CHILD DEVELOPMENT

Child Development, January/February 2011, Volume 82, Number 1, Pages 311–330

A Large-Scale Evaluation of the KiVa Antibullying Program: Grades 4–6

Antti Kärnä
University of Turku

Marinus Voeten
Radboud University Nijmegen

Todd D. Little
University of Kansas

Elisa Poskiparta, Anne Kaljonen, and Christina Salmivalli
University of Turku

This study demonstrates the effectiveness of the KiVa antibullying program using a large sample of 8,237 youth from Grades 4–6 (10–12 years). Altogether, 78 schools were randomly assigned to intervention (39 schools, 4,207 students) and control conditions (39 schools, 4,030 students). Multilevel regression analyses revealed that after 9 months of implementation, the intervention had consistent beneficial effects on 7 of the 11 dependent variables, including self- and peer-reported victimization and self-reported bullying. The results indicate that the KiVa program is effective in reducing school bullying and victimization in Grades 4–6. Despite some evidence against school-based interventions, the results suggest that well-conceived school-based programs can reduce victimization.

Bullying is a common problem in schools, affecting the lives of a large number of students. It is commonly characterized as systematic abuse of power (Smith & Sharp, 1994). More specifically, bullying is defined as repeated aggressive behavior against a victim who cannot readily defend himself or herself (Olweus, 1999). Victims of bullying often experience insecurity and various forms of psychosocial maladjustment, such as depression and anxiety; they sometimes even exhibit self-destructiveness (for meta-analyses, see Card, 2003; Hawker & Boulton, 2003). For a number of victims, their experiences continue to affect their lives later on in the forms of depression, low self-esteem, and difficulty in trusting other people (Isaacs, Hodges, & Salmivalli, 2008; Olweus, 1994). Not only are victims at risk: Compared to other children, bullies often become involved in delinquency and alcohol abuse (Kaltiala-Heino, Rimpelä, Rantanen, & Rimpelä, 2000; Loeber & Dishion, 1983; Nansel et al., 2001; Nansel et al., 2004; Olweus, 1993a, 1993b). The need to intervene effectively in bullying is thus clear and urgent. Accordingly, numerous antibullying programs have been initiated by researchers, practitioners, and governments. The present study is the first evaluation of a new antibullying program, designed for national use in Finnish comprehensive schools.

Antibullying Programs

Several whole-school intervention programs have been developed to reduce bullying in schools (for reviews, see Baldry & Farrington, 2007; Farrington, & Ttofi, 2009; Ferguson, San Miguel, Kilburn, & Sanchez, 2007; Merrell, Gueldner, Ross, & Isava, 2008; J. D. Smith, Schneider, Smith, & Ananiadou, 2004; P. K. Smith, Ananiadou, & Cowie, 2003; P. K. Smith, Pepler, & Rigby, 2004; Vreeman & Carroll, 2007). A whole-school approach views bullying as a systemic problem with multiple causes at the individual, classroom, and school levels (J. D. Smith et al., 2004). This layered perspective suggests that an intervention must target the entire school context, rather than just individual bullies and victims (J. D. Smith et al., 2004). In this regard, whole-school interventions differ from

This research is part of the KiVa project for developing an antibullying intervention program for the Finnish comprehensive schools. The KiVa project is financed by the Finnish Ministry of Education and Culture. In addition, the present study was supported by the Academy of Finland Grants 134843 and 135577 to Christina Salmivalli. We thank the whole KiVa project team, and especially Marita Kantola and Jonni Nakari, for their contribution in the data-gathering process.

Correspondence concerning this article should be addressed to Antti Kärnä, Department of Psychology, University of Turku, Assistentinkatu 7, 20014 Turun yliopisto, Finland. Electronic mail may be sent to ankarna@utu.fi.

 0009-3920/2011/8201-0021
DOI: 10.1111/j.1467-8624.2010.01557.x

more narrowly focused interventions, such as curriculum interventions, social-skills groups, and counseling (Vreeman & Carroll, 2007).

Most whole-school antibullying programs were inspired by Dan Olweus's first Bergen study. Olweus (1991) utilized a multilayered approach by targeting individual, class, and school levels with different intervention components, such as serious talks with bullies and victims, classroom discussions, and staff meetings. The evaluation was conducted using a cohort-longitudinal design with time-lagged comparisons. With this design, students after the intervention were compared with students from the same grades in the same schools before the intervention. For instance, Grade 5 pretest data served as a baseline against which the posttest data from students in Grade 4 were compared after 12 months of intervention. Most comparisons showed reductions in victimization and bullying rates of 50% or more from the baseline frequency. Substantial decreases also emerged for other antisocial behaviors, such as vandalism, theft, and truancy, in addition to an increase in general satisfaction with school life.

Since the first Bergen project, several effectiveness studies have been conducted in various countries (e.g., Baldry & Farrington, 2004; Cross, Hall, Hamilton, Pintabona, & Erceg, 2004; Frey et al., 2005; O'Moore & Minton, 2004; Pepler, Craig, Ziegler, & Charach, 1994; Pitts & Smith, 1995; Roland, 1989; Salmivalli, Kaukiainen, & Voeten, 2005; Smith & Sharp, 1994; Stevens, De Bourdeaudhuij, & Van Oost, 2000). Unfortunately, these studies have shown very inconsistent results, with the majority of studies nonsignificant, some negative, and only a few with beneficial outcomes (J. D. Smith et al., 2004). J. D. Smith et al. (2004) concluded that the amassed evidence is simply too variable to justify adopting such programs to the exclusion of other procedures.

One potential explanation of the inconsistencies in the evaluation findings is that Olweus's remarkable success is due to the high quality of Scandinavian schools, with, for instance, particularly well-trained teachers (J. D. Smith et al., 2004). In addition, Baldry and Farrington (2007) proposed that the inconsistent findings may be associated with variations in assessment methods and evaluation designs (see also Farrington & Ttofi, 2009; Vreeman & Carroll, 2007).

The variable and often weak results may be, at least partly, explained by methodological weaknesses of the studies (Baldry & Farrington, 2007). Several authors have lamented the methodological problems inherent in the effectiveness studies of antibullying intervention programs (Baldry & Farrington, 2007; J. D. Smith et al. 2004; Vreeman & Carroll, 2007). Somewhat surprisingly, all previous bullying intervention studies lack at least one and often several methodologically important features, such as an appropriate control condition, random assignment, multilevel modeling of hierarchical data, multimethod and multi-informant outcome assessment, psychometrically sound measures, systematic implementation monitoring, proper sample size, attrition analysis or missing data imputation (Baldry & Farrington, 2007; Farrington & Ttofi, 2009; J. D. Smith et al., 2004; Vreeman & Carroll, 2007). As a result, the studies clearly fall short of the standards of evidence required for interventions to be considered efficacious (see Flay et al., 2005, for the standards); therefore, only limited empirical support exists for the effectiveness of school-based antibullying programs (J. D. Smith et al., 2004). Numerous reviews of the effectiveness of the antibullying programs have called for further research using higher methodological standards to rigorously investigate whether such programs actually are effective or not (Baldry & Farrington, 2007; J. D. Smith et al., 2004; Vreeman & Carroll, 2007).

KiVa Antibullying Program

The Finnish Ministry of Education and Culture funded the development and evaluation of a new, national antibullying program named KiVa (an acronym for Kiusaamista Vastaan, ''against bullying''). The program was developed at the University of Turku, in collaboration between the Department of Psychology and the Centre for Learning Research. It was introduced in the intervention schools across Grades 4 through 6 during the 2007–2008 school year.

Theoretical Background of the KiVa Program

KiVa enjoys a multifaceted theoretical background (e.g., Salmivalli, Kärnä, & Poskiparta, 2010a). The program is built on a view of bullying that is based on two lines of research: (a) studies on the social standing of aggressive children in general (e.g., Cillessen & Mayeux, 2004; Rodkin, Farmer, Pearl, & Van Acker, 2000) and bullies in particular (Juvonen, Graham, & Schuster, 2003) and (b) research on participant roles in bullying (Salmivalli, Lagerspetz, Björkqvist, Österman, & Kaukiainen, 1996). Furthermore, social-cognitive theory (Bandura,

1989) is used as a framework for understanding the processes of social behavior.

Recent research suggests that bullying behavior is at least partly motivated by a pursuit of high status and a powerful position in the peer group (e.g., Juvonen & Galván, 2008; Salmivalli & Peets, 2008). Bullying is also a group phenomenon, in which bystanders have an effect on the maintenance of bullying and on the adjustment of the victims (Salmivalli, 2009; Salmivalli et al., 1996). More specifically, bystanders can contribute to the maintenance of bullying by assisting and reinforcing the bully, which provides bullies with the position of power that they seek. On the other hand, defending the victim may make bullying an unsuccessful strategy for attaining and demonstrating high status. KiVa is predicated on the idea that a positive change in the behaviors of classmates can reduce the rewards gained by bullies and consequently their motivation to bully in the first place. KiVa places concerted emphasis on enhancing the empathy, self-efficacy, and antibullying attitudes of onlookers, who are neither bullies nor victims. This strategy is based on sound evidence relating these characteristics to defending and supporting victimized peers (Caravita, DiBlasio, & Salmivalli, 2009; Pöyhönen, Juvonen, & Salmivalli, 2010; Pöyhönen & Salmivalli, 2008; Salmivalli & Voeten, 2004). The aim is to make bystanders show that they are against bullying and to make them support the victim, instead of encouraging the bully. As another equally important component, the KiVa program includes procedures for handling the acute bullying cases that come to the attention of the school personnel.

A prior Finnish bullying intervention study (Salmivalli et al., 2005) was also based on similar principles. That program, however, mainly consisted of teacher education (making the actual program content rather loose) without concrete materials that teachers could utilize when working with students and classrooms. It also lacked a program manual needed for accurate replication.

KiVa Program Components

KiVa includes both universal and indicated actions to prevent the occurrence of bullying as well as to intervene in individual bullying cases (e.g., Salmivalli et al., 2010a, 2010b). The program has three different developmentally appropriate versions for Grades 1–3, 4–6, and 7–9 (i.e., for 7–9, 10–12, and 13–15 years of age).

Universal actions. The KiVa program for Grades 4–6 includes 20 hr of student lessons (10 double lessons) given by classroom teachers during a school year. The central aims of the lessons are to: (a) raise awareness of the role that the group plays in maintaining bullying, (b) increase empathy toward victims, and (c) promote children's strategies of supporting the victim and thus their self-efficacy to do so. The lessons involve discussion, group work, role-play exercises, and short films about bullying. As the lessons proceed, class rules based on the central themes of the lessons are successively adopted one at a time.

A unique feature of KiVa is an antibullying computer game included in the primary school versions of the program. Students play the game during and between the lessons described earlier. The game involves five levels, each of them consisting of three components named: I KNOW, I CAN, and I DO. Students acquire new information and test their existing knowledge about bullying (I KNOW), learn new skills to act in appropriate ways in bullying situations (I CAN), and are encouraged to make use of their knowledge and skills in real-life situations (I DO).

KiVa provides prominent symbols such as bright vests for the recess supervisors to enhance their visibility and signal that bullying is taken seriously in the school and posters to remind students and school personnel about the KiVa program. Schools get presentation graphics they can use to introduce the program for the whole personnel and for parents. Parents also receive a guide that includes information about bullying and advice about what parents can do to prevent and reduce the problem.

Indicated actions. In each school, a team of three teachers (or other school personnel), along with the classroom teacher, addresses each case of bullying that is witnessed or revealed. Cases are handled through a set of individual and small group discussions with the victims and with the bullies, and systematic follow-up meetings. In addition, the classroom teacher meets with two to four prosocial and high-status classmates, encouraging them to support the victimized child.

Training days and school network meetings. Support to implement the program is given to teachers and schools in several ways. In addition to 2 full days of face-to-face training, networks of school teams are created, consisting of three school teams each. The network members meet three times during the school year with one person from the KiVa project guiding the network.

KiVa naturally shares some features with existing antibullying programs, such as Olweus's bullying prevention program (OBPP). These features are shared principles, or ideas, rather than actual program contents. For instance, both OBPP and KiVa include actions at the level of individual students, classrooms, and schools, both tackle acute bullying cases through discussions with the students involved, and both suggest developing class rules against bullying. KiVa, however, has at least three features that, when taken together, differentiate it from OBPP and other antibullying programs. First, KiVa includes a broad and encompassing array of concrete and professionally prepared materials for students, teachers, and parents. Rather than offering ''guiding principles'' or ''philosophies'' to school personnel, it provides them with a whole pack of activities to be carried out with students. Second, KiVa harnesses the powerful learning media provided by the Internet and virtual learning environments. Third, while focusing on the bystanders, or witnesses of bullying, KiVa goes beyond ''emphasizing the role of bystanders,'' mentioned in the context of several intervention programs, by actually providing ways to enhance empathy, self-efficacy, and efforts to support the victimized peers. Furthermore, students' private attitudes are made salient in order to reduce the (often false) impression that ''others think that bullying is OK'' (so-called pluralistic ignorance; see Juvonen & Galván, 2008). Although other programs share some of these features, none of them has assembled these features into the coordinated whole-school, multilayered intervention that is the hallmark of the KiVa program. With regard to research, the clear structure, well-defined content and concrete materials of KiVa make it easy to use and amenable to replication in further studies. These features distinguish KiVa from some other antibullying programs, for which such specific components are not described in sufficient detail to enable accurate replication (Vreeman & Carroll, 2007).

The Present Study

The present study expands knowledge about the effectiveness of antibullying interventions by examining the effects of the new KiVa antibullying program on bullying, victimization, and other key outcomes. We focused in this study on Grades 4–6 because these were included in the first phase (2007–2008) of program evaluation. Results involving Grades 1–3 and 7–9 from the second phase (2008–2009) will be presented in upcoming reports.

We used several outcome measures to assess the effectiveness of the KiVa program. The program effects were examined by comparing intervention-school students with control school students at two time points: in the middle and in the end of the school year (i.e., 4 and 9 months after beginning of program implementation; 7 and 12 months after pretest measures). As the main outcomes, we used self-reported and peer-reported bullying and victimization. We hypothesized that the KiVa program would yield substantial reductions in these problem behaviors. We also expected beneficial changes in other outcomes; specifically, we expected increases in defending victims and decreases in assisting and reinforcing bullies. We also hypothesized that the intervention would increase antibullying attitudes, empathy toward victims, and self-efficacy for defending. Finally, we expected the program to improve students' well-being at school.

The evaluation was done using best practice methodology and stringent standards for effectiveness (Flay et al., 2005). The present study is responsive to Baldry's and Farrington's (2007) call for high-quality evaluations with theoretically grounded interventions, randomized designs, and multiple measures of effectiveness.

Method

Sampling and Design

To recruit schools, letters describing the KiVa project were sent in the fall of 2006 to all 3,418 schools providing basic education in mainland Finland. These included both Finnish-language and Swedish-language schools, because the basic education in Finland is given in both official languages. The letter included information about the goals and content of KiVa and an enrollment form. In this first phase of program evaluation (Grades 4–6), the 275 volunteering schools were stratified by province and language and 78 of them were randomly assigned to intervention or control condition (special-education-only schools were excluded). We oversampled Swedish-language schools (15.3% of the sample schools were Swedish, whereas 9.4% of all Finnish comprehensive schools are Swedish-language schools), but adjusted for this in the analyses. The participating schools were located throughout the country and resembled other comprehensive schools in such characteristics as class size and proportion of immigrant students. As such,

they can be considered representative of Finnish comprehensive schools.

Procedure

The school year in Finland ranges from mid-August to the end of May. Data collection took place three times: in May 2007, December 2007 or January 2008, and May 2008. Students filled out Internet-based questionnaires in the schools' computer labs during regular school hours. The process was administered by the teachers, who were supplied with detailed instructions about 2 weeks prior to data collection. In addition, teachers were offered support through phone or e-mail prior to and during data collection. Teachers distributed individual passwords to the students, who used them to log in to the questionnaire. At the beginning of the session, the term bullying was defined for the students in the way formulated in the Olweus Bully/Victim Questionnaire (Olweus, 1996), which emphasizes the repetitive nature of bullying and the power imbalance between the bully and the victim. Additionally, to remind the students of the meaning of the term *bullying*, a shortened version of the definition appeared on the upper part of the computer screen while the students responded to any bullying-related question. The order of questions, items, and scales was extensively randomized to alleviate any systematic order effect. Students were assured that their answers remain strictly confidential and are not revealed to teachers or parents.

Sample

The 78 participating schools represent all five provinces in mainland Finland. The target sample at Wave 1 included 429 classrooms and a total of 8,237 students in Grades 3–5 (mean ages = 9–11 years). To recruit the children, their parents were sent information letters including a consent form. A total of 7,564 students (91.7% of the target sample) received active consent to participate in the study. One whole school dropped out before the data collection because of problems related to their school facilities. By Waves 2 and 3 some changes in the student composition had taken place, with 251 students leaving the schools and 463 entering them. Between Waves 1 and 2 two control schools (51 students) dropped out, and five more (640 students) between Waves 2 and 3. There were no missing values in predictor variables, and for outcome variables percentages of missing values were not high, except for control schools at Wave 3 (for details on attrition analysis, see http://www.kivakoulu.fi/english). Missing data were imputed using the SAS Proc MI (SAS 9.2; SAS Institute, Cary, NC) utility employing dummy codes for classrooms and for cross-classifications of classrooms as well as all interactions of these dummy codes with study variables. We conducted 100 imputations using the Markov Chain Monte Carlo algorithm. The means of these 100 imputations were used in the analyses (for more details of the imputation process, see Appendix A). Students were excluded from the analyses if: (a) they were denied permission to participate in the study but had somehow answered the questionnaire and (b) they left school after Wave 1. The final sample size for the analyses was 8,166 (4,201 in the intervention and 3,965 in the control condition). Altogether, 50.1% of the respondents were girls and 49.9% boys. Most students were native Finns (i.e., Caucasian), with the proportion of immigrants being 2.4%.

As the evaluation is about the school year in which the intervention took place, we assigned all students to the classrooms they belonged to during that school year. Classroom changes were not taken into account in the models, as the data indicated that about 82% of the classrooms remained at Wave 2 the same as they had been at Wave 1.

Variables and Instrumentation

Self-Reported Bullying and Self-Reported Victimization

The questionnaire started with demographic questions (e.g., gender and age) followed by questions about bullying and victimization. To measure bullying and victimization, we used the global items from the revised Olweus Bully/Victim Questionnaire (Olweus, 1996): ''How often have you been bullied at school in the last couple of months?'' and ''How often have you bullied others at school in the last couple of months?'' Students answered on a 5-point scale (0 = *not at all*, 4 = *several times a week*).

Participant Roles in Bullying Situations and Peer-Reported Victimization

When answering the Participant Role Questionnaire (Salmivalli et al., 1996), students were instructed to think of situations in which someone was bullied. They were presented with items describing different ways to behave in such situations, and they were asked to nominate, from a list of classmates presented on the computer screen, an

unlimited number of classmates that usually behave in the way described in each item. They were allowed also to choose ''no one.'' The 12 items used in this study form four scales reflecting different participant roles: bullying (''Starts bullying,'' ''Makes the others join in the bullying,'' ''Always finds new ways of harassing the victim''), assisting the bully (''Joins in the bullying, when someone else has started it,'' ''Assists the bully,'' ''Helps the bully, maybe by catching the victim''), reinforcing the bully (''Comes around to watch the situation,'' ''Laughs,'' ''Incites the bully by shouting or saying: Show him/her!''), and defending the victim (''Comforts the victim or encourages him/her to tell the teacher about the bullying,'' ''Tells the others to stop bullying,'' ''Tries to make the others stop bullying''). To measure peer-reported victimization, students nominated classmates treated in the following ways: ''He/She is being pushed around and hit,'' ''He/She is called names and mocked,'' ''Nasty rumors are spread about him/her'' (Kärnä, Voeten, Poskiparta, & Salmivalli, 2010). They were allowed to make an unlimited number of nominations, or to answer ''no one.''

Peer nominations received were totaled and divided by the number of classmates responding, resulting in a score ranging from 0.00 to 1.00 for each student on each item. The proportion scores were averaged across the three items for each scale, and the Cronbach's α coefficients were .91 for the bully scale, .90 for the assistant scale, .85 for the reinforcer scale, .91 for the defender scale, and .84 for the victim scale.

Antibullying Attitudes

The original 20-item Provictim scale (Rigby & Slee, 1991) was modified into a 10-item version to better fit the present context. Students responded on a 5-point scale (0 = *I disagree completely*, 4 = *I agree completely*) to items such as: ''It's okay to call some kids nasty names.'' All 10 items loaded highly on one factor in an exploratory factor analysis. After six negatively keyed items were reversely coded, scores on all 10 items were averaged (α = .79).

Empathy Toward Victims

We used a seven-item empathy scale (Pöyhönen, Kärnä, & Salmivalli, 2008) consisting of items such as ''When a bullied child is sad I feel sad as well.'' Students evaluated how often the statements were true for them, responding on a 5-point scale (0 = *never*, 4 = *always*). An exploratory factor analysis supported a single factor. The items were averaged, creating a single empathy score (ranging from 0 to 4), with higher numbers indicating greater empathy toward victims (α = .84).

Self-Efficacy for Defending Behavior

With a new self-efficacy for defending scale (Pöyhönen et al., 2010), students evaluated how easy or difficult it would be for them to defend and support the victim of bullying. The three items used in the scale were derived from the participant role questionnaire items for defending behavior, for instance ''Trying to make the others stop the bullying would be ... '' The answers were given on a 4-point scale (0 = *very difficult for me*, 3 = *very easy for me*). Internal consistency was satisfactory (α = .69), and scores were averaged across the three items to create a single self-efficacy score.

Well-Being at School

Students' well-being at school was measured with items that were initially developed by the Finnish National Board of Education (Metsämuuronen & Svedlin, 2004), including general liking of school (e.g., ''My school days are generally nice''), academic self-concept (e.g., ''Learning brings me joy''), classroom climate (e.g., ''There is a good climate in our class''), and school climate (e.g., ''I feel safe at school''). Students responded to 14 items on a 5-point scale (0 = *I disagree completely*, 4 = *I agree completely*). All items loaded highly on one factor and thus were combined into one scale by averaging the item scores (Cronbach's α = .88).

Results

Descriptive Statistics for Outcome Variables

As a preliminary step, we examined the means and standard deviations of the imputed data for all dependent variables separately for the intervention and control groups at the three time points (Table 1). Comparing the intervention and control group means, several positive trends could be noted from the sample statistics. The biggest change took place in the mean of self-reported victimization, for which a substantial decrease occurred in the intervention group (from 0.741 to 0.485), with a much smaller change in the control group (from 0.782 to 0.657). Likewise, there was a

Table 1
Descriptive Statistics for the Criterion Variables: Means and Standard Deviations

	Intervention			Control		
Criterion	Wave 1	Wave 2	Wave 3	Wave 1	Wave 2	Wave 3
Self-reported victimization						
M	0.741	0.738	0.485	0.782	0.829	0.657
SD	1.071	1.068	0.843	1.064	1.101	0.909
Self-reported bullying						
M	0.475	0.355	0.273	0.514	0.432	0.348
SD	0.748	0.647	0.565	0.732	0.708	0.597
Peer-reported victimization						
M	0.063	0.059	0.049	0.065	0.070	0.065
SD	0.091	0.081	0.075	0.096	0.091	0.081
Peer-reported bullying						
M	0.069	0.060	0.054	0.071	0.070	0.070
SD	0.119	0.109	0.097	0.120	0.120	0.112
Peer-reported assisting						
M	0.080	0.077	0.071	0.083	0.091	0.086
SD	0.111	0.114	0.102	0.113	0.126	0.115
Peer-reported reinforcing						
M	0.107	0.116	0.107	0.105	0.127	0.120
SD	0.114	0.122	0.107	0.109	0.130	0.118
Peer-reported defending						
M	0.195	0.215	0.189	0.189	0.194	0.171
SD	0.145	0.146	0.147	0.143	0.145	0.128
Antibullying attitudes						
M	3.248	3.186	3.134	3.205	3.078	3.049
SD	0.635	0.677	0.698	0.625	0.685	0.654
Empathy toward victims						
M	2.023	2.003	1.673	1.990	1.912	1.608
SD	0.610	0.611	0.726	0.576	0.630	0.685
Self-efficacy for defending						
M	1.815	1.799	1.880	1.794	1.773	1.809
SD	0.706	0.700	0.677	0.684	0.694	0.613
Well-being at school						
M	3.026	3.004	2.871	2.978	2.902	2.748
SD	0.716	0.664	0.825	0.711	0.710	0.785

Note. Intervention n = 4,201; control n = 3,965; imputed data.

change favoring the intervention group in all the other outcomes from Wave 1 to Wave 3, albeit some of the differences were small (e.g., for empathy toward victims).

Variances and Intraclass Correlations

For each dependent variable, we estimated the variance at four levels: waves, students, classrooms, and schools (Table 2). There was statistically significant variance for all variables at each level (mostly $p < .001$). We calculated intraclass correlations (ICCs), which provide estimates of the proportion of variance due to differences between students, classrooms, and schools (for notation and formulas, see the note for Table 2). ICCs at the student level were generally higher for peer-reported than for self-reported data, suggesting that peer reports are less amenable to change than self-reports. Nevertheless, all variables show an appreciable proportion of variance associated with waves of measurement (i.e., 1 minus student-level ICC). For all variables, the classroom-level variance was higher than the school-level variance, which may indicate that classrooms are more important social contexts for bullying-related phenomena than schools. But note that ICCs at the classroom level include both classroom- and school-level variance, as classrooms are nested within schools. The highest proportions of variance associated with

Table 2
Variance Estimates and Intraclass Correlations for Dependent Variables: Wave (e), Student (u), Classroom (v), and School (f) Levels

	Variances				Intraclass correlations		
	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_v^2$	$\hat{\sigma}_f^2$	ICC_1	ICC_2	ICC_3
Self-reported victimization	0.448	0.285	0.033	0.015	.43	.06	.02
Self-reported bullying	0.466	0.242	0.024	0.012	.37	.05	.02
Peer-reported victimization	0.407	0.349	0.110	0.033	.55	.16	.04
Peer-reported bullying	0.291	0.517	0.029	0.016	.66	.05	.02
Peer-reported assisting	0.310	0.541	0.044	0.016	.66	.07	.02
Peer-reported reinforcing	0.316	0.519	0.078	0.064	.68	.15	.07
Peer-reported defending	0.309	0.441	0.224	0.037	.69	.26	.04
Antibullying attitudes	0.490	0.444	0.035	0.010	.50	.05	.01
Empathy toward victims	0.245	0.174	0.014	0.006	.44	.05	.01
Self-efficacy for defending	0.304	0.140	0.011	0.006	.34	.04	.01
Well-being at school	0.489	0.398	0.066	0.032	.50	.10	.03

Note. σ_e^2 = variance between waves of measurement; σ_u^2 = variance between students; σ_v^2 = variance between classrooms; σ_f^2 = variance between schools. All variances were statistically significant (at least $p < .01$, but mostly $p < .001$). ICC = intraclass correlation. ICC_1 = proportion of total variance at the student level and higher: $ICC_1 = (\sigma_u^2 + \sigma_v^2 + \sigma_f^2)/(\sigma_e^2 + \sigma_u^2 + \sigma_v^2 + \sigma_f^2)$. ICC_2 = proportion of total variance at the classroom and school level: $ICC_2 = (\sigma_v^2 + \sigma_f^2)/(\sigma_e^2 + \sigma_u^2 + \sigma_v^2 + \sigma_f^2)$. ICC_3 = proportion of total variance at the school level: $ICC_3 = (\sigma_f^2)/(\sigma_e^2 + \sigma_u^2 + \sigma_v^2 + \sigma_f^2)$. Wave-level N = 24,498, student-level N = 8,166, classroom-level N = 431, and school-level N = 77.

classroom or school factors were obtained for the peer reports of defending (ICC = .26), victimization (ICC = .16), and reinforcing (ICC = .15). Between-school variance was highest for peer-reported reinforcing (ICC = .07). Overall, the ICCs show that students sharing the same social environment were more alike than students from other classrooms or schools.

Outcomes

We used multilevel modeling with MLwiN 2.11 (Rasbash, Charlton, Browne, Healy, & Cameron, 2009) to estimate the intervention effects in the presence of the nested data structures. We fitted four-level models, with the first level representing change over time, the second level representing individual student differences, the third level representing differences between classrooms, and the fourth level representing between-school differences. The differences between KiVa schools and control schools were examined after controlling for baseline levels of the variable of interest, gender, age, and language of instruction at school (Finnish or Swedish). Model specification is described in Appendix B.

As previous studies have shown that gender and age are important predictors of bullying and victimization (e.g., Salmivalli & Voeten, 2004; Whitney & Smith, 1993), we included them as covariates in all models. This enabled us to control for their effects and investigate their potential interactions with the intervention. We added also the language of instruction into our models, because: (a) Swedish schools were overrepresented in the sample and (b) preliminary analyses showed that Swedish-speaking minority students reported lower levels of bullying and victimization than Finnish-speaking students.

There were several dummy-coded variables in the models. The three waves of data collection were coded with two variables, T2 (*Wave 2* = 1, *other waves* = 0) and T3 (*Wave 3* = 1, *other waves* = 0). In addition, gender (*girls* = 0, *boys* =1), language of instruction (*Finnish* = 0, *Swedish* = 1), and intervention (*control school* = 0, *intervention school* = 1) were entered into the models with dummy coding. Whereas the other covariates were left uncentered, the age variable was grand-mean centered to facilitate the interpretation of possible interaction effects (Aiken & West, 1991) and of the random parts of the models.

Before testing the intervention effects, we simplified the models to achieve parsimony and good model convergence. The models were first specified as described in Appendix B. Next, the number of parameters was reduced, if the simplification did not make the model fit worse and if it did not result in convergence problems. The simplifications included omitting random slopes for gender and/or age. If random gender or age slopes could not be omitted, simplification was sought by fixing

to zero their covariances with random slopes for T2 and T3. Mostly, we ended up having a random slope for gender and its covariances with random slopes of T2 and T3 fixed to zero. The random slope for gender means that the gender differences in the criterion variables differed by classroom. As our main purpose was to estimate intervention effects we did not further explore possible predictors for random classroom slopes of the control variables.

The intervention effects were examined as follows. We tested the statistical significance of the intervention effect at Wave 2 by deleting the Intervention × T2 interaction term from the model and conducting a deviance test (e.g., Snijders & Bosker, 1999). Next, we entered the Intervention × T2 interaction back to the model, and the significance of the intervention effect at T3 was examined in a similar way, after which the Intervention × T3 term was reentered into the model.

Next, we investigated whether the intervention effects differed depending on gender or age of the student. We entered second-order interaction terms to the equation: Boy × Intervention × T2; Age × Intervention × T2; Boy × Intervention × T3; Age × Intervention × T3. We conducted a deviance test with 4 *df*, which provided an omnibus test for statistical significance indicating whether any of the four interactions was statistically significant (Snijders & Bosker, 1999). If a statistically significant result was found, we tested the statistical significance of each of these interactions by deleting one interaction term at a time. After observing the change in the deviance, we added each interaction term back into the model to test the other interactions in a similar way.

We used 11 criterion variables: self-reported and peer-reported bullying and victimization, three bystanders' behaviors in bullying situations, antibullying attitudes, empathy toward victims, self-efficacy for defending, and well-being at school. On the basis of the distributions of the variables, skew corrections were used, except for empathy toward victims and self-efficacy for defending. Variables with skewed distributions were transformed into normal scores.

Tables 3–5 present the parameter estimates (unstandardized regression coefficients) for the final models for each criterion variable. Because our procedure for aggregating the imputed data leads to underestimated standard errors, we set the alpha level for the Wald tests at an adjusted .001. The tests of significance for the central hypotheses about intervention effects and interactions involving intervention effects were conducted separately with the log-likelihood ratio chi-square difference tests (deviance tests), which are not biased because of the imputations (Wu, Lang, & Little, 2009). In the tables, results from Wald tests ($p < .001$) are indicated in bold, whereas asterisks indicate the significance of the deviance test results.

There were residual variances at Levels 2, 3, and 4 but not at Level 1, because two dummy variables were used to represent the three time points. There were also covariances between random components in the models. All intercept–slope covariances at student, classroom, and school levels were included, but some covariances between random slopes were excluded to simplify the models, as described earlier. Due to space limitations, the tables contain only the residual variances at the student, classroom, and school level omitting all covariances.

Baseline Effects

Intervention and control schools did not differ statistically on the criterion variables (see intervention at baseline, Tables 3–5). Of the control variables, only gender had consistent effects ($p < .001$). Boys were higher than girls on peer-reported victimization ($b = 0.130$) and bullying ($b = 0.768$) as well as on self-reported bullying ($b = 0.338$) and, although just not significant, on self-reported victimization ($b = 0.096$, $p = .001$; Table 3). Boys also acted in more probullying ways as bystanders compared to girls, doing more assisting ($b = 0.921$) and reinforcing ($b = 0.982$) and less defending ($b = -0.894$; Table 4). In addition, boys had less antibullying attitudes ($b = -0.439$), less empathy toward victims ($b = -0.307$), and less self-efficacy for defending ($b = -0.099$), while also reporting a lower well-being at school than girls ($b = -0.168$; Table 5).

Intervention Effects

We examined the intervention effects at two time points during the school year. Gender and age were also used as control variables in estimating intervention effects at Wave 2 and Wave 3, even when not statistically significant at baseline. The control variables did not have any consistent pattern of effects on the change in the dependent variables.

Intervention results concerning the main outcomes are reported in Table 3. Compared with the control school students at Wave 2, students in KiVa

Table 3
Hierarchical Linear Modeling Results: Intervention Effects for Self- and Peer-Reported Victimization and Bullying

	Self-reported victimization	Self-reported bullying	Peer-reported victimization	Peer-reported bullying
	Estimate	Estimate	Estimate	Estimate
Baseline				
Intercept	0.030	−0.088	−0.002	**−0.320**
Student level				
Boy	0.096	**0.338**	**0.130**	**0.768**
Age	**−0.094**	0.018	−0.071	0.012
School level				
Swedish	−0.115	−0.130	−0.137	−0.102
Intervention	−0.280	−0.075	−0.519	0.002
Intervention × Boy	0.000	−0.052	−0.028	−0.093
Intervention × Age	0.021	0.003	0.049	0.003
Change by Wave 2				
T2	0.024	0.011	0.097	0.024
Student level				
Boy × T2	0.001	0.041	−0.042	0.045
Age × T2	0.001	0.015	0.002	0.031
School level				
Intervention × T2	−0.031	−0.040	−0.167**	−0.087
Change by Wave 3				
T3	−0.019	−0.001	−0.170	−0.239
Student level				
Boy × T3	−0.023	−0.072	**−0.121**	−0.006
Age × T3	0.009	0.006	0.033	0.027
School level				
Intervention × T3	−0.154***	−0.085*	−0.309***	−0.130
Variance components				
Student level				
Intercept	**0.681**	**0.614**	**0.677**	**0.615**
T2	**0.810**	**0.824**	**0.542**	**0.438**
T3	**0.904**	**0.958**	**0.642**	**0.552**
Classroom level				
Intercept	**0.040**	**0.041**	**0.196**	**0.076**
T2	**0.045**	**0.039**	**0.244**	**0.105**
T3	**0.043**	**0.041**	**0.330**	**0.130**
Slope for boy	**0.045**	**0.044**	**0.076**	**0.135**
Slope for age				
School level				
Intercept	0.013	0.009	0.024	0.010
T2	0.003	0.004	0.000[a]	0.000[a]
T3	0.006	0.001	0.000[a]	0.011

Note. See Appendix B for an explanation of the models. Estimates of covariances omitted. Wave-level N = 24,498, student-level N = 8,166, classroom-level N = 431, and school-level N = 77. Statistically significant results ($p < .05$) from Wald tests are in boldface, whereas statistically significant results from deviance tests are indicated with asterisks.
[a]Both estimates and their standard errors were zero up to three decimals.
*$p < .05$. **$p < .01$. ***$p < .001$.

schools had a lower level of peer-reported victimization ($b = -0.167$, $p < .008$). At Wave 3, positive intervention effects emerged for self-reported victimization ($b = -0.154$, $p < .001$) and for self-reported bullying ($b = -0.085$, $p = .012$), as well as for peer-reported victimization ($b = -0.309$, $p < .001$). Students in KiVa schools were less victimized and, according to self-reports, bullied others less than control school students. The intervention seemed to decrease also peer-reported bullying, but this effect did not reach statistical significance ($b = -0.130$, $p = .095$).

Table 4
Hierarchical Linear Modeling Results: Intervention Effects for Peer-Reported Bystander Behaviors

	Peer-reported assisting	Peer-reported reinforcing	Peer-reported defending
	Estimate	Estimate	Estimate
Baseline			
Intercept	**−0.432**	**−0.405**	**0.407**
Student level			
Boy	**0.921**	**0.982**	**−0.894**
Age	0.031	0.065	−0.003
School level			
Swedish	0.006	**−0.516**	**0.321**
Intervention	−0.043	0.073	0.087
Intervention × Boy	−0.126	−0.049	0.049
Intervention × Age	0.010	−0.005	−0.002
Change by Wave 2			
T2	0.029	0.004	−0.037
Student level			
Boy × T2	0.051	**0.073**	−0.005
Age × T2	0.000	0.000	**−0.071**
School level			
Intervention × T2	−0.114	−0.116	0.110*
Change by Wave 3			
T3	0.057	−0.028	0.169
Student level			
Boy × T3	0.032	0.040	0.043
Age × T3	0.000	0.008	−0.019
School level			
Intervention × T3	−0.131*	−0.168*	0.080
Variance components			
Student level			
Intercept	**0.572**	**0.464**	**0.444**
T2	**0.466**	**0.378**	**0.366**
T3	**0.565**	**0.459**	**0.420**
Classroom level			
Intercept	**0.138**	**0.197**	**0.365**
T2	**0.149**	**0.266**	**0.271**
T3	**0.178**	**0.266**	**0.346**
Slope for boy	**0.227**	**0.263**	**0.251**
Slope for age		0.013	0.020
School level			
Intercept	0.014	0.040	0.019
T2	0.000[a]	0.017	0.002
T3	0.000[a]	0.033	0.020

Note. See Appendix B for an explanation of the models. Estimates of covariances omitted. Wave-level N = 24,498, student-level N = 8,166, classroom-level N = 431, and school-level N = 77. Statistically significant results ($p < .05$) from Wald tests are in boldface, whereas statistically significant results from deviance tests are indicated with asterisks.
[a]Both estimates and their standard errors were zero up to three decimals.
*$p < .05$.

The intervention had some positive effects on the bystanders' behaviors as well (Table 4). At Wave 2, the KiVa school students defended victims more (b = 0.110, p = .046), compared to the control school students. By Wave 3, however, the intervention effect had diminished (b = 0.080, p = .251) rendering the result nonsignificant. Positive effects emerged at Wave 3 for assisting the bully (b = −0.131, p = .011) and reinforcing the bully (b = −0.168, p = .019). This means that after 9 months of intervention, KiVa school students assisted and reinforced the bully less than the control school students.

Results concerning antibullying attitudes, empathy toward victim, self-efficacy for defending and well-being at school are presented in Table 5. Compared to the control school students at Wave 2, students in KiVa schools had more antibullying attitudes (b = 0.088, p = .021) and empathy (b = 0.059, p = .002). However, by Wave 3, these intervention effects had diminished, making the results statistically nonsignificant (b = 0.056, p = .139 and b = 0.039, p = .065 for attitudes and empathy, respectively). At the posttest assessment, KiVa school students reported having more self-efficacy for defending (b = 0.052, p = .026) and well-being at school (b = 0.096, p = .011), compared to the control-school students.

In general, the intervention had equal effects on boys and girls and students of different ages with only one exception. More specifically, we found Age × Intervention × T2 and Age × Intervention × T3 interactions on peer-reported bullying. By probing these interactions, we found that the intervention effects were larger for older students at both Waves 2 and 3. Because deviance tests were not possible here, we did not conduct tests of significance at different ages. The effect sizes for mean ages at Grades 4–6 are reported in Table 6 and elaborated next.

Table 6 shows the intervention effect sizes in the metric of Cohen's d at the two time points. All effects are in favor of the KiVa schools. The intervention was effective in reducing victimization according to both self- and peer reports, but the effect size was almost twice as large for peer reports (0.33) compared to self-reports (0.17). Compared to victimization, the intervention effects on bullying were smaller for both self-reports (0.10) and peer reports (0.03–0.18), with larger effects on peer reports for older students. The intervention decreased assisting the bully (0.14) and reinforcing

Table 5
Hierarchical Linear Modeling Results: Intervention Effects for Antibullying Attitudes, Empathy Toward Victims, Self-Efficacy for Defending, and Well-Being at School

	Antibullying attitudes	Empathy toward victims	Self-efficacy for defending	Well-being at school
	Estimate	Estimate	Estimate	Estimate
Baseline				
Intercept	**0.171**	**2.126**	**1.894**	−0.016
Student level				
Boy	**−0.439**	**−0.307**	**−0.099**	**−0.168**
Age	−0.040	**−0.049**	0.041	−0.062
School level				
Swedish	−0.029	0.065	−0.080	**0.243**
Intervention	0.160	−0.026	−0.060	0.026
Intervention × Boy	−0.010	0.031	0.033	0.013
Intervention × Age	−0.007	0.004	0.005	0.008
Change by Wave 2				
T2	0.000	**−0.072**	0.006	−0.033
Student level				
Boy × T2	**−0.076**	−0.011	−0.055	0.014
Age × T2	**−0.058**	**−0.034**	−0.035	−0.007
School level				
Intervention × T2	0.088*	0.059**	0.009	0.054
Change by Wave 3				
T3	**0.745**	0.337	**0.461**	−0.123
Student level				
Boy × T3	−0.074	**−0.083**	−0.002	−0.076
Age × T3	**−0.065**	**−0.060**	**−0.039**	0.010
School level				
Intervention × T3	0.056	0.039	0.052*	0.096*
Variance components				
Student level				
Intercept	**0.865**	**0.306**	**0.459**	**0.852**
T2	**0.913**	**0.334**	**0.613**	**0.866**
T3	**1.045**	**0.455**	**0.606**	**1.018**
Classroom level				
Intercept	**0.034**	**0.017**	**0.015**	**0.092**
T2	**0.032**	**0.015**	**0.018**	**0.057**
T3	**0.048**	**0.021**	**0.015**	**0.071**
Slope for boy	0.026	**0.015**		**0.052**
Slope for age				
School level				
Intercept	0.005	0.005	0.006	0.029
T2	0.009	0.000	0.000[a]	0.008
T3	0.007	0.000[b]	0.001	0.001

Note. See Appendix B for an explanation of the models. Estimates of covariances omitted. Wave-level $N = 24{,}498$, student-level $N = 8{,}166$, classroom-level $N = 431$, and school-level $N = 77$. Statistically significant results ($p < .05$) from Wald tests are in boldface, whereas statistically significant results from deviance tests are indicated with asterisks.
[a]Both estimates and their standard errors were zero up to three decimals.
[b]Due to convergence problems, the parameter was fixed to zero.
*$p < .05$. **$p < .01$.

the bully (0.17), and there were positive effects on other dependent variables as well (0.06–0.10). For defending the victim, antibullying attitudes and empathy toward victims, the effects actually decreased slightly from Wave 2 to Wave 3.

To make the results comparable with previous studies (e.g., Baldry & Farrington, 2007), we investigated how much the KiVa program reduced the prevalence of bullying and victimization and the odds for these problems. To this end, we

Table 6
Effect Sizes (Cohen's ds*) for the Intervention Effects Compared to the Control Schools*

Dependent variable	Wave 2	Wave 3
Self-reported victimization	0.03	0.17
Self-reported bullying	0.05	0.10
Peer-reported victimization	0.18	0.33
Peer-reported bullying: Overall	0.10	0.14
Peer-reported bullying: Grade 4	0.03	0.03
Peer-reported bullying: Grade 5	0.10	0.10
Peer-reported bullying: Grade 6	0.18	0.18
Peer-reported assisting	0.12	0.14
Peer-reported reinforcing	0.12	0.17
Peer-reported defending	0.11	0.08
Antibullying attitudes	0.09	0.06
Empathy toward victims	0.10	0.06
Self-efficacy of defending	0.01	0.08
Well-being at school	0.05	0.10

Note. All effects are in favor of the intervention. Cohen's d was calculated as the adjusted group mean difference divided by unadjusted pooled within-group standard deviation:

$$d = \frac{\gamma}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2}}}$$

where γ is the coefficient for the intervention's effect, which represents the group mean difference adjusted for student- and school-level covariates (Intervention × T2 or Intervention × T3); n_1 and n_2 are the student-level sample sizes; and S_1 and S_2 are the student-level unadjusted posttest standard deviations for the intervention group and the comparison group, respectively.

categorized students into victims, bullies, and non-involved children by dichotomizing the self-reported bullying and victimization. The cut-point of two or three times a month (score = 2) was used to indicate victimization and bullying, respectively (for a conceptual and empirical justification, see Solberg & Olweus, 2003). The resulting prevalence rates for victimization at the three waves in the control schools were 16.8%, 19.1%, and 12.7%, which can be compared to the KiVa schools with prevalence rates of 16.6%, 16.4%, and 8.9% (for Waves 1–3, respectively). The corresponding figures for bullying were for control schools 7.9%, 6.9%, and 3.8%, whereas for KiVa schools they were 8.0%, 4.6%, and 3.1%. In KiVa schools, there was from pretest to posttest a reduction of 46% in victimization and a reduction of 61% in bullying others. Compared to the control schools at Wave 3, reductions in KiVa schools amounted to approximately 30% for self-reported victimization and 17% for self-reported bullying others.

We also calculated odds ratios for victimization and bullying, while correcting the standard errors for clustering at the school level (for the formulas, see Farrington & Ttofi, 2009). At Wave 3, the odds ratio for victimization was 1.47, 95% CI [1.10, 1.96], and for bullying the odds ratio was 1.22, 95% CI [0.78, 1.90]. When the effect sizes for peer reports are converted into odds ratios (Lipsey & Wilson, 2001), we get odds ratios of 1.83 for victimization and 1.29 for bullying. Taken together, the odds of being a victim were about 1.5–1.8 times higher for a control school student than for a student in an intervention school, and the odds of being a bully were 1.2–1.3 times higher for a control school student than for a student in an intervention school.

Discussion

The present study is the first evaluation of the effectiveness of the KiVa antibullying program for Grades 4–6. On the whole, the results give clear support to the effectiveness of the KiVa program. These findings are very relevant both for schools' antibullying policies and for antibullying research. First, they provide stakeholders guidance for adopting an effective way to reduce bullying and victimization, and second, they give empirical support for the view that school-based antibullying interventions can make a difference, despite some modest and contradictory results of previous trials (for meta-analyses, see Baldry & Farrington, 2007; Farrington & Ttofi, 2009; Ferguson et al., 2007; Merrell et al., 2008; J. D. Smith et al., 2004). Of course, much more needs to be done to answer many critical questions about the effectiveness of antibullying interventions; the present study only shows the core results of one specific program. Further studies are needed to understand the mechanisms and prerequisites of change.

In our models, covariates included gender, age, and language used at school. We controlled for their effects on the baseline levels of dependent variables, and in addition, we accounted for the effects of gender and age on change over time. This statistical controlling gives some additional credibility for our results, which are based on a randomized design. Gender had very consistent and relatively large effects at baseline, such that boys were in a disadvantaged position with regard to every dependent variable: They bullied others more, they were victimized more (although this effect reached significance only in peer reports), they assisted and reinforced bullying more, and so on. These gender findings are in line with several previous studies (e.g., Salmivalli & Voeten, 2004; Veenstra et al., 2005) showing the importance of gender in bullying-related phenomena. For other

control variables, the effects on the dependent variables at baseline and on change were inconsistent and mostly nonsignificant.

Turning to the intervention results, we found some effects already at Wave 2. Compared to students in control schools, students in intervention schools defended the victims more and they had more antibullying attitudes and empathy toward victims. According to peer reports, students in intervention schools were also less victimized, but self-reports of victimization did not confirm this finding. A positive intervention effect emerged on 4 of the 11 dependent variables at Wave 2.

By Wave 3 intervention and control-school students no longer differed significantly in defending, attitudes, or empathy. Mean differences, though, were still in the expected direction. Also as hypothesized, intervention school students were less victimized, they assisted and reinforced the bully less, and they had higher self-efficacy for defending and well-being at school. According to self-reports, students in intervention schools also bullied others less, but peer reports about bullying confirmed this finding only for older students in our sample. Therefore, after 1 year of intervention, the KiVa program reduced victimization and bullying, but the results for bullying were clear and consistent only for students in Grades 5 and 6. Results from the multilevel models at Wave 3, showed positive intervention effects on 7 of the 11 dependent variables.

Previous research has suggested the possibility that intervention could cause more reporting of bullying and victimization by increasing awareness of bullying, without an actual increase in bullying or victimization. This phenomenon has been called ''sensitization effect'' (Frey et al., 2005; P. K. Smith et al., 2003; Stevens et al., 2000). It might provide a partial explanation for the statistical nonsignificance of the findings for bullying and victimization at Wave 2. Our Wave 2 data were collected some months after the implementation started. During these months, the contents of the lessons included discussions about what bullying is, how frequent a problem it is, what kind of negative consequences it can have, and why it should not be tolerated. Therefore, learning about these issues and becoming sensitized for them could mean that intervention effects are underestimated at Wave 2.

To examine the magnitude of the effects, we computed standardized effect sizes (Cohen's *d*s) from the multilevel modeling results. From these results we can see that the program reached at Wave 3 a victimization-reducing effect, which can be considered practically significant: 0.33 for peer reports and 0.17 for self-reports (e.g., Merrell et al., 2008, set the limit at $d \geq 0.20$). We interpret this as an important reduction, considering the persistence and seriousness of the problem. For other criterion variables effect sizes were mainly smaller (0.06–0.14), except for reinforcing the bully (0.17). Nevertheless, all of them consistently favored the intervention.

The KiVa antibullying program was successful in reducing the prevalence of bullying and victimization. At Wave 3, there was a reduction of 30% in self-reported victimization and a reduction of 17% in self-reported bullying, compared with control schools. These results are consistent with previous studies (e.g., Salmivalli et al., 2005; Whitney, Rivers, Smith, & Sharp, 1994), and close to results from a recent meta-analysis, which showed that antibullying programs have reduced bullying and victimization by about 17%–23% in experimental schools compared to control schools (Farrington & Ttofi, 2009).

Contrary to most previous evaluation studies, we utilized both self- and peer reports of bullying and victimization as outcome measures. A common view is that peer reports are especially resistant to change, partly due to the high stability of reputations that do not always accurately reflect actual current behaviors or experiences (Juvonen, Nishina, & Graham, 2001; Olweus, 2009). Our findings partly support this view, in that within-student variance was smaller and proportions of student-level variance were larger for peer reports than self-reports. Nevertheless, the strongest effects were obtained for peer-reported victimization. Clearly, the influence of an effective intervention can be seen in peer reports as well. Considering how peer reports were collected (nominations of classmates who behave in certain ways in situations of bullying), these beneficial intervention effects mean that students in KiVa schools have changed their actual behaviors to an extent that can be observed by classmates. In this way, the KiVa program overcomes one key limitation of several previous programs, namely, their inability to influence behavior rather than just beliefs or intentions (Merrell et al., 2008).

Shadish, Cook, and Campbell (2002) have identified the main threats to the internal validity of an evaluation study. Most of the eight threats discussed by them can be readily excluded in the present study due to random assignment of schools to intervention and control conditions. The most important remaining threat is differential attrition,

or differential loss of units from the intervention condition compared to the control condition. By using modern missing data estimation (unlike most previous bullying intervention studies), we were able to mitigate the impact of selective attrition to the degree that the process is related to variables on our data set (Enders, 2010). It is worth pointing out, and important for comparability between evaluations of antibullying programs that with full information maximum likelihood (FIML) analysis applied to the nonimputed data consistently significant effects were found on *all* criterion variables in favor of the experimental condition. In addition, most effects were stronger than the ones reported here on the basis of the imputed data. For instance, the model-based odds ratios for self-reported victimization and bullying, controlling for a large set of covariates, were in the neighborhood of two or more. In experimental schools, compared with control schools, self-reported victimization and bullying were reduced by 40% and 33%, respectively. For victimization, the results from the two ways of analyzing the data were quite comparable, but for self-reported bullying the effect seemed clearly lower when computed from the imputed data. Missing values and selective attrition are likely to be present in most evaluation studies. It is important to realize that the way studies deal with it or do not deal with it can have an influence on the effect sizes obtained. Even two methods that are both considered valid (i.e., FIML and multiple imputation, see Jeličić, Phelps, & Lerner, 2009) apparently may give somewhat different results. Overall, given the experimental design and the best practice analyses in our study, the outcomes showed a consistent pattern favoring the intervention schools, which makes any alternative explanations speculative at best.

Our sample was nationally representative in the sense that all provinces and both Finnish- and Swedish-language schools in the mainland Finland were represented. All schools involved in the evaluation either as intervention or control school volunteered to do so. Our findings are therefore generalizable only to schools willing to implement an antibullying program. This limit in generalization is reasonable because schools with a similar motivation are likely to be future implementers of the program. In other words, the program effects can be generalized to schools that are willing to implement the program (at the time of this writing, almost two thirds of all Finnish comprehensive schools have registered and either have started or will start implementing KiVa in the near future).

The findings reported here represent the core results of the intervention. As the effects were not identical across schools and classrooms, we hope to gain insight into the association of implementation and outcome and into the relative effectiveness of the different components (see, e.g., Olweus & Alsaker, 1991; Salmivalli et al., 2005; Whitney et al., 1994). Another focus of future interest will be the mechanisms of change. With longitudinal data we can examine, for example, what kinds of changes in bystander behaviors mediate the program effects on bullying and victimization, and how changes in victimization (or witnessing victimization) are related to the well-being of students.

Another future task will be to evaluate the long-term effects of KiVa. Such effects have often been discouraging (e.g., Eslea & Smith, 1998). Nevertheless, the whole idea behind KiVa is that rather than being a project that lasts for a year and then ends, it will be part of the schools' continuous antibullying work. Schools that have adopted KiVa will continue implementing it (for strategies to encourage program maintenance, see Salmivalli et al., 2010b). We will monitor the effects of KiVa in the forthcoming years not only in new KiVa schools but in the first intervention schools as well, to see whether the effects will remain or even strengthen over time.

Raising Healthy Children: Implications for Policy and Practice

School bullying is a serious problem because it poses a risk for students' current (Hawker & Boulton, 2003) as well as future psychosocial well-being (e.g., Isaacs et al., 2008). For instance, according to a recent longitudinal study on Finnish boys, bullying and victimization during early school years are risk factors for psychiatric disorders in early adulthood: Victims of bullying are at risk for anxiety disorder, and bullies are at risk for antisocial personality disorder (Sourander et al., 2007). Bullying can even contribute to school shootings; there is some evidence that most school shooters have experienced prolonged marginalization and victimization by their peers (Leary, Kowalski, Smith, & Phillips, 2003). Even apart from the most serious and rare tragedies, it is clear that bullying and victimization threaten the healthy development of children around the world. Therefore, it is important not only to reduce bullying once it has taken hold, but also to prevent it in the first place.

Fortunately, intervention programs have been developed which provide teachers with means to

prevent and reduce bullying (e.g., Olweus, 1991; Salmivalli et al., 2010). Our results concerning the effectiveness of the KiVa program suggest that theoretically well-grounded interventions, which include both universal and indicated actions, can reduce bullying problems in schools. Universal actions, targeting not only individual children but also the classroom and school levels are important, because influencing the bystanders and the classroom as a whole seems to be an essential part of an effective strategy, especially for prevention. Indicated actions, in turn, are needed to intervene in ongoing bullying.

Despite our first results on the effects of the KiVa program, detailed analyses of the effectiveness of different program components still remain an important topic for future research. In addition, it should be investigated, whether the good results can be achieved in other countries. Some factors that may have contributed to the results include the facts that Finnish schools are quite homogeneous with respect to bullying, teachers have a good training, and they even have a legal obligation to tackle bullying. These things can vary across countries, which may make reducing bullying more difficult in other contexts. On the other hand, the KiVa program is very concrete and easy to adopt, and it has detailed manuals, which makes the implementation of the intervention possible also elsewhere. A replication of the present study would actually provide valuable evidence of the effectiveness of the program under different conditions.

According to our experience from the present trial, both political decisions and commitment from part of the participating schools are needed to enable a successful implementation of the KiVa program. The school principals are in a key role in motivating and enabling high-quality implementation of any school-based intervention program, which is emphasized throughout the process of introducing KiVa to schools and training the school personnel. It is essential that school authorities see the school as an arena for positive psychosocial development in addition to fostering academic achievement. We believe that eventually, a program such as KiVa may have an influence on students' learning outcomes as well: Even the present findings showed positive effects on school well-being, such as general liking of school and academic self-concept.

Sufficient support materials and teacher training should be provided to facilitate program implementation because professionally prepared teacher manuals and other tools ease the teachers' workload and thereby facilitate accurate program implementation. It is important also to organize hands-on training for teachers in using the systematic discussion techniques for addressing bullying cases. Simulation exercises during the training provide concrete learning experiences, which creates good chances for adopting the necessary skills. Furthermore, during the training it is also beneficial to foster teachers' motivation and their sense of ownership of the program. Teachers should view preventing and reducing bullying as one of their basic tasks, not some additional work imposed on them by the education authorities.

In the present study, we implemented the principles described earlier, with rather robust and consistent positive results. We therefore suggest that a program like KiVa can reduce bullying and victimization in the middle childhood years in schools similar to those in this study. Thus, whole-school antibullying programs have significant potential for making an important contribution to educators' common goal, raising healthy children.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Baldry, A. C., & Farrington, D. P. (2004). Evaluation of an intervention program for the reduction of bullying and victimization in schools. *Aggressive Behavior*, *30*, 1–15.

Baldry, A. C., & Farrington, D. P. (2007). Effectiveness of programs to prevent school bullying. *Victims & Offenders*, *2*, 183–204.

Bandura, A. (1989). Social cognitive theory. In R. Vasta (Ed.), *Annals of child development: Six theories of child development* (pp. 1–60). Greenwich, CT: JAI Press.

Caravita, S., DiBlasio, P., & Salmivalli, C. (2009). Unique and interactive effects of empathy and social status on involvement in bullying. *Social Development*, *18*, 140–163.

Card, N. A. (2003, April). *Victims of peer aggression: A meta-analytic review*. Presented at the Society for Research in Child Development biennial meeting, Tampa, FL.

Cillessen, A. H. N., & Mayeux, L. (2004). From censure to reinforcement: Developmental changes in the association between aggression and social status. *Child Development*, *75*, 147–163.

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). New York: Cambridge University Press.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Eslea, M., & Smith, P. (1998). The long-term effectiveness of anti-bullying work in primary schools. *Educational Research*, *40*, 203–218.

Farrington, D. P., & Ttofi, M. M. (2009). School-based programs to reduce bullying and victimization. *Campbell Systematic Reviews*, 2009: 6. Retrieved from http://www.crim.cam.ac.uk/people/mt394/c09.pdf

Ferguson, C., San Miguel, C., Kilburn, J., & Sanchez, P. (2007). The effectiveness of school-based anti-bullying programs: A meta-analytic review. *Criminal Justice Review*, *32*, 401–414.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., et al. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151–175.

Frey, K. S., Hirschstein, M. K., Snell, J. L., Edstrom, L. V. S., MacKenzie, E. P., & Broderick, C. J. (2005). Reducing playground bullying and supporting beliefs: An experimental trial of the steps to respect program. *Developmental Psychology*, *41*, 479–491.

Hawker, D. S. J., & Boulton, M. J. (2003). Twenty years' research on peer victimization and psychosocial maladjustment: A meta-analytic review of cross-sectional studies. In M. E. Hertzig & E. A. Farber (Eds.), *Annual progress in child psychiatry and child development: 2000–2001* (pp. 505–534). New York: Brunner-Routledge.

Isaacs, J., Hodges, E. V. E., & Salmivalli, C. (2008). Long-term influences of victimization: A follow-up from adolescence to young adulthood. *European Journal of Developmental Science*, *11*, 387–397.

Jeličić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*, 1195–1199.

Juvonen, J., & Galván, A. (2008). Peer influence in involuntary social groups: Lessons from research on bullying. In M. J. Prinstein & K. A. Dodge (Eds.), *Understanding peer influence in children and adolescents* (pp. 225–244). New York: Guilford.

Juvonen, J., Graham, S., & Schuster, M. A. (2003). Bullying among young adolescents: The strong, the weak, and the troubled. *Pediatrics*, *112*, 1231–1237.

Juvonen, J., Nishina, A., & Graham, S. (2001). Self views versus peer perceptions of victims status among early adolescents. In J. Juvonen & S. Graham (Eds.), *Peer harassment in school: The plight of the vulnerable and victimized* (pp. 105–124). New York: Guilford.

Kaltiala-Heino, R., Rimpelä, M., Rantanen, P., & Rimpelä, A. (2000). Bullying at school—An indicator of adolescents at risk for mental disorders. *Journal of Adolescence*, *23*, 661–674.

Kärnä, A., Voeten, M., Poskiparta, E., & Salmivalli, C. (2010). Vulnerable children in varying classroom contexts: Bystanders' behaviors moderate the effects of risk factors on victimization. *Merrill-Palmer Quarterly*, *56*, 261–282.

Leary, M. R., Kowalski, R. M., Smith, L., & Phillips, S. (2003). Teasing, rejection and violence: Case studies of the school shootings. *Aggressive Behavior*, *29*, 202–214.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Loeber, R., & Dishion, T. (1983). Early predictors of male delinquency: A review. *Psychological Bulletin*, *94*, 68–99.

Merrell, K., Gueldner, B., Ross, S., & Isava, D. (2008). How effective are school bullying intervention programs? A meta-analysis of intervention research. *School Psychology Quarterly*, *23*, 26–42.

Metsämuuronen, J., & Svedlin, R. (2004). *Kouluviihtyvyyden muuttuminen peruskoulussa ja lukiossa iän funktiona* [Change in perceived well-being in primary and secondary schools as a function of age]. Manuscript submitted for publication.

Nansel, T. R., Craig, W., Overpeck, M. D., Saluja, G., Ruan, W. J., & the Health Behavior in School-Aged Children Bullying Analyses Working Group. (2004). Cross-national consistency in the relationship between bullying behaviors and psychosocial adjustment. *Archives of Pediatrics & Adolescent Medicine*, *158*, 730–736.

Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Morton, B., & Scheidt, P. (2001). Bullying behaviors among US youth: Prevalence and association with psychosocial adjustment. *Journal of the American Medical Association*, *285*, 2094–2100.

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

Olweus, D. (1993a). Bullies on the playground: The role of victimization. In C. H. Hart (Ed.), *Children on playgrounds: Research perspectives and applications* (pp. 85–128). Albany: State University of New York Press.

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

Olweus, D. (1994). Bullying at school: Long-term outcomes for the victims and an effective school-based intervention program. In L. R. Huesmann (Ed.), *Aggressive behavior: Current perspectives* (pp. 97–130). New York: Plenum.

Olweus, D. (1996). *The Revised Olweus Bully/Victim Questionnaire*. Research Center for Health Promotion (HEMIL Center). Bergen, Norway: University of Bergen.

Olweus, D. (1999). Sweden. In P. K. Smith, Y. Morita, J. Junger-Tas, D. Olweus, R. Catalano, & P. Slee (Eds.), *The nature of school bullying: A cross-national perspective* (pp. 7–27). London: Routledge.

Olweus, D. (2009). Understanding and researching bullying: Some critical issues. In S. Jimerson, S. Swearer, & D. Espelage (Eds.), *The international handbook of school bullying*. (pp. 9–33) New York: Routledge.

Olweus, D., & Alsaker, F. (1991). Assessing change in a cohort-longitudinal study with hierarchical data. In D. Magnusson, L. Bergman, G. Rudinger, & B. Törestad (Eds.), *Problems and methods in longitudinal research: Stability and change* (pp. 107–132). New York: Cambridge University Press.

O'Moore, A. M., & Minton, S. J. (2004). Ireland: The Donegal primary schools' anti-bullying project. In P. K. Smith, D. Pepler, & K. Rigby (Eds.), *Bullying in schools: How successful can interventions be?* (pp. 275–287). New York: Cambridge University Press.

Pepler, D. J., Craig, W. M., Ziegler, S., & Charach, A. (1994). An evaluation of an anti-bullying intervention in Toronto schools. *Canadian Journal of Community Mental Health*, *13*, 95–110.

Pitts, J., & Smith, P. (1995). *Preventing school bullying*. London: Home Office.

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.

Pöyhönen, V., Kärnä, A., & Salmivalli, C. (2008, July). Classroom-level moderators of the empathy-defending link. In René Veenstra (Chair), *Bullying and victimization*, Symposium conducted at the biennial meeting of the International Society for the Study of Behavioural Development, Würzburg, Germany.

Pöyhönen, V., & Salmivalli, C. (2008). New directions in research and practice addressing bullying: Focus on defending behavior. In D. Pepler & W. Craig (Eds.), *An international perspective on understanding and addressing bullying* (PREVNet Publication Series, 1, pp. 26–43). Bloomington, IN: AuthorHouse.

Rasbash, J., Charlton, C., Browne, W.J., Healy, M., & Cameron, B. (2009). *MLwiN Version 2.11*. Centre for Multilevel Modelling, University of Bristol.

Rigby, K., & Slee, P. T. (1991). Bullying among Australian school children: Reported behavior and attitudes toward victims. *Journal of Social Psychology*, *131*, 615–627.

Rodkin, P. C., Farmer, T. W., Pearl, R., & Van Acker, R. (2000). Heterogeneity of popular boys: Antisocial and prosocial configurations. *Developmental Psychology*, *36*, 14–24.

Roland, E. (1989). Bullying: The Scandinavian tradition. In D. P. Tattum & D. A. Lane (Eds.), *Bullying in schools* (pp. 21–32). Stoke-on-Trent, UK: Trentham Books.

Salmivalli, C. (2010). Bullying and the peer group: A review. *Aggression and Violent Behavior*, *15*, 112–120.

Salmivalli, C., Kärnä, A., & Poskiparta, E. (2010a). From peer putdowns to peer support: A theoretical model and how it translated into a national anti-bullying program. In S. R. Jimerson, S. M. Swearer, & D. L. Espelage (Eds.), *Handbook of bullying in schools: An international perspective* (pp. 441–454). New York: Routledge.

Salmivalli, C., Kärnä, A., & Poskiparta, E. (2010b). Development, evaluation, and diffusion of a national anti-bullying program, KiVa. In B. Doll, W. Pfohl, & J. Yoon (Eds.), *Handbook of youth prevention science* (pp. 238–252). New York: Routledge.

Salmivalli, C., Kaukiainen, A., & Voeten, M. (2005). Anti-bullying intervention: Implementation and outcome. *British Journal of Educational Psychology*, *75*, 465–487.

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.

Salmivalli, C., & Peets, K. (2008). Bullies, victims, and bully-victim relationships. In K. Rubin, W. Bukowski, & B. Laursen (Eds.), *Handbook of peer interactions, relationships, and groups* (pp. 322–340). New York: Guilford.

Salmivalli, C., & Voeten, M. (2004). Connections between attitudes, group norms, and behaviors associated with bullying in schools. *International Journal of Behavioral Development*, *28*, 246–258.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.

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., Ananiadou, K., & Cowie, H. (2003). Interventions to reduce school bullying. *Canadian Journal of Psychiatry*, *48*, 591–599.

Smith, P. K., Pepler, D., & Rigby, K. (Eds.). (2004). *Bullying in schools: How successful can interventions be?* New York: Cambridge University Press.

Smith, P. K., & Sharp, S. (Eds.). (1994). *School bullying: Insights and perspectives*. New York: Routledge.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

Solberg, M. E., & Olweus, D. (2003). Prevalence estimation of school bullying with the Olweus Bully/Victim questionnaire. *Aggressive Behavior*, *29*, 239–268.

Sourander, A., Jensen, P., Rönning, J. A., Niemelä, S., Helenius, H., Sillanmäki, L., et al. (2007). What is the early adulthood outcome of boys who bully or are bullied in childhood? The Finnish ''From a Boy to a Man'' study. *Pediatrics*, *120*, 397–404.

Stevens, V., De Bourdeaudhuij, I., & Van Oost, P. (2000). Bullying in Flemish schools: An evaluation of anti-bullying intervention in primary and secondary schools. *British Journal of Educational Psychology*, *70*, 195–210.

Veenstra, R., Lindenberg, S., Oldehinkel, A. J., De Winter, A. F., Verhulst, F. C., & Ormel, J. (2005). Bullying and victimization in elementary schools: A comparison of bullies, victims, bully/victims, and uninvolved preadolescents. *Developmental Psychology*, *41*, 672–682.

Vreeman, R. C., & Carroll, A. E. (2007). A systematic review of school-based interventions to prevent bullying. *Archives of Pediatrics & Adolescent Medicine*, *161*, 78–88.

Whitney, I., Rivers, I., Smith, P., & Sharp, S. (1994). The Sheffield project: methodology and findings. In P. Smith & S. Sharp (Eds.), *School bullying: Insights and perspectives* (pp. 20–56). London: Routledge.

Whitney, I., & Smith, P. K. (1993). A survey of the nature and extent of bullying in junior/middle and secondary schools. *Educational Research*, *35*, 3–25.

Wu, W., Lang, K. M., & Little, T. D. (2009, October). *Does it fit? a simple method of assessing model fit and testing*

significance in multiply imputed data. Paper presented at the annual meeting of the Society for Multivariate Experimental Psychology. Shalishan, Oregon.

Appendix A: Missing Data Imputation Details

The task of imputing the missing data involved a number of steps. Our goal was to capture the MAR missing data mechanism by using information available from all data sources on the data set. To do so feasibly and parsimoniously, we created aggregate factor scores to represent the information from all variables on the data set. Three factors at each wave of measurement were created after an exploratory factor analysis of the variables. This step was done as a simple data reduction procedure. We then created dummy codes to represent the different nested patterns of students in schools including the cross-classification patterns of students changing classroom between the first and second time point, if more than five students followed a similar pattern. For the intervention group, we created 249 dummy-coded group-classification variables and for the control group, we created 224 of them. We then computed 4,257 interaction terms of these variables with the nine factor scores. To reduce the number of variables and yet still capture the inherent information of the interaction terms, we did a principal component analysis of these interaction variables and output the first 120 component scores, separately for the control versus intervention conditions. The nine factor scores, the basic demographic variables, the group classification variables, and the 120 component scores for the interaction information were then imputed to create a complete-case ''block'' of information about the overall data set. This block was then used to inform the imputation of the variables that we used in the analyses. We imputed the missing data 100 times, separately for control versus intervention conditions. We then calculated the average imputed value for each missing data point, which represents the best population estimate of the value needed to reproduce the population parameters. Because aggregating data in this way leads to underestimated standard errors, we conducted significance tests for the intervention effects by using the log-likelihood ratio deviance test, which is not biased in this manner (Wu , Lang, & Little, 2009). Because scale means rather than scores for single items were imputed, all reliability estimates are based on the nonimputed data.

Appendix B: Multilevel Model to Estimate Intervention Effects

The multilevel models were specified as described in the following equations at four levels of data.

Level 1: Change Over Time

$$Y_{tijk} = \beta_{0ijk} + \beta_{1ijk}\mathrm{T2} + \beta_{2ijk}\mathrm{T3},$$

where t is used to indicate time points, i is used for individual students, j is used to denote classrooms, and k to denote schools.

In this model specification, β_{0ijk} represents the intercept or baseline (May 2007). T2 and T3 are dummy variables representing Wave 2 (December 2007 or January 2008) and Wave 3 (May 2008) of data collection. The regression coefficients for T2 and T3 represent average change compared with baseline. The Level 1 model includes two dummy variables for time because we wanted to estimate the intervention effects separately at the two time points. As the two dummies fully represent the three time points, there is no residual variance left at Level 1.

The intercept and the slopes for the time variables were allowed to vary randomly at Levels 2, 3 and 4 without any restrictions on their variances and covariances. This specification implies that the baseline and change in the criterion variables may differ between students, classes, and schools.

Level 2: Individual Student Differences

$$\beta_{0ijk} = \pi_{00jk} + \pi_{01jk}\,\mathrm{Boy} + \pi_{02jk}\,\mathrm{Age} + u_{0ijk},$$
$$\beta_{1ijk} = \pi_{10jk} + \pi_{11jk}\,\mathrm{Boy} + \pi_{12jk}\,\mathrm{Age} + u_{1ijk},$$
$$\beta_{2ijk} = \pi_{20jk} + \pi_{21jk}\,\mathrm{Boy} + \pi_{22jk}\,\mathrm{Age} + u_{2ijk}.$$

For each regression coefficient of the Level 1 model we had an equation at the student level. The slopes for the control variables, Boy and Age, were specified to be random at the classroom level but not at the school level, because it seemed not plausible that the gender or age difference would differ by school. Boy and Age in the equation for

the student-level intercepts (β_{0ijk}) represent the effects of gender and age at baseline. Boy is a dummy variable and Age is defined as (grand-mean centered) age in years at the start of the intervention. Boy and Age were also specified as predictors for the changes across time. These effects are shown in Tables 3–5 as Boy × T2, Age × T2, Boy × T3, and Age × T3 at the student level for Change by Wave 2 and Change by Wave 3, respectively. The product terms result from substituting the higher level equations into the lower level equation.

The random effects at Level 2 (u_{0ijk}, u_{1ijk}, and u_{2ijk}) are assumed to be multivariate normally distributed with zero mean and a constant (3 × 3) variance–covariance matrix. All intercept–slope and slope–slope covariances were allowed to vary freely.

Level 3: Classroom Differences

$$\pi_{00jk} = \gamma_{000k} + v_{00jk},$$
$$\pi_{01jk} = \gamma_{010k} + v_{01jk},$$
$$\pi_{02jk} = \gamma_{020k} + v_{02jk},$$
$$\pi_{10jk} = \gamma_{100k} + v_{10jk},$$
$$\pi_{11jk} = \gamma_{110k},$$
$$\pi_{12jk} = \gamma_{120k},$$
$$\pi_{20jk} = \gamma_{200k} + v_{20jk},$$
$$\pi_{21jk} = \gamma_{210k},$$
$$\pi_{22jk} = \gamma_{220k}.$$

For all nine student-level parameters there was an equation at the classroom level allowing for intercept and slope differences between classrooms. For simplicity reasons no classroom-level predictors were used in the models. The random effects at Level 3 (v_{00jk}, v_{01jk}, v_{02jk}, v_{10jk}, v_{20jk}) are assumed to be multivariate normally distributed with zero mean and a constant (5 × 5) variance–covariance matrix. The random part was simplified when random slopes or slope–slope covariances were not significant.

Level 4: School Differences

$$\gamma_{000k} = \varphi_{0000} + \varphi_{0001}\text{Intervention} + \varphi_{0002}\text{Swedish} + f_{000k},$$
$$\gamma_{010k} = \varphi_{0100} + \varphi_{0101}\text{Intervention},$$
$$\gamma_{020k} = \varphi_{0200} + \varphi_{0201}\text{Intervention},$$
$$\gamma_{100k} = \varphi_{1000} + \varphi_{1001}\text{Intervention} + f_{100k},$$
$$\gamma_{110k} = \varphi_{1100},$$
$$\gamma_{120k} = \varphi_{1200},$$
$$\gamma_{200k} = \varphi_{2000} + \varphi_{2001}\text{Intervention} + f_{200k},$$
$$\gamma_{210k} = \varphi_{2100},$$
$$\gamma_{220k} = \varphi_{2200}.$$

Intervention effects were defined at the school level. There is an intervention effect specified in the equation for the intercepts (γ_{000k}). This represents the baseline differences between intervention and control schools. The equation for the intercept has in addition a control variable to account for possible differences between Swedish- and Finnish-speaking schools.The real intervention effects are in the equations for the coefficients of T2 and T3 (γ_{100k} and γ_{200k}): differences in average change scores compared with baseline for the intervention and control schools at T2 and T3. The intervention effects are therefore represented in Tables 3–5 as interactions: T2 × Intervention and T3 × Intervention. The intervention effects for the slopes of Boy (γ_{010k}) and Age (γ_{020k}) represent the Intervention × Boy and Intervention × Age interactions at baseline.

The school-level equations were extended with product terms for possible cross-level interactions of intervention effects at T2 and T3 with gender and age of students. These three-way interactions are included by specifying Intervention as a predictor for the slopes of the changes across time on gender and age, γ_{110k}, γ_{120k}, γ_{210k}, and γ_{220k}.The random effects at Level 4 (f_{000k}, f_{100k}, f_{200k}) are assumed to be multivariate normally distributed with zero mean and a constant (3 × 3) variance–covariance matrix.