Prev Sci (2009) 10:248–259
DOI 10.1007/s11121-009-0128-y

Antibullying Programs in Schools: How Effective are Evaluation Practices?

Wendy Ryan · J. David Smith

Published online: 6 March 2009

Abstract Bullying is a problem for schools around the world, and is an important topic for research because it has been associated with negative outcomes on numerous social, psychological, and academic measures. Antibullying prevention and intervention programs have varied greatly in their outcomes, with some studies reporting positive results while others have reported little or no positive impacts. Prompted by accountability demands, many agencies have developed standards with which to assess whether social programs are effective. Antibullying program evaluations have not been systematically reviewed to determine whether these types of standards are being applied. The purpose of this study was to assess the rigor of recent peer-reviewed antibullying program evaluations. Thirty-one peer-reviewed evaluations of antibullying programs, published within the last 10 years, were identified and coded for study characteristics. Shortcomings were identified in many of these program evaluations. In order to improve evaluation practices, researchers should consider using more rigorous designs to identify cause-effect relationships, including control conditions and random assignment, using more appropriate pre-post intervals, using more advanced methods of analyses such as hierarchical linear modeling, and systematically verifying program integrity to obtain dosage data that can be used in the outcome analyses.

Keywords Antibullying · Program · Evaluation practices

W. Ryan (✉) · J. D. Smith
University of Ottawa,
Ottawa, ON, Canada
e-mail: wendyryan.ottawa@gmail.com

Bullying is a subtype of aggressive behavior characterized by the intent to harm, repetition of attacks, and abuse of power over a weaker victim (Olweus 1991). Besides direct physical or verbal aggression, bullying can include indirect forms such as group exclusion or gossip (Crick and Bigbee 1998), and sometimes occurs through electronic means such as email or cellular phones (Patchin and Hinduja 2006). Research evidence indicates clearly that involvement in bullying is detrimental to children's academic success and their physical and mental health (Orpinas and Horne 2006). Victimized children tend to display internalizing symptoms, including anxiety, depression, diminished self-esteem, and social withdrawal (Nansel et al. 2001). Children who bully as well as those who are victimized appear to be more vulnerable to depression and suicidal ideation than their non-involved peers (Roland 2002).

A worrisome consequence for children who bully others is susceptibility to future problems of violence and delinquency. For example, one study revealed that adolescent bullies viewed their dating partners less equitably and reported higher rates of aggression in those relationships than non-bullies (Connolly et al. 2000). Childhood aggression and peer rejection, both of which are operative in bullying, have been identified as the most powerful predictors of future social and behavioral problems, including adult aggression and criminality (Coie 2004).

It is not surprising in this context that educators, parents, and students are very concerned about bullying and, as a consequence, that schools have directed increasing amounts of resources to reducing bullying (J. D. Smith et al. 2007). In fact the problem is deemed to be so serious that many U. S. states are required by law to implement antibullying interventions (Limber and Small 2003).

The Olweus Bullying Prevention Program was the first comprehensive "whole-school" antibullying program

implemented on a large scale and systematically evaluated (Olweus 1993), and many other antibullying programs have been based on this model (see P.K. Smith, Pepler, & Rigby 2004). Whole-school programs include: information on bullying for school staff, students, and parents; clear, consistent school-wide policies for bullying; classroom activities that promote antibullying attitudes and teach prosocial methods of conflict resolution; and interventions for students affected by bullying.

Despite the positive results shown by Olweus (1991) and the widespread use of such programs, recent research questions their effectiveness. J. D. Smith et al. (2004) quantitatively synthesized the results of 14 evaluation studies of whole-school antibullying programs. Outcomes were mostly negligible (i.e., effect size $r \leq .09$) or negative. Only one study yielded an outcome that was categorized as medium (i.e., the Olweus Bullying Prevention Program), and none was categorized as large. Two other reviews of antibullying program evaluations have also shown mixed results. Baldry and Farrington (2007) reviewed 16 antibullying program evaluations and found that 8 yielded desirable results, 4 produced small or negligible effects, 2 showed mixed results, and 2 produced undesirable effects. In Vreeman and Carroll's (2007) review, only 3 of the 21 studies that measured direct behavioral outcomes for bullying and victimization yielded consistently positive outcomes.

In this context, educators have the difficult task of selecting bully prevention programs that have the best chance of succeeding in their schools. The imperative to use empirical evidence in school decision-making can be linked to government initiatives that tie the allocation of funding to the use of evidence-based educational interventions; e.g., the No Child Left Behind Act (NCLB 2002). As a result, there have been many initiatives to develop criteria for assessing the evidence underlying educational and psychosocial programs. For example, The Prevention Research Center for the Promotion of Human Development at Pennsylvania State University has created a technical assistance fact sheet on evidence-based programs, outlining the standards that they use to judge programs (Kyler et al. 2005). The Society for Prevention Research has acknowledged the wide variety of criteria for evaluating program effectiveness, and has produced its own set of standards specifically for evaluating the efficacy and effectiveness of prevention programs and policies (Flay et al. 2005).

Demands for accountability have led several agencies in the United States (e.g., Office of Juvenile Justice and Delinquency Prevention, Center for Mental Health Services) to review social programs and make assessments as to whether they can be labeled "evidence-based." Each of these agencies has its own set of criteria and categories for program effectiveness, which can make it difficult to assess specific programs. In order to facilitate the search for an evidence-based program, Mihalic (2007) from the Center for the Study and Prevention of Violence Blueprints Initiative, has compiled a matrix of approximately 300 social programs showing how these have been rated in terms of effectiveness across 12 different agencies. Of the 300 programs reviewed only 2 of them, *Steps to Respect: A Bullying Prevention Program*, and the *Olweus Bullying Prevention Program*, specifically target bullying. This paucity of ratings for antibullying programs is another indication of the lack of evidence available to assist schools with their programming decisions.

We believe that the limited rigor of evaluation studies on antibullying programs conducted to date may constitute an obstacle to making conclusive statements about the effectiveness of bully prevention programs deployed in many schools today. Both Vreeman and Carroll (2007) and J. D. Smith et al. (2004) examined some aspects of the level of rigor of studies reviewed, noting, for example, whether or not studies were controlled and the methods of group assignment. However, it was not the objective of their reviews to assess the quality of evaluation methods per se, so neither gives a comprehensive picture of the current state of evaluation practices in bullying prevention programs. The primary objective of this review study is to undertake such a task. Such a review could provide a significant impetus to advancing knowledge in this domain, which seems critical at this time given the significant resources being allotted to bullying prevention programs across North America. We conclude with a series of recommendations to researchers and educators about how the quality of evaluation practices in antibullying programs can be improved and knowledge advanced on this pressing social issue.

Methods

Document Search Strategy

In order to locate a comprehensive set of recent reports of antibullying program evaluations, Medline, PsycInfo, and ERIC databases were searched using the following keywords: (bully* or antibullying or anti-bullying) and (program or intervention or evaluat*) and school. This method yielded 550 articles. The reference lists of significant review papers in the bullying prevention field as well as articles that ultimately met inclusion criteria for this study were scanned to identify other possible evaluation studies for consideration for this study. The research reports were reviewed and retained only if they (a) evaluated an intervention intended to prevent bullying in schools, (b) reported data on student outcomes directly related to bullying and/or victimization, and (c) were published in English. Because our goal was to

examine the current state of the art of evaluation practices in this domain, we sought to assemble a database of the highest quality evaluation studies. Consequently, we limited our selection to written reports that had been subjected to peer review as a standard part of the publication process. Therefore, non peer-reviewed works, such as theses, dissertations, and unpublished technical reports, were excluded from our analyses. In a number of instances, multiple published reports of the same evaluation study were located. In these cases, the earliest published report with the most complete description of the study was retained for analysis. In several instances, a second report on the same evaluation was consulted in order to attain all the necessary information for the present review study. Finally, reports that were published before 1997 and reports based on data collected before 1995 were excluded. Our search for relevant documents ended in March 2007, at which time we had identified 31 bullying prevention evaluation studies that met our criteria for inclusion in this study.

Of the original 550 articles reviewed, 414 were excluded because the article was not an evaluation of an intervention intended to prevent/reduce bullying in schools; 63 were excluded because the article was published before 1997 or the data were collected before 1995; 6 were excluded because the evaluation did not report outcomes directly related to bullying or victimization; 9 were excluded because the article was not published in English; and 11 were excluded because they were not peer reviewed (these were dissertations). That left 47 articles, 16 of which were duplicates, leaving us with 31 studies that met our criteria.

Coding Procedures

Because it would have been impractical and inefficient to code for all of the standards of evidence described by Flay et al. (2005), we selected features that we believe are most relevant to antibullying program evaluators in order to improve evaluation practices. See Table 1 for the methodological features that were coded for each article. Two raters working independently double-coded the methodological features of five randomly selected reports to assess inter-rater agreement. The kappa coefficient for inter-rater agreement for this set of double-coded studies is .76.

Results

The 31 evaluation studies reviewed in this investigation comprise a wide variety of different types of antibullying programs. Most of the programs included a classroom component (77.4%) and/or a school-wide component (61.3%). Some programs included specific interventions involving peers (38.7%), individuals (35.5%), parents (35.5%), and/or the community (9.7%). About half of the programs included at least three of these components, and, as such, may be considered "whole-school" programs (J. D. Smith et al. 2004). The mean number of components within programs is 2.5 (SD=1.48), while the mode is 1.

Sample sizes in these 31 studies ranged from under 50 participants (two studies) to over 1000 participants (nine studies). The most common sample size (characteristic of 11 studies) is in the range of 200–499 participants.

The program evaluation features for the 31 studies identified for this review are presented under the following headings: program monitoring, study design, outcomes, statistical analyses and study type. Table 1 provides an overview of these findings.

Program Monitoring

We used Dane and Schneider's (1998) coding scheme for program integrity promotion and verification in order to get a more nuanced view of the scope of program monitoring undertaken within the context of each evaluation. Integrity promotion includes providing training manuals, training facilitators, and supervising implementers. Integrity verification is the systematic documentation of efforts to ensure that a program is delivered as intended. This can take the form of monitoring how closely implementers follow the procedures outlined in the program manual (i.e., adherence), measuring the level of exposure to the program (i.e., dosage), verifying the quality of delivery, assessing participant responsiveness, and determining the degree of diffusion of interventions among treatment groups (i.e., program differentiation).

Program integrity promotion Some form of program integrity promotion was reported in all but 1 of the 31 studies in our sample. Specifically, in 64.5% of the studies, provision of a manual was mentioned; in 80.1% of the studies, training was provided for those who were administering the program; and in 22.6% of the studies, supervision of program implementers was undertaken. Only 16.1% of the studies included three forms of program integrity promotion (manuals, training, and supervision). The mean number of forms of integrity promotion across the 31 studies in this review is 1.8 (SD=.72).

Program integrity verification Some form of program integrity verification was reported in the majority of studies; however, 38.7% of the studies did not report any form of integrity verification. The mean number of methods of integrity verification used within the set of studies is 1.1 (SD=1.09).

Adherence, the extent to which program components were delivered as prescribed by program manuals, was the

Table 1 Study characteristics

Study	Program components[a]	Total N (N schools)[b]	Program Monitoring		Study Design				Outcomes			Informant		Statistical Analyses			Study type[f]
			Promotion[c]	Verification[d]	Controlled	Random	Qualitative	Months to post-test (follow-up)	Bully/ victim	Other behavior	Non-behavior	Self	Others	Scales[e]	Effect sizes	Multilevel	
Hunt (2007)	a,b,e	c(5)	m,t		●	●		12	●		●	●		r,v	●		p
Fekkes et al. (2006)	a,b,e	e(41)	m,t	a,e,d	●	●		6(18)	●	●	●	●				●	p
Gollwitzer et al. (2006)	b	b(2)	m,t,s	a,e,q,p	●		●	4(8)	●	●		●	●	r,v	●		ey
Heydenberk et al. (2006)	b	c (2)	m,t	p	●		●	12(24)	●	●	●	●	●	v			p
Kim (2006)	d	a(1)	m,t,s		●	●		1	●		●	●					p
Leadbeater et al. (2003); Leadbeater & Hoglund (2006)	a,b,c,e,f	c(17)	m,t,s	a,e	●			8(20, 32,44)	●	●		●	●	r	●	●	es
McLaughlin et al. (2006)	b	b(3)			●	●		2	●			●		r,v	●		p
Edwards et al. (2005)	b,e	c(1)	m,t	q,p			●	4		●	●	●	●	r,v	●		p
Frey et al. (2005)	a,b,e	d(6)	m,t	a,e,q	●	●		6	●	●	●	●	●	r,v	●	●	p
Jennifer & Shaughnessy (2005)	a,b	c(10)	m	q	●		●	12	●			●					p
Mooij (2005)	a,b,f	e	t	a				108	●	●		●		r,v	●		p
Olweus (2005)	a,b,c,d,e	e	m		●			8(24)	●			●				●	p
Baldry & Farrington (2004)	b	c(3)	m		●	●		4	●			●		r,v			p
Cross et al. (2004)	a,b,c,e	e(29)	m,t	a,e	●	●		8(20)	●		●	●		v			p
DeRosier (2004)	d	c(11)	m		●	●		6	●	●	●	●	●		●		p
Limber et al. (2004)	a,b,d,e,f	e(12)	m,t,s	q				12(24)	●	●	●	●					p
Newman-Carlson & Horne (2004)	a	a(1)	m,t,s	a	●			2	●	●	●		●	r,v			p
O'Moore & Minton (2004, 2005)	a,b,c,d,e	d(22)	m,t					12	●			●					p
Ortega et al. (2004)	a,b,c,d,e	e(9)	t	q	●			48	●		●	●					p
Rosenbluth et al. (2004) and Whitaker et al. (2004)	a,b,e	e(12)	m,t		●	●		3(8)	●		●	●	●	v		●	p
Salmivalli et al. (2004, 2005)	a,b,d	e(16)	t	a,e	●			6(12)	●		●	●	●	r,v	●	●	p
Menesini et al. (2003)	b,c	c(2)	t		●			8	●		●	●	●	v			p
Orpinas et al. (2003)	a,b	d(1)	t					12	●	●		●		r,v	●		p
Rahey & Craig (2002)	a,b,c,d	c(2)	m,t	a	●			4	●	●		●	●	r,v			p
Alsaker & Valkanover (2001) and Alsaker (2004)	b	c(16)	s	a,d	●			7	●			●	●				p

Table 1 (continued)

Study	Program components[a]	Total N (N schools)[b]	Program Monitoring		Study Design			Outcomes			Informant		Statistical Analyses		Study type[f]
			Promotion[c]	Verification[d]	Controlled	Random	Qualitative	Months to post-test (follow-up)	Bully/ victim	Other behavior	Non-behavior	Self	Others	Scales[e]	Effect sizes	Multilevel	
Salmivalli (2001)	a,b,c	b(1)	m,t	p				1	●		●	●	●	r,v			p
Cowie & Olafsson (2000)	c	c(1)	t,s	q,p			●	7.5	●			●					p
Meyer & Lesch (2000)	d	b(3)	t		●	●		2(3)	●			●	●				p
Stevens et al. (2000; 2001; 2004)	a,b,c,d	d(18)	m,t	a,e	●	●	●	9(21)	●	●	●	●		r,v	●		es
Peterson & Rigby (1999)	a,c,d	d(1)	t				●	24	●			●					p
Naylor & Cowie (1999)	c	e(51)	t	q,p					●		●	●	●				p

[a] Program components: a=school-wide; b=classroom; c=peer; d=individual; e=parents; f=community. [b] N: a=1–49; b=50–199; c=200–499; d=500–999; e=1000+. [c] Promotion of program integrity: m=manual; t=training; s=supervision. [d] Verification of integrity: a=adherence; e=exposure; q=quality of delivery; p=participant responsiveness; d=program differentiation. [e] Psychometric properties of scales: r=reliability; v=validity. [f] Study type: p=pilot; ef=efficacy; es=effectiveness

most commonly reported form of verification (i.e., reported in 35.5% of studies). Exposure, a measure including the frequency of program activities, and/or the number and length of sessions implemented, was reported in 22.6% of the studies. Quality of delivery, including implementer attitudes toward program, preparedness, and perceptions of session effectiveness, was reported in 22.6% of studies. Participant responsiveness, including levels of participation and enthusiasm, was reported in 19.3% of studies. Program differentiation, a check to ensure that subjects in each experimental condition received only planned interventions, was reported in 6.4% of studies. Sources of verification included questionnaires administered to teachers, students, and/or facilitators; interviews with students and/or teachers; observations; and diaries. Only three research teams (Mooij 2005; Salmivalli et al. 2004, 2005; Stevens et al. 2000, 2001, 2004) reported measuring degree of program implementation to use in their impact analysis.

Study Design

Of the 31 program evaluations included in this review, 22 were controlled studies and 9 studies were uncontrolled studies. Eleven controlled studies featured random assignment to intervention and control conditions, with seven randomized at the school level and four randomized at the individual student level.

Time to post-test/follow-up In two-thirds of the studies, outcomes were reported for only one post-test. The time to post-test ranged from 1 month to 108 months. The mode was 6 months, and the mean (after removing the outlier of 108) was 8.6 months (SD=8.97). In one-third of the studies, outcomes were reported for at least one additional follow-up time. Using the longest post-test time for each study, the time to post-test/follow-up was less than 6 months in 25.8% of the studies, between 6–11 months in 29% of the studies, between 12–23 months in 25.8% of the studies, and 2 years or more in 19.3% of the studies.

Qualitative component Less than one-fifth of the studies included a qualitative component to their design. Qualitative components included interviews, diaries, observations, and open-ended questionnaires. These were either used as the main form of data or were included to supplement quantitative data by providing richer details about the context in which the antibullying program took place and provide a means of triangulation. Gollwitzer et al. (2006) used open-ended questions to measure quality of program implementation and accomplishment of training principles. They used students' responses to hypothetical vignettes as a measure of enrichment of behavioral repertoire. In the Heydenberk et al. (2006) study, open-ended questions

administered to teachers were used to gain additional data on program effectiveness and changes in student behavior. Edwards et al. (2005) used open-ended, semi-structured interviews with students to collect data on the acceptability and impact of the *Second Step* program. Jennifer and Shaughnessy (2005) collected qualitative data using multiple sources and multiple methods, for example, semi-structured interviews with pupils, managers, and facilitators; observations of several aspects of school life; and facilitator diaries. Cowie and Olafsson (2000) conducted interviews with staff, peer supporters, users, and potential users of the peer support service in order to supplement survey data in assessing the impact of the peer support service. Stevens et al. (2001) conducted semi-structured interviews with project leaders in order to gain insights into the program implementation process at each school. Peterson and Rigby (1999) collected written comments from students in order to supplement survey data.

Outcomes

Types of outcome measures used in each study were divided into three categories: 1) behavioral measures of involvement in bullying/victimization; 2) measures of other behaviors, such as aggression, prosocial behavior, and coping; and 3) non-behavioral constructs such as attitudes or beliefs. Forty-eight percent of the studies used two types of outcome measures, 22.6% of the studies used three types of outcomes, and 29% of the studies used only one type of measure. All but one study (Edwards et al. 2005) reported a behavioral measure of involvement in bullying/victimization as an outcome measure.

Researchers used a wide variety of different surveys to measure involvement in bullying. The most common measure, used in 35.5% of studies, was a modified version of the *Olweus Bully/Victim Questionnaire*. All but one study (Newman-Carlson & Horne 2004) used student self-reports as outcome measures. In 45.2% of studies, reliability was reported for at least one measure. In 54.8% of the studies, some evidence of the validity of at least one measure was presented. Studies were also analyzed to determine how many of the four possible informants (self, peers, teacher, parent) were asked to provide data. More than half (54.8%) of the studies used only one type of informant, and 38.7% of the studies used two types of informants. Only one study (Rahey and Craig 2002) used all four types of informants, and one other study (Gollwitzer et al. 2006) used three types of informants.

Statistical Analyses

In nearly all of the studies, descriptive statistics and significance tests were reported. Effect sizes were reported in only 35.5% of the studies. Multilevel statistical techniques, such as hierarchical linear modeling (HLM), were reported in only five studies.

According to Wolff (2000), randomized controlled trials are based on the following assumptions: "standardized interventions, equal groups and equal trial environments" (p. 98). None of the randomized controlled trials in this sample reported meeting all three of these assumptions. Six of the 11 RCTs evaluated a standardized intervention. Seven of the 11 reported checking for equal groups and 4 of these presented evidence that this assumption was met. Only one study reported equal trial environments. Furthermore, there were four studies (Cross et al. 2004; Frey et al. 2005; Hunt 2007; Rosenbluth et al. 2004) that randomly assigned schools to conditions but conducted statistical analyses at the level of the individual student.

Study Type

All studies included in this review were coded based on the criteria for efficacy, effectiveness, and dissemination trials developed by the Standards Committee of the Society for Prevention Research (see Flay et al. 2005). By the strictest application of the *Standards*, none of the 31 studies included in this review meet the required criteria for an efficacy, effectiveness, or dissemination study. To extract additional information from the data that would be useful for this review, a less stringent set of criteria was developed for coding studies into these three categories. Operational definitions of study type were developed by including only the criteria from the *Standards* that can be met at the inception of the evaluation study and excluding criteria that in most cases could be met after data have been collected; for example, type of statistical analyses and specification of sample characteristics. Hence, the following criteria were used for coding of study type:

1. Efficacy: a detailed program description that permits program replication; measures of relevant behavioral outcomes; a control condition; adequate procedures for group assignment (randomization or matching groups); psychometrically sound instruments; long-term follow-up on outcomes.
2. Effectiveness: the previous efficacy criteria; program delivered under real-world conditions (e.g., by school personnel); measures of program fidelity and exposure.
3. Dissemination: implementation of an efficacious program; complete program materials; monitoring and evaluation tools; cost information.

The final column in Table 1 displays the results of coding of study type on these criteria. Studies that did not meet the criteria for any of these three categories were coded as a pilot study. As these results show, only one

study was coded as an efficacy study and two were coded as effectiveness studies. The remaining 28 studies did not meet criteria for efficacy, effectiveness, or dissemination and were therefore classified as pilot studies.

Discussion

Governments world-wide have invested significant resources to address bullying problems that afflict schools, and school authorities want to use those resources effectively. Since 2004, several reviews of antibullying program outcomes have been published (e.g., Merrell et al. 2008; J. D. Smith et al. 2004) and all point to the same conclusion: They have not been shown to be effective in reducing bullying. Accepting this conclusion as final presumes that the methods used to assess the program effectiveness have been rigorous and comprehensive. We tested the validity of this premise in a review of 31 studies, selected following an exhaustive search for reports on antibullying program effects. Because we were most interested in assessing the current state of knowledge in this domain, we limited our review to recent studies (1997–2007) published in peer-reviewed journals, assuming that this is where we would find the highest quality examples of this work.

The results of our analyses lead us to conclude that antibullying program evaluations conducted to date are not adequately rigorous and comprehensive in their design and execution to accept their results as final. The current state of knowledge about the effectiveness of bullying prevention programs is perhaps best reflected in the fact that no evaluation included in this review met the complete criteria for an efficacy trial or an effectiveness trial, as defined by the Standards of Evidence Committee (2004). Even after coding studies on less stringent criteria, less than 10% of the sample qualified as either efficacy or effectiveness under these conditions. This suggests that the knowledge base on bullying prevention is in an early phase of development and much work remains to be done. This conclusion evokes a number of important implications, two of which we address in the remainder of this paper: How can evaluation practices in the antibullying domain be improved, and what steps must researchers take to generate superior quality information on which school officials can base their programming decisions?

Program Monitoring

The outcomes of a prevention program can really only be interpreted accurately in light of information about how it was implemented. Program integrity monitoring should be standard in all evaluation studies so that one can distinguish between programs that do not work from those that are inadequately implemented (Greenberg et al. 2005). Dane and Schneider (1998) discuss implementation in terms of integrity promotion and integrity verification. While all studies in this review reported at least some form of program integrity promotion (i.e., manuals, training, supervision), there were some shortcomings in the area of program integrity verification. Dane and Schneider (1998) assert that each of the five aspects of integrity verification (i.e., adherence, exposure, quality of delivery, participant responsiveness, and program differentiation) is important to monitor. However, more than one-third of studies in our review reported no integrity verification strategies, and across studies the average number of methods used for integrity verification was approximately 1. Because pure experimental research designs in school settings are always impractical and often impossible, using program dosage measures as covariates in impact analyses can bolster arguments for causality of program effects in the context of correlational studies. Results of this review indicate that this technique is too infrequently used, as only 3 of the 31 evaluations used data on the degree of program implementation or a dosage measure in their analyses of outcomes. Clearly, program monitoring is an aspect of antibullying program evaluations that requires further attention. It is recommended that researchers gather sufficient data on program implementation, and use these data in impact analyses. Linking outcomes to implementation and dosage, particularly in the context of correlational designs, can strengthen arguments of causality about program effects.

Study Design

For efficacy and effectiveness studies, the randomized controlled trial (RCT) is considered the "gold standard" for evaluating the impact of interventions, as it is the only design that can adequately control for external confounding factors (Torgerson and Torgerson 2001). However, many studies using RCTs fail to conduct empirical checks on assumptions such as group equivalence (Chatterji 2007). In fact, research by Wolff (2000) indicates that evaluations of "socially complex services," which subsume antibullying programs, often violate the assumptions of the RCT, thereby drawing into question "the validity, reliability, and generalizability of inferences of socially complex service trials" (p. 97). RCTs are based on the assumptions of standardized interventions, equivalent groups and trial environments. None of the 11 RCTs in our sample met all of these assumptions, which unfortunately undermines the value of their findings. In order to verify whether or not assumptions are met, it is important that interventions, group characteristics, and trial environments are clearly defined and systematically monitored throughout the trial.

Olweus (2005) argued that an "extended selection cohorts quasi-experimental design" is a viable alternative

when random assignment is not possible or desirable. This design provides time-lagged comparisons between age-equivalent groups: Pre-test measures are compared to data collected from students at the same grade at subsequent post-test and follow-up intervals. The design appears to have sufficient rigor to make valid conclusions about the effects of a school-based intervention. It is particularly critical that implementation is thoroughly assessed to rule out alternate explanations for outcomes beyond the effects of the prevention program.

For programs without manuals or that are loosely defined in conceptual terms (e.g., peer support initiatives), efficacy and effectiveness trials are not possible, as the programs cannot be replicated with precision. In fact, some programs "involve a flexible, iterative and evolving process" and have manuals that "make it clear that programme components should be devised to suit local conditions" (Pawson and Tilley 1998, p. 212). Since randomized controlled trials are based on the assumption of standardized interventions, such loosely defined interventions cannot be subjected to this type of evaluation. However, evaluations of these types of programs can be buttressed by qualitative methods that provide details about the nature of the intervention and insights into what works for whom under what circumstances. It is also important to state the logic model that the intervention is based on, and use data collected to provide evidence supporting or refuting the program theory.

Qualitative data Qualitative data can provide rich description about environmental factors to improve the "meaning and clarity of statistical effects" in experiments and quasi-experiments (Chatterji 2007, p. 252). However, relatively few studies (less than one-fifth) in this review included any qualitative methods in their design. The qualitative data that were collected took a variety of forms and served several functions. Methods included open-ended questionnaires, interviews, observations, and diaries; data were collected from students, teachers, facilitators, project leaders, and managers. In some cases, qualitative methods were used to measure the quality of program implementation or to assess program impact. In these studies, the addition of qualitative data served to contextualize the results, and facilitated the interpretation of evaluation data. Researchers should consider integrating qualitative methods in their designs when feasible to help contextualize quantitative results and to probe implementation issues in more depth.

Time to post-test/follow-up Another weakness of many of the antibullying program evaluations was that the length of time to post-test was short. According to the *Standards*, outcomes should be measured "at least 6 months after the intervention" (Flay et al. 2005, p. 155). More than one-quarter of the antibullying program evaluations reviewed for this study failed to meet this standard. According to some researchers it takes at least 2 years for measurable change to occur as a result of program implementation, suggesting that even the 6-month standard is rather short (e.g., Hall and Hord 2006). In about one-fifth of the evaluation studies, post-test or follow-up data were collected 2 years or more after program implementation. Short post-test intervals can confound the interpretation of evaluation data for several reasons. Firstly, the programs may simply not have had enough time to produce change in the outcomes being measured. Secondly, implementation issues in the early phase of a program will likely affect outcomes and may mask the potential of promising programs. Consequently, it is usually more efficacious to do impact evaluations on mature programs instead of new ones. Therefore, it is recommended that longer pre-post intervals should be used so that the antibullying program has time to take effect, and its impacts should be assessed at subsequent follow-up intervals. Ideally, data should be collected at annual intervals over a 3-year period to control for seasonal effects on outcomes. Annual follow-ups may not be necessary for pilot studies and general implementations because these types of evaluations generally aim to refine methods and explore what works for whom under what circumstances.

Outcomes

Leff et al. (2004) recommend that evaluators use multiple methods and multiple informants in order to get a clearer picture of program effectiveness. In the current sample of antibullying program evaluations, most studies used two or more different outcome measures, but nearly a third included only one type of measure. Additionally, more than half of the evaluations included only one type of informant. Data from multiple sources can increase confidence in findings. This is particularly important in bullying prevention studies, as research indicates that there are systematic differences among informants (e.g., students, parents, and teachers) on reporting rates of bullying and victimization (Pellegrini and Bartini 2000). Therefore, collecting data about bullying from a variety of informants is necessary to construct a reasonably complete picture of bully/victim problems in a school and, by implication, to understand the influence of interventions in reducing their occurrence. It is, therefore, recommended that antibullying program evaluators use multiple methods and multiple informants in collecting outcome data.

Researchers in our sample used a wide variety of outcome measures, which increases the difficulty of making comparisons across studies, particularly as small differences in the scales (reporting bullying over the previous month

versus the previous semester) can have large effects on results. The field of bullying prevention would benefit from a concerted and coordinated effort from researchers to develop a common measure of bullying and victimization for evaluation purposes and that is preferably available in the public domain. Compounding the problem of numerous measures, validity and reliability of these scales were too infrequently addressed. In more than half of the studies, no data on instrument reliability or validity were reported. Clearly, antibullying program evaluators need to carefully choose measures that are reliable and valid, and present evidence of this in their reports.

Statistical Analyses

The American Psychological Association lists the failure to report effect sizes as a common defect in reporting of research results (APA 2001), and we found that our sample of antibullying program evaluations is no exception. Effect sizes were reported in only one-third of the evaluations reviewed, and few of these reported confidence intervals. Thompson (2002) promotes the reporting of confidence intervals for effect sizes and argues that this practice facilitates meta-analytic thinking. Effect sizes are particularly important in describing program outcomes, as program users are likely to be more interested in knowing how much of a difference a program can make, rather than simply if a program can make a difference.

Using multilevel statistical models such as hierarchical linear modeling (HLM) is indicated in cases in which units are nested within other units (Bryk and Raudenbush 1992). Schools are a prime setting for using hierarchical statistical models, given the nesting of students in classrooms and classrooms within schools and so on. While hierarchical modeling seems particularly appropriate for the evaluation of bullying prevention programs in schools, it appears to be infrequently used. Of the 31 studies reviewed for this paper, only 5 studies used multilevel statistical modeling. Consequently, it is recommended that researchers use these techniques whenever feasible. Computer software is available to determine what sample sizes are necessary at each level (school, class, individual) depending on the statistical power and confidence intervals required (Okumura 2007). Using multilevel statistical techniques can increase the precision with which program effects are measured, and may clarify where the net benefits of such prevention efforts primarily accrue.

Conclusion

This review of recent bullying prevention studies suggests that evaluation practices in this domain have not yet reached a level of rigor that permits us to accept their outcomes as conclusive. We encourage readers to consider this conclusion in light of several limitations of this review. The results of this review are limited to the evaluation features selected by the authors for study. We selected features that we believed would be most relevant to antibullying program evaluators and that we felt are vital to improving evaluation practices in this area. It is nonetheless possible that other relevant features were excluded that might have altered the findings of this review. The review is also limited by its focus on published peer-reviewed studies to the exclusion of all other evaluations of antibullying programs including theses and dissertations. The 11 dissertations that were excluded because of our criteria may have included high quality evaluations of antibullying programs, and it is possible that if these had been included in our review our findings may have been somewhat altered. However, we believe that given the nature of the publication enterprise and the peer-review system, our decision to exclude dissertations is a valid one, and the chance is small that such evaluations would have significantly altered our conclusions.

There are tremendous resources being committed to antibullying programs in North American schools in the absence of a body of compelling evidence in the prevention literature that these programs are actually substantially reducing bullying. In this context, program evaluators and prevention scientists have a critical and pressing role to provide clear and accurate information to inform public policy on bullying prevention in schools. To this end, we encourage evaluators and researchers to consider our recommendations, which are recapped below, for future evaluations of antibullying programs.

1. Use control conditions and random assignment where possible, and consider a rigorous quasi-experimental design (e.g., Olweus 2005) when an experimental design is not feasible.
2. Collect baseline data before the intervention has been introduced, and collect outcome data at least 6 months after implementation. For efficacy and effectiveness trials collect follow-up data annually thereafter for 2 or more years.
3. Use multiple methods and multiple informants to assess program impacts.
4. Report evidence of reliability and validity for instruments used.
5. Collect qualitative data to contextualize implementation and outcome data.
6. Systematically monitor program integrity and use program dosage data in outcome analyses.
7. Use multilevel statistical modeling to analyze data that have been collected from individuals from several

classrooms, nested within several schools when sample sizes are large enough.

We recognize that conducting research in schools presents particular challenges, and that researchers often work with constraints that can thwart their intentions to conduct rigorous research. Therefore, evaluators should routinely discuss these issues with school personnel and, where appropriate, educate them about the importance and value of methodological rigor. They should also work with schools to devise ways to accommodate the research requirements with the school's mandate to provide services to students. Researchers should also demonstrate how their involvement with a prevention initiative adds value to the program and can assist the school to achieve their educational goals. In the end, complete and valid information about school-based prevention programs is in everyone's best interest, and researchers can facilitate their work by explaining this in accessible terms to all stakeholders in prevention programs.

References

Alsaker, F. D. (2004). Bernese programme against victimization in kindergarten and elementary school. In P. K. Smith, D. Pepler, & K. Rigby (Eds.), *Bullying in schools: How successful can interventions be?* (pp. 289–306). New York: Cambridge University Press.

Alsaker, F. D., & Valkanover, S. (2001). Early diagnosis and prevention of victimization in kindergarten. In J. Juvonen & S. Graham (Eds.), *Peer harassment in school: The plight of the vulnerable and victimized* (pp. 175–195). New York: Guilford Press.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

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. doi:10.1002/ab.20000.

Baldry, A. C., & Farrington, D. P. (2007). Effectiveness of programs to prevent school bullying. *Victims and Offenders*, *2*, 183–204. doi:10.1080/15564880701263155.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Chatterji, M. (2007). Grades of evidence: Variability in quality of findings in effectiveness studies of complex field interventions. *The American Journal of Evaluation*, *28*, 239–255. doi:10.1177/1098214007304884.

Coie, J. D. (2004). The impact of negative social experiences on the development of antisocial behavior. In J. B. Kupersmidt & K. A. Dodge (Eds.), *Children's peer relations: From development to intervention* (pp. 243–267). Washington, DC: American Psychological Association.

Connolly, J., Pepler, D., Craig, W., & Taradash, A. (2000). Dating experiences of bullies in early adolescence. *Child Maltreatment*, *5*, 299–310. doi:10.1177/1077559500005004002.

Cowie, H., & Olafsson, R. (2000). The role of peer support in helping the victims of bullying in a school with high levels of aggression. *School Psychology International*, *21*, 79–95. doi:10.1177/0143034300211006.

Crick, N. R., & Bigbee, M. A. (1998). Relational and overt forms of peer victimization: A multiinformant approach. *Journal of Consulting and Clinical Psychology*, *66*, 337–347. doi:10.1037/0022-006X.66.2.337.

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.

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. doi:10.1016/S0272-7358(97)00043-3.

DeRosier, M. E. (2004). Building relationships and combating bullying: Effectiveness of a school-based social skills group intervention. *Journal of Clinical Child and Adolescent Psychology*, *33*, 196–201. doi:10.1207/S15374424JCCP3301_18.

Edwards, D., Hunt, M. H., Meyers, J., Grogg, K. R., & Jarrett, O. (2005). Acceptability and student outcomes of a violence prevention curriculum. *The Journal of Primary Prevention*, *26*, 401–418. doi:10.1007/s10935-005-0002-z.

Fekkes, M., Pijpers, F. I. M., & Verloove-Vanhorick, P. (2006). Effects of antibullying school program on bullying and health complaints. *Archives of Pediatrics & Adolescent Medicine*, *160*, 638–644. doi:10.1001/archpedi.160.6.638.

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. doi:10.1007/s11121-005-5553-y.

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. doi:10.1037/0012-1649.41.3.479.

Gollwitzer, M., Eisenbach, K., Atria, M., Strohmeier, D., & Banse, R. (2006). Evaluation of aggression-reducing effects of the "Viennese Social Competence Training." *Swiss Journal of Psychology*, *65*, 125–135. doi:10.1024/1421-0185.65.2.125.

Greenberg, M. T., Domitrovich, C. E., Graczyk, P. A., & Zins, J. E. (2005). *The study of implementation in school-based preventive interventions: Theory, research, and practice* (Vol. 3). Rockville, MD: Center for Mental Health Services, Substance Abuse and Mental Health Services Administration.

Hall, G. E., & Hord, S. M. (2006). *Implementing change: Patterns, principles, and potholes* (2nd ed.). New York: Pearson.

Heydenberk, R. A., Heydenberk, W. R., & Tzenova, V. (2006). Conflict resolution and bully prevention: Skills for school success. *Conflict Resolution Quarterly*, *24*, 55–69. doi:10.1002/crq.157.

Hunt, C. (2007). The effect of an education program on attitudes and beliefs about bullying and bullying behaviour in junior secondary school students. *Child and Adolescent Mental Health*, *12*, 21–26. doi:10.1111/j.1475-3588.2006.00417.x.

Jennifer, D., & Shaughnessy, J. (2005). Promoting non-violence in schools: The role of cultural, organisational and managerial factors. *Educational and Child Psychology*, *22*, 58–66.

Kim, J. (2006). The effect of a bullying prevention program on responsibility and victimization of bullied children in Korea. *International Journal of Reality Therapy*, *26*, 4–8.

Kyler, S. J., Bumbarger, B. K., & Greenberg, M. T. (2005). *Technical assistance fact sheet on evidence-based programs*. University Park, PA: Prevention Research Center for the Promotion of Human Development. Retrieved Oct. 1, 2007 from http://www.prevention.psu.edu/pubs/documents/EBP_factsheet.pdf.

Leadbeater, B., & Hoglund, W. (2006). Changing the social contexts of peer victimization. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *15*, 21–26.

Leadbeater, B., Hoglund, W., & Wood, T. (2003). Changing contexts? The effects of a primary prevention program on classroom levels of peer relational and physical victimization. *Journal of Community Psychology*, *31*, 397–418. doi:10.1002/jcop.10057.

Leff, S. S., Power, T. J., & Goldstein, A. B. (2004). Outcome measures to assess the effectiveness of bullying prevention programs in schools. In D. L. Espelage & S. M. Swearer (Eds.), *Bullying in American schools: A social-ecological perspective on prevention and intervention* (pp. 269–293). Mahwah, NJ: Lawrence Erlbaum.

Limber, S. P., & Small, M. A. (2003). State laws and policies to address bullying in schools. *School Psychology Review*, *32*, 445–455.

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

McLaughlin, L., Laux, J. M., & Pescara-Kovach, L. (2006). Using multimedia to reduce bullying and victimization in third-grade urban schools. *Professional School Counseling*, *10*, 153–160.

Menesini, E., Codecasa, E., Benelli, B., & Cowie, H. (2003). Enhancing children's responsibility to take action against bullying: Evaluation of a befriending intervention in Italian middle schools. *Aggressive Behavior*, *29*, 10–14.

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. doi:10.1037/1045-3830.23.1.26.

Meyer, N., & Lesch, E. (2000). An analysis of the limitations of a behavioural programme for bullying boys from a subeconomic environment. *South African Journal of Child and Adolescent Mental Health*, *12*, 59–69.

Mihalic, S. (2007). *Matrix of programs*. Boulder, CO: Center for the Study of Prevention of Violence, Blueprints for Violence Prevention. Retrieved on September 24, 2007, from www.colorado.edu/cspv/blueprints.

Mooij, T. (2005). National campaign effects on secondary pupils' bullying and violence. *The British Journal of Educational Psychology*, *75*, 489–511. doi:10.1348/000709904X23727.

Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Marton, B., & Scheidt, P. (2001). Bullying behaviors among US youth: Prevalence and association with psychosocial adjustment. *Journal of the American Medical Association*, *285*, 2094–2100. doi:10.1001/jama.285.16.2094.

Naylor, P., & Cowie, H. (1999). The effectiveness of peer support systems in challenging school bullying: The perspectives and experiences of teachers and pupils. *Journal of Adolescence*, *22*, 467–479. doi:10.1006/jado.1999.0241.

Newman-Carlson, D., & Horne, A. M. (2004). Bully Busters: A psychoeducational intervention for reducing bullying behavior in middle school students. *Journal of Counseling and Development*, *82*, 259–267.

No Child Left Behind Act of 2001, P.L. 107-110, 115 Stat.1425 (2002).

Okumura, T. (2007). Sample size determination for hierarchical linear models considering uncertainty in parameter estimates. *Behaviormetrica*, *34*, 79–93. doi:10.2333/bhmk.34.79.

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

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

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

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.

O'Moore, A. M., & Minton, S. J. (2005). Evaluation of the effectiveness of an anti-bullying programme in primary schools. *Aggressive Behavior*, *31*, 609–622. doi:10.1002/ab.20098.

Orpinas, P., & Horne, A. M. (2006). Bullies: The problem and its impact. In P. Orpinas & A. M. Horne (Eds.), *Bullying prevention: Creating a positive school climate and developing social competence* (pp. 11–31). Washington, DC: American Psychological Association.

Orpinas, P., Home, A. M., & Staniszewski, D. (2003). School bullying: Changing the problem by changing the school. *School Psychology Review*, *32*, 431–444.

Ortega, R., Del Rey, R., & Mora-Merchan, J. A. (2004). SAVE model: An anti-bullying intervention in Spain. In P. K. Smith, D. Pepler, & K. Rigby (Eds.), *Bullying in schools: How successful can interventions be?* (pp. 167–185). New York: Cambridge University Press.

Patchin, J. W., & Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, *4*, 148–169. doi:10.1177/1541204006286288.

Pawson, R., & Tilley, N. (1998). Cook-book methods and disastrous recipes: A rejoinder to Farrington. *Evaluation*, *4*, 211–213. doi:10.1177/13563899822208545.

Pellegrini, A., & Bartini, M. (2000). An empirical comparison of methods of sampling aggression and victimization in school settings. *Journal of Educational Psychology*, *92*, 360–366. doi:10.1037/0022-0663.92.2.360.

Peterson, L., & Rigby, K. (1999). Countering bullying at an Australian secondary school with students as helpers. *Journal of Adolescence*, *22*, 481–492. doi:10.1006/jado.1999.0242.

Rahey, L., & Craig, W. M. (2002). Evaluation of an ecological program to reduce bullying in schools. *Canadian Journal of Counselling*, *36*, 281–296.

Roland, E. (2002). Bullying, depressive symptoms and suicidal thoughts. *Educational Research*, *4*, 55–67. doi:10.1080/00131880110107351.

Rosenbluth, B., Whitaker, D. J., Sanchez, E., & Valle, L. A. (2004). The Expect Respect project: Preventing bullying and sexual harassment in US elementary schools. In P. K. Smith, D. Pepler, & K. Rigby (Eds.), *Bullying in schools: How successful can interventions be?* (pp. 211–233). New York: Cambridge University Press.

Salmivalli, C. (2001). Peer-led intervention campaign against school bullying: Who considered it useful, who benefited? *Educational Research*, *43*, 263–278. doi:10.1080/00131880110081035.

Salmivalli, C., Kaukiainen, A., Voeten, M., & Sinisammal, M. (2004). Targeting the group as a whole: The Finnish anti-bullying intervention. In P. K. Smith, D. Pepler, & K. Rigby (Eds.), *Bullying in schools: How successful can interventions be?* (pp. 251–273). New York: Cambridge University Press.

Salmivalli, C., Kaukiainen, A., & Voeten, M. (2005). Anti-bullying intervention: Implementation and outcome. *The British Journal of Educational Psychology*, *75*, 465–487. doi:10.1348/000709905X26011.

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

Smith, J. D., Schneider, B. H., Smith, P. K., & Ananiadou, K. (2004). The effectiveness of whole-school anti-bullying programs: A synthesis of evaluation research. *School Psychology Review*, *33*, 548–561.

Smith, J. D., Ryan, W., & Cousins, J. B. (2007). Antibullying programs: A survey of evaluation activities in public schools. *Studies in Educational Evaluation*, *33*, 120–134. doi:10.1016/j.stueduc.2007.04.002.

Standards of Evidence Committee. (2004). *Standards of evidence: Criteria for efficacy, effectiveness and dissemination*. Retrieved September 5, 2008 from http://www.jstor.org.proxy.bib.uottawa.ca/stable/pdfplus/1049953.pdf.

Stevens, V., Van Oost, P., & De Bourdeaudhuij, I. (2000). The effects of an anti-bullying intervention programme on peers' attitudes and behaviour. *Journal of Adolescence*, *23*, 21–34. doi:10.1006/jado.1999.0296.

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. doi:10.1016/S0022-4405(01)00073-5.

Stevens, V., Van Oost, P., & De Bourdeaudhuij, I. (2004). Interventions against bullying in Flemish schools: Programme development and evaluation. In P. K. Smith, D. Pepler, & K. Rigby (Eds.), *Bullying in schools: How successful can interventions be?* (pp. 141–165). New York: Cambridge University Press.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 25–32. doi:10.3102/0013189X031003025.

Torgerson, C. J., & Torgerson, D. J. (2001). The need for randomized controlled trials in educational research. *British Journal of Educational Studies*, *49*, 316–328. doi:10.1111/1467-8527.t01-1-00178.

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. doi:10.1001/archpedi.161.1.78.

Whitaker, D. J., Rosenbluth, B., Valle, L. A., & Sanchez, E. (2004). Expect Respect: A school-based intervention to promote awareness and effective responses to bullying and sexual harassment. In D. L. Espelage & S. M. Swearer (Eds.), *Bullying in American schools: A social-ecological perspective on prevention and intervention* (pp. 327–350). Mahwah, NJ: Erlbaum.

Wolff, N. (2000). Using randomized controlled trials to evaluate socially complex services: Problems, challenges and recommendations. *The Journal of Mental Health Policy and Economics*, *3*, 97–109. doi:10.1002/1099-176X(200006)3:2<97::AID-MHP77>3.0.CO;2-S.