defined to the children in the way it is presented in the OBVQ (Olweus, 1986), which emphasizes the repeated nature of bullying and the power imbalance between the bully and the victim. At pretest and posttest assessments, four items representing typical forms of bullying others derived from the OBVQ were used as indicators of bullying behavior. Each item was prompted with the following: "Have you been bullying others in this way in the last two months?" The specific questions of bullying were presented on separate pages so that four forms (verbal, exclusion, physical and manipulative) were seen one by one. The direct forms were verbal and physical: "I called someone by mean names, made fun of, laughed in someone's face or teased in a hurtful way" and "I hit, kicked or shoved someone," respectively. The two indirect forms of bullying (exclusion and manipulative behavior) were assessed by asking: "I ignored someone completely or excluded him/her from things or the group of my friends" and "I spread lies about someone and tried to get others to dislike him/her." All four items were responded to on a 5-point scale (0 = *not at all*, 1 = *once or twice*, 2 = *two or three times a month*, 3 = *every week once or twice*, and 4 = *several times a week*). At pretest, the ordinal alpha coefficient was .86; at posttest, the coefficient was .88.

2.3. Implementation measures

Prior to program implementation (in August), teachers were given lessons booklets in which all lesson-specific activities for each of the 10 lessons were listed. After a delivered lesson, they were instructed to mark whether they used (or did not use) each of the activities included and to continue doing so throughout the school year. In addition, the booklets contained questions about the time teachers had spent (in minutes) for preparing and delivering each lesson. If there were missing data on all three measures asked in the booklets, the lesson was considered as not delivered.

2.3.1. Lesson adherence

The first measure, designed to assess the total degree of lesson adherence for the KiVa curriculum, was calculated as the proportion of tasks delivered for each lesson. These proportions were averaged over the 10 lessons. The average proportion of curriculum tasks completed ranged from 3% to 100% with a mean of 68% ($SD = 20\%$).

2.3.2. Duration of lessons

The number of minutes used for teaching the lesson content was averaged across the lessons a teacher reported to have delivered. The duration of lessons ranged from 25 to 180 min, with a mean of 79 min ($SD = 19$ min).

2.3.3. Lesson preparation

Time spent in preparing the lessons was calculated by averaging the reported numbers of minutes across the lessons delivered by a teacher. The time devoted to preparing a lesson ranged from 8 to 98 min, with a mean of 29 min ($SD = 16$ min).

2.4. Design and procedures

Participating schools used the KiVa antibullying program for the first time during a period of one school year as they participated in a randomized controlled trial testing the effectiveness of the KiVa program (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; 2013). Schools had been randomly assigned to intervention and control conditions, but the 79 control schools were not part of the present sample.

Data from students were obtained by internet-based questionnaires. There were three measurement occasions: in May (T1), in December–January (T2), and in May (T3; one year after the pretest). In the present study, we used only the pretest measurement (T1; at the end of the school year before intervention) and the final posttest (T3; at the end of the school year in which the trial had taken place). All students with parental consent to participate received personal passwords to log in to the questionnaire and were assured strict confidentiality (see Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; 2013).

Implementation data were collected from teachers delivering the classroom curriculum in the intervention condition by means of lesson booklets. The teachers were asked to fill in the lesson booklet immediately after each lesson (i.e., every month) over the course of the intervention year. Schools were reminded at the end of the school year to mail the booklets to the University of Turku by using pre-addressed and pre-paid envelopes.

2.5. Analytic method

The present study explored relationships between the implementation of the KiVa curriculum—adherence, duration, and preparation—and reductions in victimization and bullying. Implementation of the curriculum occurred at the level of the classroom and therefore was the primary focus of analysis. Several models were estimated within the multilevel structural equation modeling (MSEM) framework. MSEM is an extension of structural equation modeling (SEM) that allows for latent variable model testing at multiple levels of data (Muthén & Asparouhov, 2011). This approach accounts both for measurement error—via latent variables—and for nested data (students within classrooms). One feature of MSEM that is important to this study is the decomposition of variation into components specific to classrooms and students.

Latent difference score (LDS) models were used to capture changes in bullying and victimization from pretest to posttest (McArdle & Nesselroade, 1994; McArdle & Prindle, 2008; McArdle, 2001, 2009). A simple LDS model is shown in Fig. 1. In this model, variable x is measured on two occasions. The second measurement x_2 is modeled as a function of the first measurement x_1 and a difference score Δx. The difference score Δx represents the part of x_2 not related to x_1—that is, change that occurred between the two time points.

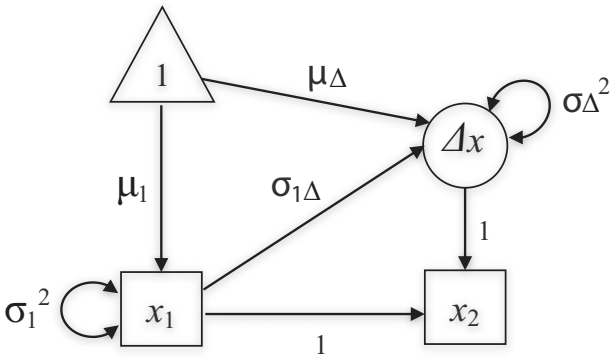

Fig. 1. Latent difference score model. *Note*. Boxes represent observed variables, the triangle represents a constant. Single-headed arrows indicate regression coefficients, intercepts, or means; double-headed arrows indicate variances, the circle represents latent (or unobserved) variable, which is created to indicate the change that occurred between two time points.

Mathematically, this is represented by the simple equation $x_2 = 1 \times x_1 + 1 \times \Delta x$. Each score on the right-hand side of this equation is pre-multiplied by a factor of 1; these correspond to the regression coefficients shown in the figure. The other parameters in this model are: (a) μ_1, the mean of x_1; (b) σ_1^2, the variance of x_1; (c) μ_Δ, the mean of the difference score Δx; (d) σ_Δ^2, the variance of the difference score; and (e), the covariance between x_1 (the initial status) and the resulting change represented by Δx. This SEM representation of difference scores has many advantages over traditional models of change (e.g., ANOVA, gain scores, and residual change scores): model parameters are explicitly defined, researchers have more flexibility in testing change hypotheses, information on global model fit is available, and unreliability in measurement is removed from the difference scores (McArdle, 2009).

Separate models were fit for victimization and bullying. A path diagram of the model is presented in Fig. 1. As discussed in the Method section, we used four variables from the OBVQ measure to capture bullying and victimization at pretest and posttest. The four variables were used to create latent constructs and thus account for measurement error. The variation in each indicator was decomposed into student- and classroom-level components—these are represented by the circles in Fig. 1. The difference scores were regressed onto the initial measurement variables. This specification removes any influence of initial scores on subsequent change and is recommended when the change process (i.e., those resulting from intervention) occurs after the initial time point (McArdle, 2009). At the classroom level, the difference score is predicted by the set of implementation variables and at the student level the difference score is predicted by a set of student demographics. All predictors were left uncentered.

The data were analyzed with Mplus 7.0 (Muthén & Muthén, 1998–2012). Parameters were estimated using a robust FIML estimator (i.e., MLR). In addition, variation at the school level was controlled for by using the TYPE = COMPLEX option—as classrooms were nested within schools, this option was used so that standard errors would not be underestimated at the classroom level. Global model fit of the measurement models was evaluated according to recommendations by Hu and Bentler (1999): (a) a standardized root mean square residual (SRMR) < .08; (b) a comparative fit index (CFI) > .95; and (c) a root mean square error of approximation (RMSEA) < .06. These cut-off values are only rough guidelines. The CFI and RMSEA were calculated separately for each level according to the partially-saturated model fit approach outlined by Ryu and West (2009). All significance tests were conducted using a Type I error rate of .05 except where noted.

3. Results

3.1. Descriptive statistics

Means and standard deviations for and correlations between student-level variables are shown in Table 1. Moderate stability from pretest to posttest (across 12 months) was observed for victimization ($r = .52$) and bullying ($r = .44$). Victimization and bullying were positively correlated within measurement occasions ($r = .54$ and .50) for pretest and posttest as well as across occasions ($r = .24$ and .34). Boys were victimized more and bullied other students more at both time points. Older students were bullied more at pretest. Gender was balanced in the sample ($M = .51$) and the mean age was 9.97. Intraclass correlation coefficients (ICC), which quantify the amount of variation between groups in nested data, are also provided in Table 1. The ICCs were moderate in size for bullying and victimization (range = .07–.15) and comparable to those commonly seen in educational research (Bliese, 2000). Almost all of the variance in age was located at the classroom level (ICC = .95), which was not surprising as the groups were defined by grade level. The ICC for gender was negligible (ICC = .01), and thus gender was omitted as a group level variable.

Classroom-level descriptive statistics are provided in Table 2. School-level ICCs for the fidelity variables ranged between .44 and .48, suggesting that implementation varied considerably between schools as well as classrooms. The classroom-level correlations between victimization and bullying were uniformly higher compared to student-level estimates. The average time of lesson preparation correlated negatively with bullying and victimization at all measurement occasions. Correlations between lesson adherence and bullying and victimization were all negative, but none were statistically significant. Duration of lessons was positively correlated with preparation time and lesson adherence; preparation time and lesson adherence were uncorrelated. Although statistically significant, the positive correlations between duration of lessons and the other two implementation variables were not strong. Therefore, including these three facets of program implementation as separate predictors in the model was justified. Grade level was negatively associated with bullying and victimization and positively associated with the preparation time of teachers as well as the duration of

Table 1
Student-level descriptive statistics at pretest and posttest.

Variable	*M*	*SD*	ICC	Vic1	Bul1	Vic3	Bul3	Boy
Vic1	[a]	.74	.12[b]	–				
Bul1	[a]	.41	.14[b]	.54**	–			
Vic3	[a]	.67	.09[b]	.52**	.34**	–		
Bul3	[a]	.46	.07[b]	.24**	.44**	.50**	–	
Boy[c]	.51	.50	.01	.10**	.19**	.07**	.16**	–
Age[c]	9.97	1.77	.95	.09	.32	.07	.22**	−.01

Notes. ICC = Intraclass correlation coefficient (at the classroom level). Vic1 = Victimization at wave 1. Bul1 = Bullying at wave 1. Vic3 = Victimization at wave 3. Bul3 = Bullying at wave 3. Boy = Gender (0 = girl; 1 = boy). Age = Age of student. ** $p < .01$.

[a] Means for constructs that vary at both levels are defined at the between-group level.

[b] Average ICC value of the four observed indicators per construct.

[c] Defined as student-level variables only.

Table 2
Classroom-level correlations and descriptive statistics at pretest and posttest.

Variable	*M*	*SD*	ICC[a]	Vic1	Bul1	Vic3	Bul3	Prep	Adhe	Dura	Grade
Vic1	.79	.36	–	1.00							
Bul1	.33	.24	–	.98 **	1.00						
Vic3	.63	.28	–	.98 **	.94 **	1.00					
Bul3	.27	.18	–	.98 **	.95 **	.99 **	1.00				
Prep	29.51	16.17	.48	−.24 **	−.22 **	−.30 **	−.27 **	1.00			
Adhe	.68	.20	.47	−.01	−.01	−.06	−.05	−.04	1.00		
Dura	79.18	19.71	.44	−.17 **	−.18	−.19 **	−.19 **	.33 **	.25 **	1.00	
Grade	3.48	1.69	.78	−.80 **	−.85 **	−.82 **	−.82 **	.25 **	.03	.20 **	1.00

Notes. Vic1 = Victimization at wave 1. Bul1 = Bullying at wave 1. Vic3 = Victimization at wave 3.
Bul3 = Bullying at wave 3. Prep = Average amount of preparation per lesson in minutes.
Adhe = Proportion of curriculum tasks completed. Dura = Average duration of lessons in minutes.
Grade = Classroom grade. ** $p < .01$.
[a] School-level ICC.

lessons. With regard to variable means, both bullying and victimization decreased approximately one-half of a standard deviation by posttest; this finding has been reported elsewhere (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; 2013).

3.2. Measurement models

As noted earlier, the four items from the OBVQ measuring verbal, exclusionary, physical, and manipulative victimization and bullying were used as indicators of the two latent outcome variables—victimization and bullying. Prior to estimating the LDS models, we conducted a series of multilevel confirmatory factor analyses (MCFA) to ensure satisfactory measurement of the outcomes. In these models, covariates were excluded and the relationships between latent variables were not directional. Separate models were fit for victimization and bullying. Fit indices for the victimization model were lower compared to the bullying model, but most were in acceptable ranges at both levels, $\chi^2_{Student}(15) = 145.32$, $p < .001$, SRMR $_{Student} = 0.02$, RMSEA $_{Student} = 0.03$, CFI $_{Student} = .90$ and $\chi^2_{Classroom}(15) = 62.69$, $p < .001$, SRMR $_{Classroom} = 0.03$, RMSEA $_{Classroom} = 0.09$, CFI $_{Classroom} = .97$. The bullying model fit the data well at the student level, $\chi^2_{Student}(15) = 76.76$, $p < .001$, SRMR $_{Student} = 0.03$, RMSEA $_{Student} = 0.02$, CFI $_{Student} = .97$, and at the classroom level, $\chi^2_{Classroom}(15) = 13.20$, $p = .59$, SRMR $_{Classroom} = 0.05$, RMSEA $_{Classroom} = 0.00$, CFI $_{Classroom} = 1.00$.

In addition to sufficient model fit, the analysis of longitudinal data requires invariance in measurement parameters over time to ensure the same construct is measured over the course of a study (Vandenberg & Lance, 2000). Therefore, we tested for measurement invariance of item factor loadings and item intercepts using nested model comparisons. Because the chi-square difference test ($\Delta\chi^2$) is sensitive to sample size, we reduced the acceptable Type I error rate to .01 (Little, 2013) and also examined change in the CFI index (ΔCFI) for which a difference of .01 or greater is considered indicative of measurement non-equivalence (Cheung & Rensvold, 2002). Loading invariance was tested first by constraining factor loadings to be equal across time—the so-called *weak-invariant* model. For victimization, the chi-square difference test was not statistically significant, and model fit actually improved according to the CFI at

Table 3
Latent difference score model: student-level results (N = 7413).

		Victimization			Bullying		
Type	Parameter	Est	Std	*SE*	Est	Std	*SE*
Factor	Item 1[a]	1.0	(.73)	–	1.0	(.67)	–
Loadings	Item 2	0.73**	(.63)	0.02	0.76**	(.56)	0.05
Pretest	Item 3	0.68**	(.61)	0.02	0.87**	(.61)	0.04
	Item 4	0.72**	(.67)	0.02	0.56**	(.57)	0.05
Factor	Item 1[a]	1.0	(.74)	–	1.0	(.74)	–
Loadings	Item 2	0.73**	(.62)	0.02	0.76**	(.64)	0.05
Posttest	Item 3	0.68**	(.62)	0.02	0.87**	(.70)	0.04
	Item 4	0.72**	(.66)	0.02	0.56**	(.61)	0.05
Regression	Pre → Post[a]	1.0	–	–	1.0	–	–
Coefficients	Diff → Post[a]	1.0	–	–	1.0	–	–
	Pre → Diff	−0.53**	(−.56)	0.02	−0.57**	(−.50)	0.04
	Boy → Pre	0.15**	(.10)	0.03	0.15**	(.19)	0.02
	Age → Pre	0.01	(.03)	0.02	0.07	(.31)	0.04
	Boy → Diff	0.02	(.02)	0.02	0.08**	(.09)	0.01
	Age → Diff	0.00	(.01)	0.01	0.02	(.09)	0.01
Residual	Pre	0.54**	(.99)	0.04	0.14**	(.87)	0.02
Variances	Post[a]	0.00	–	–	0.00	–	–
	Diff	0.34**	(.69)	0.02	0.16**	(.78)	0.02

Note. Standardized parameter estimates are shown in parentheses. Est = Parameter estimate. Std = Standardized estimate. *SE* = Standard error.
[a] Parameter fixed for model identification.
** $p < .01$.

the student level, $\Delta\chi^2_{\text{Student}}(3) = 1.61$, $p = .66$, $\Delta\text{CFI}_{\text{Student}} = -.014$, and at the classroom level, $\Delta\chi^2_{\text{Classroom}}(3) = 1.31$, $p = .73$, $\Delta\text{CFI}_{\text{Classroom}} = -.004$. The observed positive change in CFI is likely a sign of improved model parsimony. Thus, the weak-invariant model was supported. For bullying, constraints at the student level did not result in a significant loss of model fit, $\Delta\chi^2_{\text{Student}}(3) = 11.40$, $p = .01$, $\Delta\text{CFI}_{\text{Student}} = .003$. At the classroom level, $\Delta\text{CFI}_{\text{Classroom}}$ was .017. The constraints, however, were acceptable according to the chi-square difference test, $\Delta\chi^2_{\text{Classroom}}(3) = 8.34$, $p = .04$, and the model still fit the data quite well ($\text{CFI}_{\text{Classroom}} = .98$). Therefore, we decided the invariance constraints for bullying were tenable at the classroom level as well. The equality of item intercepts—specified in a *strong-invariant* model—was not required for interpreting parameters of interest in the LDS models (i.e., regression coefficients). Still, we tested the strong-invariant model (defined at the classroom level) and found the intercepts could be considered invariant for both bullying, $\Delta\chi^2_{\text{Classroom}}(3) = 7.76$, $p = .05$, $\Delta\text{CFI}_{\text{Classroom}} = .016$, and victimization, $\Delta\chi^2_{\text{Classroom}}(3) = 1.77$, $p = .62$, $\Delta\text{CFI}_{\text{Classroom}} = -.004$.

3.3. Latent difference score model: victimization

Having determined a satisfactory measurement model with invariant item parameters, the final LDS model for victimization was estimated. The student-level LDS model fit the data well. The chi-square value was significantly different from zero, $\chi^2_{\text{Student}}(30) = 351.84$, $p < .001$. However, alternative fit indices were all in acceptable ranges, $\text{SRMR}_{\text{Student}} = 0.03$, $\text{RMSEA}_{\text{Student}} = 0.03$, $\text{CFI}_{\text{Student}} = .98$. Parameter estimates and standard errors for the student-level model are presented on the left-hand side of Table 3. Standardized factor loadings were all above .50 and significantly different from zero (see Factor Loadings Pretest and Factor Loadings Posttest sections).

Regarding the latent regressions (see Regression Coefficients section in Table 3), a significant negative effect was found for the effect of pretest scores on the latent difference scores at posttest (standardized $\beta = -0.56$). This means that higher pretest levels of victimization were associated with larger decreases in victimization over the span of the study. Covariate (gender, age) effects on pretest scores and on the latent difference scores are also provided in Table 3 (see Regression Coefficients section). Consistent with previous studies (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; 2013), boys were victimized significantly more at pretest (standardized $\beta = 0.10$) than girls within the same classroom. There were no gender effects, however, on the change between pretest and posttest. Comparative age within a classroom was not a significant predictor of being victimized at pretest nor associated with decreases in victimization over time.

At the classroom level, the model for victimization also fit the data reasonably well, $\chi^2_{\text{Classroom}}(44) = 96.17$, $p < .001$, $\text{SRMR}_{\text{Classroom}} = 0.08$, $\text{RMSEA}_{\text{Classroom}} = 0.05$, $\text{CFI}_{\text{Classroom}} = .93$. Parameter estimates and standard errors for the classroom-level model are presented on the left-hand side in Table 4. The standardized factor loadings—all close to 1 in value—were much higher compared to the student level (see Factor Loadings Pretest and Factor Loadings Posttest sections), which is common in MSEM (Muthén, 1994).

With regard to the dose–outcome relation (see Regression Coefficients section on the left-hand side of Table 4), two aspects of implementation had significant effects on the posttest difference scores: Lesson adherence and the average preparation time per lesson. For each standard deviation increase in the degree of lesson adherence—approximately 20% of all possible activities—victimization would be expected to decrease by an additional 0.17 standard deviations. Likewise, for each standard deviation increase in

Table 4
Latent difference score model: classroom-level results (J = 417).

		Victimization			Bullying		
Type	Parameter	Est	Std	*SE*	Est	Std	*SE*
Factor	Item 1[a]	1.0	(.94)	–	1.0	(.95)	–
Loadings	Item 2	0.87**	(.99)	0.05	0.67**	(.95)	0.05
Pretest	Item 3	1.29**	(.99)	0.09	1.18**	(.99)	0.07
	Item 4	0.57**	(.97)	0.03	0.50**	(.94)	0.05
Factor	Item 1[a]	1.0	(.90)	–	1.0	(.95)	–
Loadings	Item 2	0.87**	(.96)	0.05	0.67**	(.80)	0.05
Posttest	Item 3	1.29**	(.99)	0.09	1.18**	(.99)	0.07
	Item 4	0.57**	(.92)	0.03	0.50**	(.89)	0.05
Regression	Pre → Post[a]	1.0	–	–	1.00	–	–
Coefficients	Diff → Post[a]	1.0	–	–	1.00	–	–
	Pre → Diff	−0.31**	(−1.03)	0.05	−0.36**	(−0.36)	0.08
	Grade → Pre	−0.15**	(−0.79)	0.02	−0.11**	(−0.82)	0.04
	Prep → Diff	−0.00**	(−0.21)	0.00	−0.00	(−0.14)	0.00
	Adhe → Diff	−0.08*	(−0.17)	0.04	−0.03	(−0.07)	0.04
	Duration → Diff	0.00	(0.08)	0.00	0.00	(0.02)	0.00
	Grade → Diff	−0.01	(−0.17)	0.01	−0.01	(−0.12)	0.02
Residual	Pre	0.04**	(.38)	0.01	0.02**	(.32)	0.01
Variances	Post[a]	0.00	–	–	0.00	–	–
	Diff	0.00	(.18)	0.00	0.00**	(.35)	0.00

Note. Standardized parameter estimates are shown in parentheses. Est = Parameter estimate. Std = Standardized estimate. SE = Standard error. Prep = Average amount of preparation per lesson in minutes. Adhe = Proportion of curriculum tasks completed. Duration = Average duration of lessons in minutes.

[a] Parameter fixed for model identification.

* $p < .05$.

** $p < .01$.

preparation time—approximately 16 min per lesson—victimization would be expected to decrease by an additional 0.21 standard deviation units. Lesson duration was not significantly predictive of decreases in victimization. Finally, grade level and pretest levels of victimization predicted reductions in classroom victimization; that is, classrooms containing older elementary school students and with higher initial levels of classroom-wide victimization saw greater subsequent decreases.

3.4. Latent difference score model: bullying

The student-level LDS model for bullying fit the data well, $\chi^2_{Student}$ (30) = 210.95, $p < .001$, $SRMR_{Student} = 0.05$, $RMSEA_{Student} = 0.03$, $CFI_{Student} = .95$. Parameter estimates and standard errors are presented on the right-hand side of Table 3. Standardized factor loadings were all above .50 and significantly different from zero (see Factor Loadings Pretest and Factor Loadings Posttest sections).

The effect of pretest scores on the latent difference score at posttest (see Regression Coefficients section) was significantly different from zero (standardized $\beta = -0.50$). It appears that students with higher pretest levels of bullying compared to their classroom peers reported larger decreases in bullying over time. Similar to the results for victimization, boys bullied significantly more at pretest (standardized $\beta = 0.19$) than girls within the same classroom. The effect of gender on the latent difference scores was also significant such that boys had smaller decreases in bullying—a 0.09 standard deviation difference—compared to girls in the same classroom. Students' age compared to others in a classroom was again a non-significant predictor of pretest levels or subsequent changes in bullying.

At the classroom level for bullying model (Table 4), the SRMR was at the acceptable model fit threshold, $\chi^2_{Classrom}$ (44) = 52.72, $p < .01$, $SRMR_{Classroom} = 0.08$, $RMSEA_{Classroom} = 0.02$, $CFI_{Classroom} = .97$. Fit was acceptable according to the RMSEA and CFI, and modification indices did not reveal any salient sources of misspecification. Therefore, we interpreted the bullying model as is. Parameter estimates and standard errors for the classroom-level model are presented on the right-hand side of Table 4. The classroom-level indicators were again highly reliable according to the standardized factor loadings (see Factor Loadings Pretest and Factor Loadings Posttest sections).

None of the classroom-level dose–outcome relations were statistically significant in bullying reduction (see Regression Coefficients section). The standardized regression weights for the proportion of tasks completed ($\beta = -0.07$) and the average preparation time per lesson ($\beta = -0.14$) were in expected directions but were smaller than in the victimization model. Regarding other model effects, grade level and pretest levels of aggregate bullying predicted reductions in classroom-level bullying over time, as in the victimization model.

4. Discussion

Scholars in the field of prevention have clearly demonstrated that programs often are not implemented with the same standard (e.g., Dane & Schneider, 1998; Durlak & DuPre, 2008). In school-based effectiveness studies, an important question is to what extent teachers, who are often responsible for program delivery, have actually put the program in practice—with good or poor fidelity. Prior to the present study, very little research on bullying interventions focused on this issue. More specifically, there have been only few studies examining whether some aspects of implementation are more relevant than others in maximizing positive outcomes in terms of reductions in bullying problems (Hirschstein et al., 2007; Low et al., 2014). Although Kärnä, Voeten, Little, Poskiparta, Alanen, et al., 2011 reported a school-level association between the KiVa antibullying curriculum delivered and reduction in victimization, the present study took a step forward in examining several aspects of implementation and relating them to outcomes at the classroom level.

4.1. Degree of implementation

During the 9 months of intervention, the average degree of lesson adherence was almost 70%, which can be considered satisfactory. There was significant variation between schools with regard to implementing the KiVa program, as indicated by the high ICCs. Thus, the commitment to use the curriculum is not only a matter of individual teachers but depends on the school context.

Regarding victimization (Hypotheses 1a–1c), higher lesson adherence predicted a larger reduction in victimization as expected. Many of the tasks included in the KiVa student lessons (Salmivalli, Poskiparta, Tikka, & Pöyhönen, 2013; Salmivalli, Pöyhönen, & Kaukiainen, 2012) emphasize collaborative learning—providing opportunities for student-to-student interaction (e.g., small-group and dyadic discussions, brainstorming, and role-play) instead of teacher-centered practices. During the implementation process across the year, a set of class rules are adopted; students gradually learn to know which behaviors constitute bullying and begin to understand the role of the group in maintaining or stopping it. Thus, with a larger proportion of tasks encountered, possibilities to practice constructive responses to bullying (whether experienced as a target or witnessed) and integrating these skills into daily interactions become more likely.

A second measure of implementation indicated the average duration allocated for lesson delivery. The average teacher-reported duration was nearly 80 min per lesson, indicating that teachers on average reduced only 10 min from the recommended duration for a KiVa lesson. Similar to the findings of Sapouna et al.'s (2010) evaluation of an online antibullying intervention, we found no dose–outcome relation in terms of the time used. This finding suggests that the amount of content delivered may be more important than the time used for lesson delivery. The duration of student lessons may consist of other things than actual antibullying content (e.g., transitions between lessons, getting students' attention, and reacting to misbehavior). In a recent smoking prevention study, observers noted that teachers devoted most of the instructional time—varying between 36 and 60 min—to classroom management and getting organized for the delivery of the intervention (Goenka et al., 2010). Accordingly, teacher-reported duration may not be

a strong measure of the degree to which students were exposed to program content. Moreover, the positive classroom-level correlation between duration and adherence suggests that adherence cannot be optimal with minimal duration (time needs to be invested to lesson delivery in order to meet satisfactory levels of adherence). When the overlap between the two is controlled for, however, adherence to lesson content (as well as preparation of lessons) is more important for success than duration as such. It is important to measure several aspects of implementation fidelity during intervention delivery in order to gain insight into dose–outcome relations (Gearing et al., 2011) and especially the unique predictors of positive outcome.

In addition to lesson adherence, the time teachers spent preparing the lessons was related to reductions in victimization. When considering preparation as a tool for effective teaching, it is worthwhile to discuss the nature of teaching as well as the mechanism by which teaching quality is improved by preparation. Namely, annotated lesson plans enhance teachers' knowledge about what to teach and the skills to do it effectively (Davis & Krajcik, 2005; Dunn et al., 2010; Morris & Hiebert, 2011). Preparation enables teachers to build knowledge of the intervention model (e.g., what are the learning objectives), confidence with implementation techniques and strategies (e.g., how to achieve learning objectives), and skills to generalize the content outside of planned lessons (e.g., integrating curriculum experiences), all of which reflect the standards of program delivery. Leinhardt and Greeno (1986) have defined teaching as a complex cognitive skill based on knowledge about both the content to be taught as well as how to construct and deliver a lesson. Also, Shulman (1986) has pointed out that the knowledge base for teaching contains content as well as pedagogical reasoning and action. With respect to knowledge related to bullying, Bauman and Del Rio (2005) have argued that teachers do not have a clear understanding of the nature of bullying. When teachers devote time to reading the KiVa manual, however, they are exposed to current research-based knowledge on bullying. Increased knowledge and confidence with the content can influence students' activity—for example, students' willingness to present different solutions for stopping bullying—and to practice skills needed in intervening and supporting victimized peers. Importantly, the way in which students perceive, interpret, and process information in the instructional situation—including content and social processes—determines what they will learn (De Corte, 2000; Shuell, 1996).

Contrary to our Hypotheses (2a–2c), none of the three implementation aspects had a significant effect on reducing bullying perpetration over time. Although this was somewhat surprising, it is in accordance with some previous findings suggesting that reducing bullying is more challenging than reducing experienced victimization (for reviews, see Merrell et al., 2008; Vreeman & Carroll, 2007). Whereas the effects of lesson adherence and lesson preparation on bullying were in the expected direction, the impact of lesson duration was practically zero. For some classrooms, the desired effects were not obtained because there was little or no bullying–victimization problems at baseline (i.e., most children were not repeatedly victimized or bullied others) thus leaving no room for improvement. Prevalence of self-reported bullying was lower than prevalence of self-reported victimization, as shown by the classroom means. Therefore, it may have been more difficult to find effects on bullying than on victimization. Another plausible explanation for the non-significant effects on bullying can be associated with the short duration (i.e., one school year) of the trial (see Brown et al., 2011), which may not have been enough to change the behavior of bullies, especially if they were powerful and popular (Garandeau, Lee, & Salmivalli, 2014; Rodkin, Farmer, Pearl, & Van Acker, 2006; Vaillancourt & Hymel, 2006). Some students involved in bullying may simply have decreased the number of peers they targeted (in other words, some victims were able to escape the victim role). In addition, some former victims may have felt less victimized if fewer classmates reinforced the bullies during bullying incidents, while more classmates communicated support for the victims. Another interpretation might be increased coping competence among the victims. If true, these are all desirable outcomes, even when it turns out impossible to prevent all bullying. Finally, it should be noted that the KiVa program also includes indicated actions by which both victims and perpetrators are targeted; such actions might be more likely to reduce bullying perpetration than the universal classroom-level components of the program, which were the focus of the present study.

With respect to individual level factors related to bullying problems, our findings at pretest showed that boys are more often involved in bullying as victims or bullies than girls (see also, Card, Stucky, Sawalani, & Little, 2008; Solberg & Olweus, 2003). Furthermore, the reduction in bullying others was also significantly associated with being a boy; boys showing on average less reduction than girls.

4.2. Limitations and further research

There were some limitations in the current study. One methodological limitation is that the self-report measures of bullying and victimization were ordinal. FIML estimation is based on the assumption of continuous and multivariate normally-distributed data. In this study, reliance was placed on the robustness of the estimation method used (robust FIML estimation) for addressing violations of these distributional assumptions. Some research supports such practices in single-level SEM—for instance, when five or more response categories are available and observed frequencies are symmetric (Rhemtulla, Brosseau-Liard, & Savalei, 2012)—but the use of robust FIML estimation with ordinal data in the MSEM case has not yet been explored.

Second, all the measures in this study were based on self-reports; peer-reports and observational measures of teacher actions were not used. Notably, we tested fidelity effects on peer-reports with a restricted sample excluding students in Grades 1, 2, and 3—peer reports were not collected in these grades—but no significant relations were found. With respect to the teacher reports, significant consistency across observation ratings and teacher self-reports has been found (Domitrovich et al., 2010; Lillehoj et al., 2004; Moskowitz, Schaps, & Malvin, 1982). However, implementation measures based on teachers' self-reports may be inflated relative to observer-reported ratings (Domitrovich et al., 2010; Lillehoj et al., 2004; Melde et al., 2006). Although observational data would have improved our assessment of fidelity, this method was not possible given the large national sample obtained. More empirical data on the quality of interactions between teachers and students in terms of emotional and instructional support during program delivery would inform our research and increase knowledge regarding implementation standards.

Furthermore, 20% of the teachers did not provide data on curriculum implementation. As a result of the attrition, the relatively high implementation rate may be inflated. On the other hand, teachers who dropped out of the study may have followed parts of the curriculum but failed to return booklets due to administrative shortcomings such as lack of coordination, organizational commitment, or personnel turnover. For some schools the decision to volunteer in the evaluation study was probably collegial (e.g., most of the teachers prioritized antibullying work), whereas some schools may have joined on the basis of the principal's decision only. Previous research suggests that teachers' motivation and perceptions of bullying (Kallestad & Olweus, 2003) as well as the school context such as the principal's commitment and support (Ahtola, Haataja, Kärnä, Poskiparta, & Salmivalli, 2013) play an important role in investing resources for antibullying work in classrooms.

A fourth limitation relates to individual students' attendance at lessons. Namely, student-level exposure to the intervention was not controlled for due to the universal nature of lessons (targeting at all students in classrooms). The presence of high-status children may be an important precondition for influencing group norms and bullying behavior in addition to implementation fidelity. Increases in pro-victim behavior may be more likely to happen through popular and well-liked students (Caravita, Di Blasio, & Salmivalli, 2009; Hodges, Malone, & Perry, 1997; Pöyhönen, Juvonen, & Salmivalli, 2010). Therefore, keeping records of student attendance should be considered standard practice in further evaluation studies (see also Ferrer-Wreder et al., 2010).

Despite these limitations, this study adds to our understanding about the effects of universal actions, targeting not only individual, "problematic" children but the peer relations within classrooms. Given that teachers are often unaware of bullying, especially when it is indirect, such as social exclusion (Bauman & Del Rio, 2005; Bradshaw, Sawyer, & O'Brennan, 2007; Craig, Henderson, & Murphy, 2000; Smith & Shu, 2000), educative action plays an important role in conveying the message that bullying is wrong and all students are responsible for stopping it. Prior to this study, very little was known whether students whose teachers show commitment and compliance for bullying prevention program benefit significantly more than students with less committed teachers.

4.3. Conclusion

Our results support the existing literature showing that better implementation of a prevention program is associated with more positive effects obtained (e.g., Durlak & DuPre, 2008). Two aspects of implementation of the KiVa antibullying curriculum stand out as predictors of better outcomes: lesson adherence and lesson preparation. Specifically, reductions in victimization were greater when more curriculum activities were completed and when teachers adequately prepared the lessons. Even if the associations between implementation and outcome were modest in size, they are important as they show that teachers' actions can have an influence on the changes obtained—a rare finding in bullying prevention studies.

We strongly recommend that further attention be paid to the fidelity of implementation of antibullying curricula when the effectiveness of programs is being evaluated. Teachers who recognize their responsibility to educate students about the social dynamics related to bullying are likely to reduce victimization in their classrooms. Finally, 100% of implementation may be unrealistic given constraints on teacher resources as well as more general school resources, but small adaptations may strengthen the ownership of the program and increase the likelihood of using the program in the future. If teachers share the goals of the program and have up-to-date knowledge about bullying, it is likely to improve the quality of universal actions and benefits students even when all tasks cannot be delivered.

Appendix A. Missing values

The data were taken from two previous studies on the effectiveness of the KiVa antibullying program. The first implementation of the program concerned Grades 4 to 6 of Finnish elementary schools (see Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011). Schools were randomly assigned to an experimental or control condition. From the teachers in the experimental condition, we collected data about the extent to which they implemented the program as intended. For the present study, we could only use the 39 schools in the experimental condition to investigate the association between the degree of implementation and student outcomes in terms of self-reported bullying and victimization. The second implementation of the KiVa program included Grades 1 to 3 of Finnish elementary schools (Kärnä et al., 2013). We chose to use the data from the 38 experimental schools participating in this study. Both files combined included 8452 students of 439 teachers in 77 schools.

In both studies, several patterns of missing values were observed. These were discussed in appendices available with the two articles. In the present appendix, we discuss the missing values in the variables selected as dependent and independent variables for studying the relation between degree of implementation and student outcomes: (a) self-reported bullying and victimization assessed at pretest and posttest and (b) degree of implementation assessed by asking teachers. In addition, some background variables were used (gender, grade level, and age of student). As noted in the Method section, data from 1039 students were deleted—some because there was no parental consent for them and thus they did not participate in the measurements; some because they, for other reasons, did not participate in the pretest nor the posttest; and some because they participated only in the pretest and left the school before the start of the intervention.

Concerning the dependent variables, missing values were due to the longitudinal design and therefore attrition. Full data on self-reported bullying and victimization were available for, respectively, 5105 and 5114 students from 315 teachers in 76 schools. Two main patterns of missing values were observed: (1) the pretest data were missing (about 21% of the students), (2) the posttest data were missing (about 10% of the students).

The high percentages missing at the pretest were primarily due to the fact that there was no pretest in the first grade, although we nevertheless included the first-graders in the present analyses. In Grades 2 to 6, the pretest data were missing for only 7.8% of the

students. There were two reasons why first-graders could not take part in the pretest. First, they were not yet in the schools at the time of the pretest (i.e., the end of the school year preceding the start of program implementation). Second, they might lack the reading skills needed to respond to the questionnaire. These reasons do not appear related to bullying. In Grades 2 to 6, the main reason for the missing pretest values was that some students entered the schools after the time of the pretest (i.e., at the beginning of or during the school year that program implementation took place). Newcomers were somewhat more involved in bullying than other children (Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al., 2011; Grades 4–6). Missing posttest values were due to whole classrooms that did not take the posttest or to individual students that were absent at the day of testing or did not answer all questions. In general, the students with missing posttest scored on average higher on self-reported bullying and victimization at pretest compared with students that had valid scores on the posttest.

Next, we turn to missing values on the independent variables. From the sample of 417 teachers included in the analyses, 332 (80%, or 76% of the total number of teachers in the intervention schools) returned the lesson booklets, from which all implementation variables were derived and used in the analyses. Of the 439 teachers in intervention schools, 22 were excluded from the analyses because of the missing values of their students. None of these teachers co-operated in the implementation study. Teachers who took part in the pre-implementation training were more inclined to return the booklets; 79% of the teachers provided information on whether they attended training, or not, and of these, 83% of teachers with one day of training, and 88% of the teachers with two days of training. But these differences were not statistically significant. For 329 of the 332 responding teachers, complete data on the three derived independent variables were obtained; for three teachers we had a missing value for the percentage of activities covered. These missing values on implementation variables may be associated with teachers' unwillingness to document implementation or the lack of recall to return the booklets at the end of the school year. The pattern of missing teacher-level variables resulted—at the student level—in implementation data for 4844 students (65% of the students). Students with and without implementation data were compared on self-reported bullying and victimization. Students with missing implementation data had a higher percentage of missing values in student outcome variables at the posttest, increasing from 4.8% for the students with complete implementation data to 31.4% for the students with missing implementation data. This finding confirms that missingness on student outcomes at posttest was to a large extent a classroom phenomenon associated with failure of some teachers to cooperate. In general, bullying and victimization were on average at pretest higher for the students with missing values on implementation variables, but bullying and victimization decreased on average at posttest compared with pretest, and these findings were true both for students with and without missing values on the implementation variables. The decrease for students from teachers with missing implementation information may, however, be biased by the higher percentage of missing values at the posttest for this category of teachers.

The four missing value patterns hardly showed differences on student background variables (gender, age, and grade level). For 4% of the students, information on age was missing. Missing values occurred only in the data for Grades 1 to 3, especially in Grades 1 and 2. In many of these cases, pretest data were also missing. When student age was missing, the percentage of pretest missingness increased from 20% to 39%. Additionally, a higher average degree of bullying and victimization was observed at pretest and less so at posttest compared with the other students. Complete data were available for the other student background variables.

As suggested by Kärnä, Voeten, Little, Poskiparta, Kaljonen, et al. (2011; 2013), the analyses of missing value patterns suggest that missing data must be taken into account in the models directed at investigating the relationship between degree of implementation and student outcomes. This suggestion was achieved by estimating the parameters of the models via full information maximum likelihood (FIML) estimation—a state-of-the-art missing data technique (Schafer & Graham, 2002). Because missing values also appeared in independent variables, these variables were treated as endogenous (i.e., were assumed to be normally distributed), and thus students with missingness on one or more independent variables were included in the analysis. This FIML approach works well (i.e., gives unbiased estimates) when the missing data can be assumed missing completed at random (MCAR) or missing at random (MAR) and when the distributional assumptions for the residuals of the model are met (see, for instance Enders, 2010, Chapter 4). Our analyses of the missing data showed differential attrition to some extent, and it is therefore unlikely that the missing data are MCAR. But MAR is a much less stringent assumption: MAR means that the probability of a missing value does not depend on the missing value itself but on covariates used in the analysis model. MAR implies that all variables related to missingness need to be in the model. One cannot know whether this is true, but it is believed that FIML is rather robust to violations of this assumption (Graham, 2009).

References

Ahtola, A., Haataja, A., Kärnä, A., Poskiparta, E., & Salmivalli, C. (2013). Implementation of anti-bullying lessons in primary classrooms: How important is head teacher support? *Educational Research*, *55*, 376–392. http://dx.doi.org/10.1080/00131881.2013.844941.

Bauman, S., & Del Rio, A. (2005). Knowledge and beliefs about bullying in schools. *School Psychology International*, *26*, 428–442. http://dx.doi.org/10.1177/0143034305059019.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analyses. In K. J. Klein, & W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extentions, and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

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*, 361–382.

Brown, E. C., Low, S., Smith, B. H., & Haggerty, K. P. (2011). Outcomes from a school-randomized controlled trial of Steps to Respect: A bullying prevention program. *School Psychology Review*, *40*, 423–443.

Caravita, S., Di Blasio, P., & Salmivalli, C. (2009). Unique and interactive effects of empathy and social status on involvement in bullying. *Social Development*, *18*, 140–163. http://dx.doi.org/10.1111/j.1467-9507.2008.00465.x.

Card, N. A., Stucky, B. D., Sawalani, G. M., & Little, T. D. (2008). Direct and indirect aggression during childhood and adolescence: A meta-analytic review of gender differences, intercorrelations, and relations to maladjustment. *Child Development*, *79*, 1185–1229. http://dx.doi.org/10.1111/j.1467-8624.2008.01184.x.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5.

Craig, W. M., Henderson, K., & Murphy, J. G. (2000). Prospective teachers' attitudes toward bullying and victimization. *School Psychology International*, *21*, 5–21. http://dx.doi.org/10.1177/0143034300211001.
Cross, D., Hall, M., Hamilton, G., Pintabona, Y., & Erceg, E. (2004). Australia: The Friendly Schools project. In P. K. Smith, D. Pepler, & K. Rigby (Eds.), *Bullying in schools. How successful can interventions be?* (pp. 187–210). Cambridge: Cambridge University Press.
Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*, 23–45. http://dx.doi.org/10.1016/S0272-7358(97)00043-3.
Davis, E. A., & Krajcik, J. S. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher*, *34*(3), 3–14. http://dx.doi.org/10.3102/0013189X034003003.
De Corte, E. (2000). Marrying theory building and the improvement of school practice: A permanent challenge for instructional psychology. *Learning and Instruction*, *10*, 249–266. http://dx.doi.org/10.1016/S0959-4752(99)00029-8.
Domitrovich, C. E., Gest, S. D., Jones, D., Gill, S., & DeRousie, R. M. S. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly*, *25*, 284–298. http://dx.doi.org/10.1016/j.ecresq.2010.04.001.
Dunn, R., Craig, M., Favre, L., Markus, D., Pedota, P., Sookdeo, G., & Terry, B. (2010). No light at the end of tunnel vision: Steps for improving lesson plans. *Clearing House*, *83*, 194–206. http://dx.doi.org/10.1080/00098650903507460.
Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–350. http://dx.doi.org/10.1007/s10464-008-9165-0.
Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*, 405–432. http://dx.doi.org/10.1111/j.1467-8624.2010.01564.x.
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*, 237–256. http://dx.doi.org/10.1093/her/18.2.237.
Dusenbury, L., Brannigan, R., Hansen, W. B., Walsh, J., & Falco, M. (2005). Quality of implementation: Developing measures crucial to understanding the diffusion of preventive interventions. *Health Education Research*, *20*, 308–313. http://dx.doi.org/10.1093/her/cyg134.
Elliott, D., & Mihalic, S. (2004). Issues in disseminating and replicating effective prevention programs. *Prevention Science*, *5*, 47–53. http://dx.doi.org/10.1023/B:PREV.0000013981.28071.52.
Enders, C. K. (2010). *Applied missing data analysis* (1st ed.). New York, NY: The Guilford Press.
Ennett, S. T., Haws, S. W., Ringwalt, C. L., Vincus, A. A., Hanley, S., Bowling, J. M., & Rohrbach, L. A. (2011). Evidence-based practice in school substance use prevention: Fidelity of implementation under real-world conditions. *Health Education Research*, *26*, 361–371. http://dx.doi.org/10.1093/her/cyr013.
Eslea, M., & Smith, P. K. (1998). The long-term effectiveness of anti-bullying work in primary schools. *Educational Research*, *40*, 203–218. http://dx.doi.org/10.1080/0013188980400208.
Fagan, A. A., & Mihalic, S. (2003). Strategies for enhancing the adoption of school-based prevention programs: Lessons learned from the Blueprints for violence prevention replications of the Life Skills Training program. *Journal of Community Psychology*, *31*, 235–253. http://dx.doi.org/10.1002/jcop.10045.
Fekkes, M., Pijpers, F. I. M., & Verloove-Vanhorick, S. P. (2006). Effects of antibullying school program on bullying and health complaints. *Archives of Pediatrics and Adolescent Medicine*, *160*, 638–644. http://dx.doi.org/10.1001/archpedi.160.6.638.
Ferrer-Wreder, L., Cadely, H. S. -E., Domitrovich, C. E., Small, M. L., Caldwell, L. L., & Cleveland, M. J. (2010). Is more better? Outcome and dose of a universal drug prevention effectiveness trial. *The Journal of Primary Prevention*, *31*, 349–363. http://dx.doi.org/10.1007/s10935-010-0226-4.
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*(3), 1–13.
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*, 44–50. http://dx.doi.org/10.1016/j.appdev.2013.10.004.
Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review*, *31*, 79–88. http://dx.doi.org/10.1016/j.cpr.2010.09.007.
Gingiss, P., Roberts-Gray, C., & Boerm, M. (2006). Bridge-It: A system for predicting implementation fidelity for school-based tobacco prevention programs. *Prevention Science*, *7*, 197–207. http://dx.doi.org/10.1007/s11121-006-0038-1.
Goenka, S., Tewari, A., Arora, M., Stigler, M. H., Perry, C. L., Arnold, J. P. S., & Reddy, K. S. (2010). Process evaluation of a tobacco prevention program in Indian schools-methods, results and lessons learnt. *Health Education Research*, *25*, 917–935. http://dx.doi.org/10.1093/her/cyq042.
Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530.
Hahn, E. J., Noland, M. P., Rayens, M. K., & Christie, D. M. (2002). Efficacy of training and fidelity of implementation of the Life Skills Training program. *Journal of School Health*, *72*, 282–287. http://dx.doi.org/10.1111/j.1746-1561.2002.tb01333.x.
Hirschstein, M., Edstrom, L., Frey, K., Snell, J., & MacKenzie, E. (2007). Walking the talk in bullying prevention: Teacher implementation variables related to initial impact of the Steps to Respect program. *School Psychology Review*, *36*, 3–21.
Hodges, E. V. E., Malone, M. J., & Perry, D. G. (1997). Individual risk and social risk as interacting determinants of victimization in the peer group. *Developmental Psychology*, *33*, 1032–1039. http://dx.doi.org/10.1037/0012-1649.33.6.1032.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling—A Multidisciplinary Journal*, *6*, 1–55. http://dx.doi.org/10.1080/10705519909540118.
Jones, S. M., Brown, J. L., & Lawrence Aber, J. (2011). Two-year impacts of a universal school-based social-emotional and literacy intervention: An experiment in translational developmental research. *Child Development*, *82*, 533–554. http://dx.doi.org/10.1111/j.1467-8624.2010.01560.x.
Kallestad, J. H., & Olweus, D. (2003). Predicting teachers' and schools' implementation of the Olweus bullying intervention program: A multilevel study. *Prevention and Treatment*, *6*, 3–21. http://dx.doi.org/10.1037/1522-3736.6.1.621a.
Kam, C. -M., Greenberg, M. T., & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. Promoting Alternative Thinking Skills Curriculum. *Prevention Science*, *4*, 55–63.
Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Alanen, E., & Salmivalli, C. (2011). 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., Alanen, E., & Salmivalli, C. (2013). Effectiveness of the KiVa anti-bullying program: Grades 1–3 and 7–9. *Journal of Educational Psychology*, *105*, 535–551. http://dx.doi.org/10.1037/a0030417.
Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Kaljonen, A., & Salmivalli, C. (2011). A large-scale evaluation of the KiVa antibullying program: Grades 4–6. *Child Development*, *82*, 311–330. http://dx.doi.org/10.1111/j.1467-8624.2010.01557.x.
Kyriakides, L., Kaloyirou, C., & Lindsay, G. (2006). An analysis of the revised Olweus Bully/Victim Questionnaire using the Rasch measurement model. *British Journal of Educational Psychology*, *76*, 781–801. http://dx.doi.org/10.1348/000709905X53499.
Leinhardt, G., & Greeno, J. G. (1986). The cognitive skill of teaching. *Journal of Educational Psychology*, *78*, 75–79. http://dx.doi.org/10.10377/0022-0663.78.2.75.
Lillehoj, C., Griffin, K., & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: Agreement and relation to youth outcomes. *Health Education & Behavior*, *31*, 242–257. http://dx.doi.org/10.1177/1090198103260514.
Little, T. D. (2013). *Longitudinal structural equation modelling*. New York, NY: Guilford Press.
Low, S., Ryzin, M. J. V., 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*, 165–176. http://dx.doi.org/10.1007/s11121-012-0359-1.
Martens, M., van Assema, P., Paulussen, T., Schaalma, H., & Brug, J. (2006). Krachtvoer: Process evaluation of a Dutch programme for lower vocational schools to promote healthful diet. *Health Education Research*, *21*, 695–704. http://dx.doi.org/10.1093/her/cyl082.
McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analyses. In R. Cudeck, S. Du. Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future* (pp. 342–380). Lincolnwood, IL: Scientific Software International, Inc.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *Vol. 60*. (pp. 577–605). Palo Alto: Annual Reviews.
McArdle, J. J., & Nesselroade, J. R. (1994). Structuring data to study development and change. In S. H. Cohen, & H. Reese (Eds.), *Life-span developmental psychology* (pp. 223–267). Hillsdale, NJ: Erlbaum.
McArdle, J. J., & Prindle, J. J. (2008). A latent change score analysis of a randomized clinical trial in reasoning training. *Psychology and Aging*, *23*, 702–719. http://dx.doi.org/10.1037/a0014349.
Melde, C., Esbensen, F. -A., & Tusinski, K. (2006). Addressing program fidelity using onsite observations and program provider descriptions of program delivery. *Evaluation Review*, *30*, 714–740. http://dx.doi.org/10.1177/0193841X06293412.
Merrell, K. W., Gueldner, B. A., Ross, S. W., & Isava, D. M. (2008). How effective are school bullying intervention programs? A meta-analysis of intervention research. *School Psychology Quarterly*, *23*, 26–42. http://dx.doi.org/10.1037/1045-3830.23.1.26.
Mihalic, S., Fagan, A. A., & Argamaso, S. (2008). Implementing the LifeSkills Training drug prevention program: Factors related to implementation fidelity. *Implementation Science*, *3*, 5. http://dx.doi.org/10.1186/1748-5908-3-5.
Morris, A. K., & Hiebert, J. (2011). Creating shared instructional products: An alternative approach to improving teaching. *Educational Researcher*, *40*(1), 5–14. http://dx.doi.org/10.3102/0013189X10393501.
Moskowitz, J. M., Schaps, E., & Malvin, J. H. (1982). Process and outcome evaluation in primary prevention. *Evaluation Review*, *6*, 775–788. http://dx.doi.org/10.1177/0193841X8200600604.
Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376–398. http://dx.doi.org/10.1177/0049124194022003006.
Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. J. Hox, & J. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 15–40). New York, NY: Taylor & Francis.
Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
Olweus, D. (1986). *The Olweus Bully/Victim Questionnaire. Mimeo*. Bergen, Norway: University of Bergen.
Olweus, D. (1993). *Bullying at school: What we know and what we can do*. Oxford: Wiley-Blackwell.
Olweus, D. (2007). *Olweus Bullying Questionnaire. Standard School Report*. Center City: Hazelden Publishing.
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*, *56*, 143–163.
Resnicow, K., Davis, M., Smith, M., Lazarus-Yaroch, A., Baranowski, T., Baranowski, J., & Wang, D. (1998). How best to measure implementation of school health curricula: A comparison of three measures. *Health Education Research*, *13*, 239–250. http://dx.doi.org/10.1093/her/13.2.239.
Rhemtulla, M., Brosseau-Liard, P. E', & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354–373. http://dx.doi.org/10.1037/a0029315.
Rodkin, P. C., Farmer, T. W., Pearl, R., & Van Acker, R. (2006). They're cool: Social status and peer group supports for aggressive boys and girls. *Social Development*, *15*, 175–204. http://dx.doi.org/10.1111/j.1467-9507.2006.00336.x.
Rohrbach, L. A., Graham, J. W., & Hansen, W. B. (1993). Diffusion of a school-based substance-abuse prevention program—Predictors of program implementation. *Preventive Medicine*, *22*, 237–260. http://dx.doi.org/10.1006/pmed.1993.1020.
Ryan, W., & Smith, J. D. (2009). Antibullying programs in schools: How effective are evaluation practices? *Prevention Science*, *10*, 248–259. http://dx.doi.org/10.1007/s11121-009-0128-y.
Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*, 583–601. http://dx.doi.org/10.1080/10705510903203466.
Sainio, M., Veenstra, R., Huitsing, G., & Salmivalli, C. (2011). Victims and their defenders: A dyadic approach. *International Journal of Behavioral Development*, *35*, 144–151. http://dx.doi.org/10.1177/0165025410378068.
Salmivalli, C., Kaukiainen, A., & Voeten, M. (2005). Anti-bullying 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.
Salmivalli, C., Poskiparta, E., Ahtola, A., & Haataja, A. (2013). The implementation and effectiveness of the KiVa antibullying program in Finland. *European Psychologist*, *18*, 79–88. http://dx.doi.org/10.1027/1016-9040/a000140.
Salmivalli, C., Poskiparta, E., Tikka, A., & Pöyhönen, V. (2013). *KiVa: Teacher's manual, Unit 1 (Research into Practice Publication Series, No. 2)*. Turku, Finland: University of Turku, Psychology Department.
Salmivalli, C., Pöyhönen, V., & Kaukiainen, A. (2012). *KiVa: Teacher's manual, Unit 2 (Research into Practice Publication Series, No. 3)*. Turku, Finland: University of Turku, Psychology Department.
Sapouna, M., Wolke, D., Vannini, N., Watson, S., Woods, S., Schneider, W., & Aylett, R. (2010). Virtual learning intervention to reduce bullying victimization in primary school: A controlled trial. *Journal of Child Psychology & Psychiatry*, *51*, 104–112. http://dx.doi.org/10.1111/j.1469-7610.2009.02137.x.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. http://dx.doi.org/10.1037/1082-989X.7.2.147.
Shuell, T. J. (1996). Teaching and learning in a classroom context. In D. C. Berliner, & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 726–764). London, England: Prentice Hall International.
Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.
Smith, J. D., Schneider, B. H., Smith, P. K., & Ananiadou, K. (2004). The effectiveness of whole-school antibullying programs: A synthesis of evaluation research. *School Psychology Review*, *33*, 547–560.
Smith, P. K., & Sharp, S. (1994). *School bullying: Insights and perspectives*. London: Routledge.
Smith, P. K., & Shu, S. (2000). What good schools can do about bullying. *Childhood*, *7*, 193–212. http://dx.doi.org/10.1177/0907568200007002005.
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.
Spoth, R., Guyll, M., Trudeau, L., & Goldberg-Lillehoj, C. (2002). Two studies of proximal outcomes and implementation quality of universal preventive interventions in a community–university collaboration context. *Journal of Community Psychology*, *30*, 499–518. http://dx.doi.org/10.1002/jcop.10021.
Stevens, V., Van Oost, P., & De Bourdeaudhuij, I. (2001). Implementation process of the Flemish antibullying intervention and relation with program effectiveness. *Journal of School Psychology*, *39*, 303–317. http://dx.doi.org/10.1016/S0022-4405(01)00073-5.
Tortu, S., & Botvin, G. J. (1989). School-based smoking prevention—The teacher-training process. *Preventive Medicine*, *18*, 280–289. http://dx.doi.org/10.1016/0091-7435(89)90075-3.
Ttofi, M. M., & Farrington, D. P. (2010). 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.
Vaillancourt, T., & Hymel, S. (2006). Aggression and social status: The moderating roles of sex and peer-valued characteristics. *Aggressive Behavior*, *32*, 396–408. http://dx.doi.org/10.1002/ab.20138.
Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. http://dx.doi.org/10.1177/109442810031002.
Vreeman, R., & Carroll, A. (2007). A systematic review of school-based interventions to prevent bullying. *Archieves of Pediatrics and Adolescent Medicine*, *161*, 78–88. http://dx.doi.org/10.1001/archpedi.161.1.78.
Weare, K., & Nind, M. (2011). Mental health promotion and problem prevention in schools: What does the evidence say? *Health Promotion International*, *26*, I29–I69. http://dx.doi.org/10.1093/heapro/dar075.
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*, 130–143. http://dx.doi.org/10.1016/j.amepre.2007.04.011.