BULLYING PREVENTION PROGRAMS FOR CHILDREN & YOUTH

A Guide to Program Selection

1st Edition

John C. LeBlanc, MSc MD

Tanya Bilsbury, MSc & Ashley Chisholm, BSc

Bullying Prevention Programs for Children & Youth: A Guide to Program Selection. – 1st ed.
ISBN (PDF): 978-0-9940205-1-2
First printing 14 March 2016
Last updated 19 October 2016

Published by CPSC Atlantic
c/o Dr. John LeBlanc
L5117, IWK Health Centre
5980 University Avenue
Halifax, NS, B3K 6R8 Canada
John.LeBlanc@dal.ca
www.cpscatlantic.org

About the Authors

John LeBlanc, MSc MD, is an Associate Professor of Pediatrics, Psychiatry, and Community Health & Epidemiology at Dalhousie University and a Staff Pediatrician at the IWK Health Centre. He is the director of CPSC Atlantic.

Tanya Bilsbury, BSc MSc, is a research associate at CPSC Atlantic and MPA candidate with the School of Public Administration.

Ashley Chisholm, BSc, is a research assistant at CPSC Atlantic and an MSc candidate with the Dalhousie Department of Community Health & Epidemiology

Acknowledgements

We thank the members of Canadian Prevention Science Cluster (CPSC) Atlantic for their partnership, collaboration, feedback and support, and Rebecca Clements for her foundational work. CPSC Atlantic included representatives from the Nova Scotia Department of Education and Early Childhood Education, the Department of Health and Wellness, and Dalhousie University. The CPSC (2009 – 2016) and this research was supported by the Social Sciences and Humanities Research Council of Canada.

DALHOUSIE UNIVERSITY

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Table of Contents

Executive Summary

Schools implement bullying prevention programs without clearly knowing whether a program will provide good value for their limited human and capital resources. This "*Bullying Prevention Programs for Children and Youth Toolkit"* has been created to help schools and other organizations working with youth evaluate bullying programs with respect to effectiveness, cost, and 'fit' with their particular context.

We reviewed bullying prevention programs used in some Nova Scotia schools as well as others identified in a search of scientific literature. Reviewed programs had at least one controlled study published in a peer-reviewed scientific journal, were focussed on reducing bullying or aggression, and were 'universal' in scope, with at least some program features addressing schools and organizations as a whole. We evaluated the evidence for program effects objectively and transparently using a rigorous structured approach called GRADE. GRADE allows only four recommendations for program implementation: 1) Strong recommendation for a program; 2) Weak (also called Provisional) recommendation for a program; 3) Weak recommendation against a program; 4) Strong recommendation against a program.

Only one of the reviewed programs received a recommendation (weak) for implementation: the WITS program. WITS delivered relatively strong effects at low cost and showed long-term benefits. It also has components that promote healthy relationships. All other programs received weak recommendations against implementation, largely because they were resource-intensive yet delivered little or no reduction in bullying or victimization, or were not feasible in a North American environment.

Most studies of bullying interventions are conducted in a way that casts doubt on their conclusions. Therefore, no program received a strong recommendation. A weak recommendation should not prevent a program from being used in a particular setting. However, it implies that the impact of a program in a new setting should be carefully measured to ensure that precious resources are not being wasted. The insufficient evidence for most programs, the high direct and opportunity costs that come with implementation, and the need to implement for several years before benefits accrue should compel organizations to carefully assess whether or not a program works in their setting.

There are many possible reasons why most bullying prevention programs show limited effectiveness. First, research studies typically only examine a small number of students across one or a few schools, when much larger samples may be required to detect effects reliably. Second, bullying is an intentional act, and structured programs may not address the internal and external factors leading some people to bully. Third, bullying prevention interventions are complex with different roles for students, teachers, other school staff, families and communities. It is difficult for complex and dynamic organizations like schools to implement programs in the exact manner intended by program creators, and to some extent, such 'fidelity' is at odds with the role of teachers as professionals trained to design and modify their own curricula to meet specific goals. Fourth, there are many factors that affect bullying (e.g., school climate, the quality of relationships within the school, and the presence or absence of a dedicated anti-bullying champion), and a program may not address all the important causes.

The authors believe that the priority for most organizations is to focus on developing assets in youth such as social and emotional skills, building positive organizational climates and promoting healthy relationships. There *may* be a place for a program focused primarily on bullying prevention once those priorities are addressed. Note that there will always be a place for strategies that target high risk individuals or groups as no program applied to an entire school or organization will prevent all individuals from bullying others.

Summary of Main Findings

Program	Recommendation	Comment
Dare to Care (Tested in grades 4 – 6)	**Against**: Weak	**Low quality evidence**. A small observed reduction in peer-reported aggression did not adjust for improvement in the control group. The program has changed since the evaluation and the curriculum-based component tested in the evaluation is no longer part of the program.
Friendly Schools and Families (Tested in grades 4 – 6)	**Against**: Weak	**Moderate quality evidence**. **Small consistent effects** on peer-reported victimization but effects on self-reported victimization and perpetration were **inconsistent** (occurring in 1 of 3 years) or null. The 2-day training requires travel to Australia.
KiVa (Tested in grades 1 – 7)	**Against**: Weak	**High quality evidence**. **Small or very small effects** in elementary students and even **fewer benefits** in middle school students. A widely-used program in Finland with ongoing evaluation in the UK and USA. Results are not available for the English version of the program, which is not currently available in North America.
Olweus Bullying Prevention Program (Tested in grades K – 12)	**Against**: Weak	**Low quality evidence**. The program is **very costly** and provides **very small benefits** at best. Available evidence suggests that the program holds more promise with high school students
Steps to Respect (Tested in grades 3 – 6)	**Against**: Weak	**Moderate to high quality evidence**. Offers relatively stronger effects at a lower price than other programs. However, small benefits observed by staff or researchers are not corroborated by student reports. This program is only available while supplies last. In its place, Second Step offers a condensed 'Bullying Prevention Unit' (K-5), which has not been evaluated.
Second Step: Student Success Through Prevention (Tested in grade 6)	**Against**: Weak	Only evaluated in **one study**, with grade 6 students only. The program had **no effect** on most aggression outcomes, all of which increased in prevalence from pre-test to post-test. However, students who received the program experienced a **smaller increase** in physical aggression than the control group.
WITS Primary Program (Tested in grades 1 – 3)	**For**: Weak	This program has shown **relatively stronger results** than other programs at a **lower cost**. There were small **beneficial effects** on physical and relational victimization observed three years after program implementation, and **a moderate reduction in physical aggression** at the classroom-level as reported by students.

Overview of the Program Guide

We use this definition of bullying by Dr. D. Olweus: "aggressive behavior that is intentional and that involves an imbalance of power, repeatedly and over time"[1]. Bullying through the use of computers or other electronic devices is defined as cyberbullying[2]. Direct bullying includes verbal bullying with derogatory names or comments, and physical bullying with shoving or hitting. Indirect bullying involves more subtle acts such as social exclusion and rumour-spreading[1]. In this toolkit we distinguish between 'bullying' and the broader construct of 'aggression.' Aggression includes bullying but doesn't necessarily require intent, power imbalance, or repetition. This distinction is not always maintained in studies that assess bullying prevention programs.

Bullying can lead to long-term consequences such as depression and suicidal behaviour for both victims and perpetrators[3–6]. Witnessing bullying is also associated with higher rates of depression, anxiety, and substance use[7]. Many organizations working with children and youth have programs and activities to prevent bullying. Some programs may have positive effects on reducing bullying, whereas others have effects that are very small or non-existent. Appropriate program selection should consider costs, since scarce money and staff time are required to implement the program properly.

Our toolkits are designed to give organizations an efficient way to select effective evidence-based programs; we also provide easy access to the evidence behind our decisions. This toolkit summarizes the evidence behind commonly-used bullying prevention programs and recommends *for* or *against* the use of these programs based on the quality of evidence for beneficial effects, and the cost of the program relative to the size of the effect.

The toolkit uses a structured approach to critically assess scientific evidence called GRADE (see p. 9) in order to evaluate the evidence as objectively and transparently as possible. The toolkit presents a summary of the evidence and recommendations. The CPSC Atlantic website (www.cpscatlantic.org) provides access to the detailed worksheets describing the study information upon which the recommendations are based.

Bullying Prevention Programs for Children and Youth is the second guide arising from the work done by the Canadian Prevention Science Cluster (CPSC Atlantic). It is targeted to schools and other front-line organizations that directly work with children and youth. It follows a similar format to the *Social and Emotional Learning Programs for Schools* toolkit[8] (May 2013), a review of programs that claim to enhance social and emotional skills, and which identified five evidence-based effective and feasible programs that did so.

What are the Distinguishing Features of this Program Guide?

There are several easily accessible compendia of bullying prevention programs such as the Public Health Association of Canada's Canadian Best Practices Portal (CBPP)[i], the National Registry of Evidence-based Programs and Practices (NREPP) from the U.S. Substance Abuse and Mental Health Services Administration (SAMHSA)[ii], the Blueprints for Healthy Youth Development (previously known as the Blueprints for Violence Prevention)[iii] and the Collaborative for Academic, Social, and Emotional Learning (CASEL)[iv]. These compendia are useful and most are excellent. They have limitations for busy professionals:

1) All compendia assess and report the quality of evidence for programs but some don't tell the reader how they assessed quality (e.g., CBPP).
2) Each has developed a unique rating system that helps readers assess the quality of the evidence and the strength of the recommendation for or against implementation. One in particular (CBPP) does not have any negative ratings.
3) Summary of outcomes varies tremendously from restatements of the researchers' published papers (CBPP) to effect sizes (NREPP) to a narrative description with directions but no numbers, e.g., higher or lower (CASEL & Blueprints).
4) As specified by the GRADE approach, this toolkit focuses directly on evidence that a program reduces the occurrence of bullying or aggression, and does not extrapolate to bullying published evidence for other beneficial outcomes such as improved school climate or social and emotional skills. Some compendia (CBPP, CASEL) provide an overall recommendation for a program even when not all relevant outcome measures have been evaluated.
5) Evidence-based programs may well have a positive impact on bullying and aggression, but perhaps very small effects at high personnel and financial cost. SAMSHA comments on costs for a US context and below, we make a GRADE recommendation based on benefits versus risks and costs.
6) Compendia vary in presenting programs in a form that is immediately useful to policymakers, e.g., the brief assessment of programs, expected outcomes, quality of evidence, feasibility, and both human and capital costs. This is a requirement of the GRADE approach.

[i] www.cbpp-pcpe.phac-aspc.gc.ca
[ii] www.nrepp.samhsa.gov
[iii] http://www.blueprintsprograms.com
[iv] www.casel.org

Program Selection

We used three criteria to select programs for this toolkit:

1) The program is designed for implementation at a group level such as a school, or classroom
2) The program focuses on the reduction of bullying or aggression
3) At least one study published in a peer-reviewed scholarly journal assessed the effect of the program on bullying or aggression
4) At least one study compared participants who experienced the program to participants who did not
5) The study focuses on at least one quantitative measure of bullying or aggression.

A comparison (or control) group is a critical feature of good experimental design and is essential to assess measures that change naturally over time or as children age. Sometimes bullying increases (i.e., in middle school) or decreases (i.e., in high school) simply because children are getting older, growing into, and later out of, different kinds of behaviours. For this reason, we excluded studies that used a simple pre-intervention vs. post-intervention design with no control group, also called a 'before-after' design. For more information about why control groups are essential, please see the appendix.

There is no single accepted measure for bullying or for aggression. The focus of a measure might be on overall bullying, or it might focus on specific forms such as physical bullying. These outcomes may be self-reported, peer-reported, or observer-reported (e.g., by teachers or researchers). Different observers typically don't agree[10,11] because not all observers see the same activities, and observers disagree as to what constitutes bullying. Some types of bullying, such as rumour-spreading or cyberbullying, can be hard for third-party observers to detect. Peer and teacher reports of bullying are more consistently associated with school disciplinary infractions[10] than self-reports, which are better at predicting personal, internal outcomes such as low self-worth and anxiety[12]. We assessed all reported outcomes related to bullying or aggression because they represent diverse and complementary perspectives and no single measure captures the phenomena[12]. We would have preferred to restrict our analysis to studies focussing on bullying alone but that would have greatly reduced the number of studies available for our analysis.

Evaluating the Strength of Evidence: Statistical Significance and Effect Size

This toolkit summarizes the effects reported by program evaluations and comments on their impact and statistical significance. Briefly, a 'statistically significant' effect is one that would only have occurred by chance 1 in 20 times or less if the program under study did not actually do anything. Therefore, if a study shows an effect that should only occur by chance 1 in 20 times, it would be reasonable to consider an alternative explanation such as the effect being due to the program under study and not to chance. For a variety of reasons, high quality evidence for school-based programs that try to change human behaviour is rare. These reasons include the difficulty of conducting research in school settings, the expense of including enough groups to show that an effect exists, the difficulty of measuring bullying and aggression, and the inherent difficulty of changing how humans behave.

The size of a program effect can be stated in many different ways that can be confusing and make it difficult to compare studies. These include differences between groups stated as percentages, odds ratios, and Cohen's *d*. We prefer *absolute differences* in percentage because this doesn't overstate the impact of programs[v]. It can help decision-makers perceive a clear gain or loss for the amount invested in a program and percentage reduction is easy to understand for the public and professionals alike. For example, if a program claims to reduce bullying by a *relative difference* of 50%, a principal of a school of 100 students that has 20 students who bully, can expect, after running this program, a reduction to 10 students who bully. However, under the same conditions, a school of 100 students where the number of students who bully is 6 can only expect this to be reduced to 3 students after running the program. It is against these *absolute* gains of 10 fewer students who bully in the first scenario or 3 fewer in the second example that a school must balance the costs in money and time for this example program.

Table 1 (p. 7), and the more complete Table 3 (p. 51) in the Appendix, are guides to interpret and compare different measures. Cohen's *d* and the odds ratio (OR) are the most common, so we'll explain them here. The odds ratio (OR) is a measure of association between an exposure (e.g., program or no program) and an outcome (e.g., bullied or not bullied). The OR represents the odds that an outcome will occur under a particular condition or exposure (e.g., likelihood of being a victim in a school that has received a bullying prevention program), compared to the odds of the outcome occurring in the absence of that exposure (i.e., being a victim in a school that has not received a bullying prevention program). An OR of 1 indicates that there was no program effect (i.e., the odds of being a victim with a bullying prevention program is the same as with no program). An OR above 1 indicates that the exposure is associated with a higher likelihood of experiencing the outcome, whereas an OR below 1 means that the exposure is associated with a lower likelihood of experiencing the outcome. ORs hovering near 1 (e.g., 0.90 or 1.2) would be considered very small effects. Please see the Appendix to learn more about statistical significance and effect size.

[v] This is the same as 'absolute risk reduction' used in Public Health and Medicine and is closely related to the concept of 'number needed to treat' (NNT), or the number of people that must be exposed to an intervention for each individual who benefits.

Table 1. A guide to interpreting and comparing effect sizes

Effect Size	Very Small	Small	Moderate	Large
Description	Little or no noticeable difference; e.g., average heights of girls aged 15 & 15½, reducing bullying from 60% to 55%.	The difference is apparent but not immediately noticeable; e.g., average heights of girls aged 15 and 16, or reducing bullying from 60% to 50%.	The difference is 'visible to the naked eye of a careful observer'; e.g., average heights of girls aged 14 and 18, or reducing bullying from 60% to 30%.	The difference is obvious and 'grossly perceptible'; e.g., average heights of 10 vs. 18 year old girls, or reducing bullying from 60% to 10%
Frequency Difference *Absolute*	<10%	10%	30%	≥50%
Odds Ratio	<1.5	1.5	3.5	≥9
Standardized Difference *Cohen's* d	<0.2	0.2	0.5	≥0.8

Adapted from Cohen[17] *& Watson*[18]*.* See Appendix for an expanded table of effect sizes

One must use both statistical significance and the size of the intervention impact on an outcome (effect size) to assess a study since an effect may be statistically significant but small and unimportant. This is an issue with studies that included many more research participants, because a study's ability to detect effects increases with the number of subjects (i.e., it's much easier for real but unimportant effects to be statistically significant as study sample sizes increase). Accordingly, studies with very large samples can report 'statistically significant' effects that have no practical value because those who received the intervention are only marginally different from those who didn't. Consider a very large-scale evaluation of the Olweus Bullying Prevention Program (OBPP) in Pennsylvania[13], which included over 10,000 elementary school students.

This study examined, among other outcomes, bullying perpetration and victimization in students of different ages and at different levels of program implementation: considering results for perpetration and victimization only, there a total of 18

outcomes. Eight of these outcomes were not reported and three of the effects were not statistically significant. One GRADE criterion is that all pertinent outcomes should be reported so that the reader can evaluate the overall significance of the results and not just the results chosen by the authors. The remaining six outcomes all showed benefits in the 'very small' category of effect size. The smallest of these effects was observed with elementary school students in the high implementation condition after one year (the effect was lost in the second year); the size of this effect is not described in the paper, but it can be calculated from one of the figures, which shows an absolute reduction in bullying by *one third of one percent*, from about 7.5% to 7.2%. The largest of these effects was observed with high school students in the high implementation condition after two years (following a smaller beneficial effect in the first year). It is an absolute reduction of about 5.5%, from about 14.5% to 9%. This is certainly more than 0.3% reported above but both are difficult to interpret without knowing all reported results.

Another problem arises when more than one outcome is measured; it becomes more likely that at least one outcome will be significant *even if there were no impact of a program on bullying or victimization.* Let's say that the OBBP has no impact and a study had a single measure of bullying, say, a 10% reduction in the rate of bullying before and after the intervention, and that the study investigators set the p-value at the usual 5% level. This means that there is a 5% risk that they will find a significant result when the program doesn't do anything at all. Now say that there are 18 outcomes as in the previous results and that each outcome is measuring something different (i.e., they are statistically independent). Now the risk of having at least one statistically significant result when the program does nothing goes up to 60%[vi]! This is called the 'multiple comparisons' problem.

Small effect sizes such as 0.3% or 5.5% are also sensitive to small changes in study execution or in how bullying is measured, factors that could sway the effect in either direction. Finally, even if one accepts these changes as being true and not affected by bias or chance, one must question whether such small reductions in bullying are worth the overall costs of the program: a start-up cost of $9,000 CAD price based on bringing in an external trainer in addition to purchasing curriculum materials (see p. 25).

Given the scarce resources available to schools and youth organizations, decisions about the costs versus benefits of a program should involve a careful judgement by administrators. For example, two evaluations of the OBPP[14,15] detected no overall effect of the program on bullying, yet the authors concluded that schools should nevertheless make a long-term commitment to the program and establish committed funding mechanisms to support it. One paper[14] stated: "we encourage schools not to stop implementing the OBPP. One reason is that this program is the only available bullying prevention program that is comprehensive and encompasses a whole school approach". Some advocates claim that having a bullying prevention program is always better than not having one, and these interventions should be based on comprehensive school and community-wide approaches[16]. However, someone who must use scarce resources in as effective a manner as possible to help children and youth should ask what other uses there could be for the time and money invested in a bullying prevention program that has such a small impact. These could include,

[vi] From binomial probability: $1 - 0.95^{18}$ (or one minus probability that all 18 outcomes are not statistically significant)

for example, programs that develop social and emotional skills or improve school climate, more resource teachers or support personnel, more extracurricular activities, or new playground equipment to encourage fitness and social interaction.

We describe bullying prevention programs as ‘universal’, ‘targeted’, or ‘indicated’, also called 'Tier 1', 'Tier 2' and 'Tier 3' interventions respectively. Universal programs target the entire school population and ideally families of students in the school. This typically includes all staff, whether or not they have contact with students, since all staff contribute to the social health (school climate) of a school. Targeted programs reach out to a specific group, such as children with self-regulation or anger management issues or with poor social skills. Indicated programs target the particular student and possibly his or her family. These may be students who bully that don't respond to universal or targeted interventions or students who have serious mental or emotional issues as a result of witnessing or experiencing bullying.

This toolkit attempts to provide a critical, independent review of the effects of bullying prevention programs and makes explicit the rationale by which recommendations were made. The toolkit also attempts to summarize costs and outcomes for decision makers who must decide whether or not to implement a bullying prevention program in a particular setting. It also provides an avenue to a more detailed analysis of programs if the reader desires to explore the research more deeply.

Evaluating the Strength of Evidence: Experimental Design and the GRADE Approach

Our recommendations for or against programs are based on a structured and systematic review of published evidence. To arrive at our recommendations, we used the Grading of Recommendations, Assessments, Development and Evaluation (GRADE) approach, which is an outcome-based, rigorous, and transparent[19] system for assessing evidence. The GRADE approach is endorsed by leading scientific organizations including the World Health Organization, the Cochrane Collaboration, the British Medical Journal, the Canadian Task Force on Preventive Health Care and the American College of Physicians[20]. This was implemented as follows: 1) we selected bullying prevention programs according to the criteria listed on page 5 and searched for all experimental or quasi-experimental studies that assessed that program; 2) two to four reviewers independently evaluated each study publication and abstracted information about the study design and results; 3) reviewers assessed the quality of evidence for each of the main outcomes; and 4) two reviewers independently made a strong or weak recommendation for or against program implementation based on evidence quality, effect size, and program cost. If the agreements did not agree, they met with a third reviewer and arrived at a consensus.

The GRADE approach puts a strong emphasis on the quality of evidence; for example, it favours randomized control trials (RCTs) over observational studies. We modified the GRADE approach to give more emphasis to quasi-experimental (QE) studies (*see glossary in Appendix*), frequently used in school-based research. These studies are more susceptible to bias than RCTs but offer more opportunity to ensure intervention and control groups are comparable at baseline than do observational studies, where researchers simply assess the 'natural experiments' of schools deciding whether or not to implement a program. We therefore assigned QE studies a study design score that was between RCTs and observational studies.

There are four possible GRADE recommendations. A 'strong' recommendation *in favour* of a program indicates that we are confident that the program has benefits that consistently outweigh the risks or costs if implemented as designed. A 'strong' recommendation *against* a program indicates that we are confident that the program creates harm, has no benefit, or has small and unimportant benefits that are outweighed by the risks or costs. 'Weak' recommendations for or against a program indicate that the quality or quantity of evidence is insufficient to make firm conclusions based on available studies. However, the program may be effective (or not) under certain conditions. Note that a weak recommendation for a program does not indicate that the program is less effective than one with a strong recommendation (although it may well be); it simply indicates that ther is too little research available at this time to make that judgement. 'Low quality evidence' means that the effect of the program on bullying/aggression in the cited studies is very uncertain because there were serious limitations to the research. 'Moderate quality evidence' means that the evidence consists of only a few quality quasi-experimental studies or lower-quality randomized trials. Finally, 'high quality evidence' indicates that the outcome of the program is supported by one or more well-done RCTs.

Table 2: Bullying Prevention Program Effects and Recommendations

LEGEND		SR	Student-reported	PR	Peer-reported	AR	Adult-reported	RR	Researcher-reported
	Small benefit		Very small benefit		Not statistically significant/ Inconsistent effect				Harmful effect

Program	Gr.	Victimization	Perpetration	Recommend	Comments
Dare to Care Est. cost[vii]: $4800	4-6	**Aggression, SR:** No effect	**Aggression, PR:** Small effect **pre/post comparison**	**Against**: Weak	**Low quality evidence**. The **small observed benefit** did not adjust for improvement in the control group. The program has changed since the evaluation. The curriculum-based component tested in the evaluation is no longer part of the program.
Friendly Schools & Families Est. cost: $6400	4-6	**Bullying, SR :** Small effect for only one year	**Bullying, SR:** No effect	**Against**: Weak	**Moderate quality evidence**. **Small consistent effects** on peer-reported victimization but effects on self-reported victimization and perpetration were **inconsistent** (occurring in 1 of 3 years) or null. 2-day training requires travel to Australia.
		Frequent Bullying, SR : Small effect or only one year	**Frequent Bullying, SR:** No effect		
		Bullying, PR: Very small or small effects			
KiVa Est. cost: *Not available*	1-3 & 4-6	**Bullying, SR:** Very small effects	**Bullying, SR:** Very small effects	**Against**: Weak	**Low quality evidence**. **Very small effects** in elementary students and even **fewer benefits** in middle school students. A widely-used program in Finland with ongoing evaluation in the UK and USA. Results are not available for the English version of the program, which is not currently available in North America.
		Bullying, PR: Small effects	**Bullying, PR:** No effect		
		Cyberbullying, SR: Very small effect	**Cyberbullying, SR:** No overall effect		
	7-9	**Bullying, SR:** No effects	**Bullying, SR:** No effects		
		Bullying, PR: Very small effect	**Bullying, PR:** No overall effect		
		Cyberbullying, SR: Very small effect	**Cyberbullying, SR:** No effect		

[vii] All costs are in Canadian dollars as of 2015. Estimated costs are for the first year of implementation based on the full price of new materials and, when in-person training is required, the approximate cost of travel, food and accommodation) to and from Nova Scotia.

Program	Gr.	Victimization	Perpetration	Recommend	Comments
Olweus Bullying Prevention Program Est. cost: $8730	Elem	**Bullying, SR:** No effect	**Bullying, SR:** Very small effects; one not maintained	**Against**: Weak	**Low quality evidence**. Very small benefits largely restricted to secondary school students; program costs are high in terms of both money and staff time. Major benefit that is often reported and found only in the original studies done by Olweus himself. Benefits of that magnitude have not been reported.
	Mid	**Bullying, SR:** No overall effects or very small effects	**Bullying, SR:** No overall effects or very small effect		
		Aggression, SR: No overall effect	**Bullying, PR:** Harmful effect		
	High	**Bullying, SR:** Very small effects	**Bullying, SR:** Very small effects		
Steps to Respect Est. cost: $700 *While supplies last*	3-6	**Aggression, SR:** No effects	**Aggression, SR:** No effects	**Against**: Weak	**Moderate quality evidence**. Offers relatively **stronger effects** at a **lower price** than other programs. However, **small benefits** observed by staff or researchers are not corroborated by student reports. This program is only available while supplies last. In its place, Second Step offers a condensed 'Bullying Prevention Unit' (K-5), which has not been evaluated.
		Bullying, RR: No effect	**Verbal/Relational Aggression, AR**: No effect		
		School Aggression-Related Problems, SR: No effect	**Physical Aggression, TR:** Small effect		
		School Aggression-Related Problems, AR: Small effect	**Aggression, RR:** No effect		
			Bullying, RR: Small effect		
Second Step: Student Success Through Prevention Est cost. $1000	6	**Verbal Aggression, SR** *Homophobic name-calling* No effect	**Verbal Aggression, SR** No effect	**Against**: Weak	**Low quality of evidence.** Only evaluated in **one study**, with grade 6 students only. The program had **no effect** on most aggression outcomes, all of which increased in prevalence from pre-test to post-test. However, students who received the program experienced a **smaller increase** in physical aggression than the control group.
			Relational Aggression, SR: No effect		
		Aggression, SR: No effect	**Physical Aggression, SR:** Small effect		

Program	Gr.	Victimization	Perpetration	Recommend	Comments
		Sexual Aggression, SR: No effect	**Sexual Aggression, SR:** No effect		
WITS Primary Program Est. cost. $635	1-3	**Individual-level Relational Aggression, SR:** Very small or small effect **Classroom-level Relational Aggression, SR:** Small effect **Individual-level Physical Aggression, SR :** Very small or small effect **Classroom-level Physical Aggression, SR:** Small or moderate effects	**Physical Perpetration, AR :** No effect	**For:** Weak	**Moderate quality of evidence.** This program was designed and tested in Canada, and has shown **relatively stronger results** than other programs at a **lower cost**. There were small **beneficial effects** on physical and relational victimization observed three years after program implementation, and a **moderate reduction** in **physical aggression** at the classroom-level as reported by students.

Note: The 'grade' column refers to the grades in which the program was *tested*, not the grade range to which the program applies. For example, Dare to Care is available for grades K-12, but has only been tested in grade 4-6 students. Thus, the only available evidence for the program apples to grade 4-6 students only.

Dare to Care

www.daretocare.ca

General Description & Outcomes	Program Resources	Program Specifics
General Descriptions: • Aims to transform the 'silent majority' of bystanders to a 'caring majority' by increasing their awareness of bullying and their ability to effectively respond to incidents. • The program advertises that it 'eliminates' bullying, 'bully-proofs' your school and creates a 'bully-free' environment. • Information sessions with parents, teachers, and students to encourage open dialogue with common language for all community members • School community members (students, parents, & school staff members) collaborate to develop discipline policy on bullying with a focus on reparation rather than punishment. • Classroom curriculum about the nature of bullying and strategies to avoid victimization involves role-plays, artwork, books, videos, and skits[viii]. • Emphasizes individual or group counselling for bullies & victims *Primary Outcomes:* • Aggression: Self-reported victimization • Aggression: Peer-reported perpetration	*Duration of Program:* • 1 day professional development • 1 student development day (optional follow-up day for grade 5-9 students) • 2 hour parent session *Financial Resources:* • Teacher professional development day: $1650 (1 -3 schools); $2500 (3-10 schools); $3500 (≥ 10 schools) • Student day: $1250 • 'Take the time' follow-up day: $2500 (max. 120 students) • Parent sessions: $650 • Cost of the facilitator(s)' travel from Alberta, including mileage, airfare, accommodation, and $60/day meal allowance (est. $1250 for 2 nights) *Instructor:* • Trained facilitator	*Grade Range* Grades K to 12 Language English ☑ French ☐ Spanish ☐ *Focus* Universal ☑ Targeted ☐ Indicated ☑ *Scope* Whole school ☑ Classroom based ☐ Parent component ☑

[viii] The evaluation of the program reported that there was a classroom curriculum; however, this no longer appears to be part of the program, which is now centered on workshops. The manual '*Bully Proofing Your School*'[50] (1996, 2000) belongs to an earlier version of the program and is no longer listed as a component of the program[51]

Dare to Care Program Recommendation

Factors	Decision	Explanation
High or moderate quality evidence?	Yes ☐ No ☑	Research: • Very low quality evidence • One quasi-experimental study[21] with grade 4-6 students in Calgary, Canada. Comments: • Data were collected in 2004 • The study compared schools implementing the program for 3-months, 1 year, or 2 years. • The study had a low response rate and data were collected from a small number of students. • The pre/post differences observed in the program and control groups were compared without adjusting for baseline differences between the program and control groups. • There was no defined reference period (e.g., past year) for the prevalence measures • Self-reported perpetration was measured[22] but not reported • The evaluation included a curriculum-based component that is no longer part of the program
Certainty: *Benefits outweigh the downsides?*	Yes ☑ No ☐	Findings: **Aggression: Self-Reported Victimization** • No statistically significant reduction in victimization after 3 months of implementation compared to pre-test in the treatment or control group; no increase in the effect with longer implementation (up to 2 years). **Aggression: Peer-Reported Perpetration** • Use of the program was associated with a small (*d*=0.33) reduction after 3 months of implementation compared to pretest in the treatment group consisting of 3 points on a 30-point scale. There was a 1 point improvement in the control group that did not reach statistical significance. The effect increased over up to 2 years of implementation. Comments: • Intervention and control groups were not equivalent at baseline; the analysis did not adjust for this. The 3-point improvement in peer-reported perpetration for the intervention group is simply a pre/post comparison that does not adjust for the 1-point improvement in the control group; if such an adjustment were made, the treatment effect would be even smaller. • The measurement of experiencing lifetime aggression makes it difficult to show effects of a short-duration program. • No reported negative effects to implementation of Dare to Care

Resource Implications: *Benefits worth the costs?*	Yes ☐ No ☑	• Program costs are high and recurring. While there is no evidence of harm, there was no effect on self-reported victimization and only low-quality evidence of a small effect on witnessing bullying. Additionally, the current workshop-based program is not the same as the curriculum-based program evaluated in the study.
Summary and Recommendation:	**WEAK** recommendation **AGAINST** using 'Dare to Care' as a bullying-prevention program. The quantity and quality of research on the program is not sufficient.	

Friendly Schools & Families

www.friendlyschools.com.au

General Description & Outcomes	Program Resources	Program Specifics
General Descriptions: • A whole-school approach involving administrators, teachers, school staff, students, and parents. The goal is to create a cohesive community dedicated to supporting one another through school-based and home-based activities that involve all members of the school community. • Strengths-based approach focuses on creating positive health rather than eliminating risk factors. • Focuses on creating a positive school community and teaching social & emotional skills to students. • Formal lessons relate to emotions, social knowledge, and social skills. *Primary Outcomes:* • Bullying: Self- and peer-reported victimization • Bullying: Self-reported perpetration	*Duration of Program:* • Before initial implementation the fall semester is spent training staff and establishing school-wide policies. • 9 hours of specific in-class lessons (January-May). *Financial Resources:* • $660 for the complete package • $2200 tuition for two-day train-the-trainer session. Trainer must travel to Australia for two days, which considerably increases start-up costs (e.g., about $3000 for Halifax-Perth return flight, extra for hotel, taxi, meals, etc.) *Instructor:* • Teacher-led using available materials and manuals	*Grade Range* Grades K to 8 Language English ☑ French ☐ Spanish ☐ *Target Population* Universal ☑ Targeted ☐ Indicated ☐ *Program Components* Whole school ☑ Classroom based ☑ Parent component ☑

Friendly Schools & Families Recommendation

Factors	Decision	Explanation
High or Moderate Quality Evidence?	Yes ☑ No ☐	Research: • Moderate to high quality evidence • Two longitudinal randomized control trials (RCTs)[23,24] based in public schools in Perth, Western Australia Comments: • The program was evaluated in Australian samples 10-15 years ago (2000-2004) • The studies used a clustered randomized assignment, a large sample size, a careful definition of bullying, and an appropriate reference period of the past 3 months. • One study[23] followed grade 4 students until grade 6 and compared with a no-treatment group. • The other study[24] compared levels of implementation (low vs. high & moderate vs. high), and followed grade 4 students for 2 years and grade 6 students for 1 year). • Schools that had low implementation received a simplified version of the program manual with no practical detail, resources, or training, and were the closest approximation to a no-treatment group possible given wide exposure to the program. Schools that had moderate implementation received a comprehensive guide with full implementation and monitoring strategies for the whole school. Schools that had high implementation received training for teachers to engage parents and provided family activities. • The study assessing levels of implementation was limited by low compliance[24].
Certainty: *Benefits outweigh the risks?*	Yes ☑ No ☐	Findings: **Bullying: Self-Reported Victimization** • An RCT[23] found a small statistically-significant reduction in victimization (any vs. none) in year 1 (gr 4) of the intervention (OR= 1.49, *d*=0.22, r^2= 1.1%, an absolute difference between groups of 7% at post-test), but no statistically significant effects in year 2-3 (gr 5-6). When viewed as frequent victimization vs. low/no victimization, there were no statistically significant effects in years 1-2, but there was a small effect (OR=1.50, *d*=0.22, r^2=1.2%, an absolute difference between groups of 5% at post-test) in year 3. • An RCT[24] found that high implementation had a moderate effect compared to low implementation in year 1 for grade 4 students (OR= 1.76, *d*= 0.31, r^2= 2.4%; a small effect) and grade 6 students (OR= 1.54, d= 0.24, r^2= 1.4%; a small effect). There were no statistically significant differences in years 2-3

		or between the moderate and high implementation conditions[ix]. **Bullying: Self-Reported Perpetration** • There were no statistically significant effects on perpetration compared to no treatment.[23] • Higher levels of implementation were not associated [vi] with improved effects.[24] **Bullying: Peer-Reported Perpetration** • There were consistent statistically-significant improvements in bullying witnessed, which increased from a very small effect in year 1 when students in grade 4 (OR= 1.36, *d*=0.17, r^2=0.7%, an absolute difference between groups of 5% at post-test) to small effects in year 2/gr 5 (OR= 1.48, *d*=0.22, r^2=1.1%, an absolute difference between groups of 8% at post-test) and year 3/gr 6 (OR= 1.67, *d*=0.29, r^2=2.0%, an absolute difference between groups of 10% at post-test)[24]. Comments: • The program showed no effect on perpetration, small effects on victimization, and consistent effects on witnessing bullying that increased over time. • High implementation was associated with better outcomes for victimization than low implementation in year 1 only, but there was no difference between moderate and high levels of implementation, and no effects of implementation on perpetration. • The intervention had low compliance. In the first study[23], parents implemented a median of only 16.5% of family activities in the first two years of the study, while teachers covered a median of 75% of the classroom lessons. In the second study[24], 55% of program components were implemented at the classroom level, 63% at the school level, and 22% at the family level.
Resource implications: *Benefits worth the costs?*	Yes ☐ No ☑	• The program shows small but consistent effects on peer-reported bullying, however there were no effects on self-reported perpetration and only inconsistent effects on victimization. • The full cost of the program is prohibitive, and low implementation suggests that the program is too resource-intensive for a North American school setting.
Summary and Recommendation:	**Weak** recommendation **AGAINST** using Friendly Schools and Families as an evidence-based school bullying prevention program in North America. There is high quality evidence for a moderate reduction in victimization and a small reduction in bullying but the cost of the program would be very high given the initial dependence on Australian trainers.	

[ix] However, there were a number of marginally significant (p<0.10) effects supporting implementation of the program when implemented.

KiVa

www.kivaprogram.net

<table>
<tr><th>General Description & Outcomes</th><th>Program Resources</th><th>Program Specifics</th></tr>
<tr><td rowspan="2"><u>General Descriptions:</u>

• The program focuses on training bystanders to support victims rather than reinforce bullies, raising awareness of bullying as a group process, increasing empathy for victims, promoting bystanders' ability to help victims, and increasing the skills of targeted children.
• Universal actions include student lessons with discussion, group work, role-playing, and short films; a virtual environment with an anti-bullying computer game (elementary) an online forum (middle school) students; a parent guide; and reminder symbols (e.g., posters, vests) for staff and students.
• Indicated actions include individual or group discussions with bullies, victims, and prosocial classmates held by a KiVa team (three school staff members). Classroom teachers meet with potential supporters of the victim.
• Three school teams form a network that meets three times per year with a program representative.

<u>Primary Outcomes:</u>

• Bullying: Self and peer-reported victimization
• Bullying: Self and peer-reported perpetration
• Cyberbullying: Self-reported victimization & perpetration</td><td rowspan="2"><u>Duration of Program (Per Year):</u>

• 10 semi-structured lessons (20 hours total)
• Virtual environment with computer game/online forum
• 3 versions are available:
o Grades 1 to 3
o Grades 4 to 6
o Grades 7 to 9

<u>Financial Resources:</u>

• Training and resources for the English-version of the program are not yet available in North America.

<u>Instructor:</u>

• Teacher-led; teachers are given two full days of face-to-face training and access to manuals and materials.</td><td><u>Grade Range:</u>
Grades 1 to 9

<u>Language</u>
English ☑
French ☑
Spanish ☐</td></tr>
<tr><td><u>Target Population:</u>

Universal ☑
Targeted ☐
Indicated ☑

<u>Program Components:</u>

Whole school ☑
Classroom based ☑
Parent component ☑</td></tr>
</table>

Kiva Program Recommendation

Factors	Decision	Explanation
High or moderate quality evidence?	Yes ☐ No ☑	Research: • Low quality evidence • Two RCTs[25,26] and one quasi-experimental study[27] with two[x] secondary analyses[28] Comments: • Research conducted in Finland from 2007-2009 • The quasi-experimental study[27] used an age-cohort design, comparing post-test students who received the intervention to the previous year's students of the same grade who had not received the intervention. • All evaluations tested the Finnish-language version of the program in Finnish comprehensive schools. The English-language version of the program is being implemented in the UK, USA, and NZ but has not yet been formally evaluated. Training and resources are not presently available in North America. • Only students who were involved more than 2-3 times in a month were classified as victims. • The analysis used in the RCTs controlled for confounders (e.g., baseline scores, age, and gender). However, results were not adjusted and may be due in part to confounders (e.g., student characteristics that influence the study results but are not due to the intervention).
Certainty: Benefits outweigh the downsides?	Yes ☑ No ☐	Findings: **Bullying: Self-reported victimization:** *Grades 1-3:* • An RCT[26] showed a statistically-significant, but very small reduction in victimization in grade 2-3 students after 1 year of implementation. The effect varied by gender, with small positive effects for girls (girls in the control group were 1.6 times more likely to be victimized, *d*= 0.27), and mixed effects for boys[xi] . • The quasi-experimental study[27] showed very small effects for grade 1-3 students after 1 year of implementation (ORs showing higher victimization in the control group ranged from 1.17 to 1.23; *d*s = 0.09-0.11). Prevalence was 30% in the control group compared to 25% in the program

[x] We did not include one secondary analysis (Salmivalli et al., 2011[28]), which examined program effects on different types of bullying (e.g., physical, social) because the authors used a criterion of α= 0.20 to establish statistical significance, which is much higher than the standard value of α= 0.05.

[xi] Overall, there was a very small negative effect for boys (an increase in victimization of +0.44 on a 5-pt scale), but this effect was reversed to a change of -1.7 in classrooms with a majority of boys (over 65%).

		group, an absolute reduction of 5%. *Grades 4-6*: • An RCT[25] showed statistically-significant but very small (*d*=0.17) reduction in victimization, consisting of a sixth of a point on a 5-point scale. The unadjusted analysis showed higher victimization in the control group (OR=1.47), with prevalence values of 13% vs. 9% (absolute reduction = 4%). • A quasi-experiment[27] showed very small effects for grade 1-3 students after 1 year of implementation (ORs showing higher victimization in the control group ranged from 1.22 to 1.33, *d*s = 0.11-0.16). Prevalence was 16% in the control group compared to 13% in the program group, an absolute reduction of 3%. *Grades 7-9:* • In grade 8-9 students, an RCT[26] showed that one-year implementation of the program had no statistically-significant effect on self-reported victimization. • A quasi-experiment[27] also showed no statistically-significant effects in grade 7-9 students. **Cyberbullying: Self-reported victimization** *Grades 4-6 & 8-9:* • An RCT[29] showed that after 1 year of implementation, the program decreased average cyber-victimization by a quarter of a point on a 5-point scale. The odds ratio showed that unexposed students had 1.29 times the odds of being victimized, a very small effect size (*d*=0.14). **Bullying: Peer-reported victimization:** *Grades 4-6:* • An RCT[25] showed that one-year implementation of the program resulted in a small (*d*=0.33) statistically significant reduction in average victimization, consisting of a third of a point on a 5-point scale. *Grades 8-9:* • An RCT[26] showed that one-year implementation of the program resulted in a statistically-significant, but very small reduction in victimization, consisting of a tenth of a point on a 5-point scale; a very small effect (*d*=0.10) was found in grade 8 students only. **Bullying: Self-reported perpetration:** *Grades 1-3:* • An RCT[26] showed a very small statistically-significant reduction in victimization among grade 2-3 students, consisting of a third of a point on a 5-point scale. Students who didn't receive the intervention had 1.43 times the odds of being victimized (*d*= 0.197). • A quasi-experiment[27] showed very small effects for grade 1-3 students after 1 year of

		implementation (ORs showing higher victimization in the control group ranged from 1.15 to 1.30, *d*s = 0.08-0.15). Prevalence was 13.5% in the control group compared to 10.9% in the program group, an absolute reduction of 2.6%. *Grades 4-6:* • An RCT[25] showed that one-year implementation of the program resulted in a very small (*d*=0.10) but statistically significant reduction in average perpetration, consisting of a twelfth of a point on a 5-point scale. • A quasi-experiment[27] showed very small effects for grade 1-3 students after 1 year of implementation (ORs showing higher victimization in the control group ranged from 1.19 to 1.34, *d*s = 0.10-0.16). Prevalence was 10% in the control group compared to 8% in the program group, an absolute reduction of 2%. *Grades 7-9* • An RCT[26] and a quasi-experiment[27] showed that one-year implementation had no significant effects. **Cyberbullying: Self-reported perpetration** *Grades 4-6 & 8-9* • An RCT[29] showed that after 1 year, the program had no statistically-significant effect[xii] on cyberbullying. **Bullying: Peer-reported perpetration:** *Grades 4-6* • An RCT[25] showed no statistically significant change in scores after a year of implementation. *Grades 8-9* • An RCT[26] showed that one-year implementation had no statistically-significant overall effect. However, there was a statistically-significant but very small (*d*=0.11) reduction in perpetration for boys. Comments: • The grade 1-3 version of the program had very small effects on self-reported bullying • The grade 4-6 version of the program had very small effects on self-reported bullying and cyberbullying, a small or null effect on peer-reported bullying. • The grade 7-9 version of the program has little or no effect on primary outcomes • The research is limited by 'floor effects': the baseline rates of bullying were very low (e.g., mean

[xii] There was a marginally significant benefit; OR= 1.34 [very small]; 95% CI = 0.99-1.79. The small effect may be explained by chance.

		bullying values that were already near zero), which makes it difficult for an intervention to make further improvements. • While implementation in the quasi-experiment[27] was low (teachers had given 45% of lessons and 60-70% of classrooms used the virtual environment), effect sizes were consistent with the randomized trials, which had good implementation[26]
Resource implications: Benefits worth the costs?	Yes ☐ No ☑	• At this time, barriers to implementation are absolute (resources and training not available in North America). • The program may be too resource-intensive for schools to implement properly • Current evidence does not support the use of the grade 7-9 version
Summary and Recommendation:	**Weak** recommendation **AGAINST** using 'KiVa' as a bullying prevention program. While the effects of the elementary versions of the program show some promise, a recommendation would require a positive evaluation of the English version of the program.	

Olweus Bullying Prevention Program

www.kids-can.ca

General Description & Outcomes of Interest	Program Resources	Program Specifics
General Description: • The program is based on four *adult-centred* key activities within the school environment: 1) displaying warmth and a positive interest in students; 2) setting firm boundaries for unacceptable behaviour; 3) consistently enforcing consequences when rules are broken; and 4) functioning as positive role models and authority figures. • School-wide activities include training all staff members, creating a committee, posting school rules against bullying throughout the school, and surveying students annually. • Classroom activities include group discussion, class meetings, and role-play scenarios. • Targeted interventions include developing individualized strategies, following up with involved students, and engaging parents. • Parents are encouraged to become engaged in school and community activities as well as individual interventions • Community participation includes local government, law enforcement, community agencies, and other relevant community partners *Primary Outcomes:* • Victimization (bullying and aggression): Self-report • Perpetration (bullying): Self-report and observer-report	*Duration of Program:* • One school year • Can be integrated into the class curriculum *Financial Resources:* • $3500: Two day training course • $1720: One year telephone consultation with a certified trainer • Schools pay for trainer's travel costs, lodging, meals, and local transportation (est. $800[xiii]) • $2330[xiv]: Core program including school-wide and teacher guides • $680: Questionnaire[xv] (*Annually)* • Supplemental materials (e.g., class meeting guide, DVDs) extra ($200-$2400) *Instructor:* • Led by a school-wide committee and individual classroom teachers • A certified trainer is required for staff training and consultation	*Grade Range* Grades K to 12 Language English ☑ French ☐ Spanish ☐ *Target Population* Universal ☑ Targeted ☐ Indicated ☑ *Program Components* Whole school ☑ Classroom-based ☑ Parent component ☑

[xii] There are no certified Olweus Trainers available locally; the closest available trainer for Nova Scotia is located in New Brunswick

[xiv] Estimated for an 'average sized' school with 500 students, 30 teachers, and 12 co-ordinating committee members

[xv] Online questionnaires: $500, interactive whiteboard questionnaires: $590

Olweus Bullying Prevention Program Recommendation

Factors	Decision	Explanation
High or moderate quality evidence?	Yes ☐ No ☑	Research: • Three quasi-experimental studies: two age-cohort designs[13,15] and one non-randomized controlled trial[14] • One study[13] evaluated elementary to secondary students, two[14,15] evaluated middle school students, and one[15] also included teacher reports. • A large scale age-cohort study[13] (in Pennsylvania) with over 50,000 students compared effects in three grade levels (elementary, middle, and high school). A group with district-wide implementation and high access to training and support was followed for 1 or 2 years (HALT 1 and HALT 2) and a group with less support and school-level implementation only was followed for 1 year (PA CARES). Comments: • Low quality evidence: Non-randomized designs, little control, and many unreported outcomes. • Age-cohort studies[13,15] compared post-test students to different students of the same age a year previously who had not yet received the intervention. • One study[14] compared the program to less formal bullying prevention initiatives (e.g., school policy). • Two studies[13,15] assessed bullying and one[17] assessed frequent victimization by aggression. At least two studies[14,15] defined victims/perpetrators as those involved at least 2 times in a month.
Certainty: *Benefits outweigh the downsides?*	Yes ☐ No ☑	Findings: **Bullying: Self-reported victimization** • The Pennsylvania study[13] did not report significant effects in elementary or middle school students in HALT 1, HALT 2, or PA Cares cohorts. Results for Halt 1 were not reported for middle or high school students, but there was a very small significant decrease in middle/high students in the HALT 2 cohort (about 17.5% to 14.5%; absolute difference = 3%. There was no effect on middle school students in PA cares but a very small benefit for high school students (about 16.5% to 13.5%; absolute difference = 3%. • An age-cohort study[15] with grade 7 and 8 students in one school found no overall effect of the program on victimization; however, there were some small-moderate mixed effects in girls[xvi]. For example, there was a 31% decrease in victimization for grade 7 girls, but a 36% increase in perpetration for grade 8 girls. **Aggression: Self-reported victimization**

[xvi] This study conducted unplanned tests on a large number of variables by grade and gender; there was no statistical correction for the large number of comparisons. The sample size for these observations would have been small (around 40). Interactions are not reported in studies with multiple comparisons unless they are stated a priori.

		• A non-randomized[14] controlled trial with grade 6-8 students from 10 schools found no overall effect of the program on relational or physical victimization compared to less formal bullying prevention approaches; positive effects were observed in Caucasian students only (a 28% decrease in relational victimization and a 37% reduction in physical victimization). **Bullying: Self-reported perpetration** • The following results were observed in the large-scale Pennsylvania study[13]: o *Elementary*: A statistically-significant, but very small (0.3%) absolute reduction in perpetration was observed for elementary students after 1 year in the HALT cohort, but there was no significant effect after 2 years. There was a very small (0.7%) absolute reduction in PA CARES. o *Middle/High School*: Effects were not reported for middle school students in HALT 1, but there was a very small statistically significant reduction from 13% to 11% (absolute difference = 2%) in high school students in HALT 1, and a very small statistically significant reduction from 14.5% to 8.0% (absolute difference = 6.5%) in middle/high school students after 2 years. There was a very small significant reduction from 14% to 11% in high school students in PA CARES (absolute difference = 3%) but no effect in middle school students. • An age-cohort study[15] with grade 7 and 8 students in one school found no overall effect of the program on taking part in bullying, but a statistically significant *increase* of 36% was noted in grade 8 females. • The same study[15] found that teachers observed a 49% relative increase in bullying Comments: • Few to no benefits in elementary or middle school; small consistent effects in high school • There is some evidence of mixed and negative effects in middle school girls (based on 1 school only)
Resource implications: Benefits worth the costs?	Yes ☐ No ☑	• The program is resource-intensive and has shown null effects or very small effects that are mixed in direction
Summary and Recommendation:	**WEAK** recommendation **AGAINST** the Olweus Bullying Prevention Program. The program is very expensive; low quality evidence reports a mix of null, very small positive, or very small negative effects.	

Steps to Respect

www.cfchildren.org/steps-to-respect

(*Program no longer published but still used in many schools as of 2015*)

General Description & Outcomes	Program Resources	Program Specifics
General Descriptions: • The program aims to change attitudes about bullying, increase empathy for victims, and educate bystanders to prevent bullying. • The program targets multiple levels of the school environment through intervention components directed at the school, peer, and individual levels. • School-wide components focus on fostering a positive school climate and norms (e.g., monitoring, discipline, and training to intervene with students involved in bullying). • All staff members are trained using a core instructional session. All counsellors, administrators, and teachers receive two additional training sessions. • The program has a semi-structured curriculum with hands-on and literature-based lessons intended to promote socially responsible norms and behaviour and increase social and emotional skills. *Primary Outcomes:* • Aggression: Self-reported victimization • Bullying: Observed victimization • Aggression: Self-, teacher-, and observer- reported perpetration • Bullying: Observer-reported perpetration • Aggression: Student and staff reported school aggression-related problems	*Duration of Program:* • 11 classroom lessons (45 minutes) and a 15-minute booster lesson taught within the same week. • Designed to have a cumulative effect over 3 years of implementation *Financial Resources:* • $700: Complete School Program including: • School-wide Implementation Support Kit (includes program guide, training manual, staff and parent training video) • 3 Curriculum Kits (grades 3-6) (includes classroom lessons, literature units, classroom DVD, posters) *Instructor:* • Teacher-led using available materials and manuals	*Grade Range* Grades 3 to 6 Language English ☑ French ☐ Spanish ☑ *Target Population* Universal ☑ Targeted ☐ Indicated ☑ *Program Components* Whole school ☑ Classroom based ☑ Parent component ☑

Steps to Respect Program Recommendation

Factors	Decision	Explanation
High or moderate quality evidence?	Yes ☑ No ☐	Research: • Moderate quality evidence. • Two RCTs[30,31], one follow-up of an RCT[32], and two secondary analyses[33,34] conducted in California, USA. See note[xvii] Comments: • Studies evaluated children from grades three to six in economically and ethnically diverse samples • Playground observations of students were carefully conducted; however, it is not clear whether observers were blinded (e.g., if they knew whether the students were assigned to the intervention or control group). • Student-self report and teacher-report measures all focussed on aggression; researchers in the playground observations assessed bullying. • Outcomes were measured since the beginning of the school year in Fall 2008 (pre-test) and Spring 2009 (post-test)
Certainty: Benefits outweigh the downsides?	Yes ☐ No ☑	Findings: **Aggression: Self-reported victimization** • Two randomized trials[30,31] found no statistically-significant effect of the program after 1 year of implementation. **Bullying: Observer-reported victimization** • A randomized trial[30] found no statistically-significant effect of the program on playground observations of victimization after 1 year of implementation. **Aggression: Self-, teacher-, and observer-reported perpetration** • Two randomized trials[30,31] found no statistically-significant effect on self-reported direct or indirect aggression after 1 year of implementation. • A randomized trial[31] found no statistically-significant effect on teacher-reported non-physical aggression, but a small (*d*= 0.25) positive effect on teacher-reported physical aggression. Reports of physical aggression increased from 21% to 23% in the treatment group (absolute difference= 2%, relative difference = 9%) and from 17% to 29% in the control group (absolute difference= 12%, relative

[xvii] We focussed on the results of an evaluation[30] comparing one year of treatment to one year of control, rather than on an evaluation[32] that compared two years of treatment to one year of control and uncontrolled longitudinal effects. We also did not review a study[52] that focused on a single element of indirect aggression (malicious gossip), because the broad domain of indirect aggression was already analyzed in this cited study[30].

		difference = 41%). • A randomized trial[30] found no statistically-significant effect of the program on playground observations of aggression after 1 year of implementation. **Bullying: Observer-reported perpetration** • A randomized trial[30] found a small positive effect (est. *d* =0.21) on playground observations of bullying perpetration after 1 year of implementation, consisting of a relative reduction of one incident in three hours. Reports of bullying increased from an average of 0.85 incidents per hour to 0.97 in the treatment group, and from 0.73 to 1.19 in the control group. The difference is equivalent to one incident of bullying prevented in three hours of playground time. **School aggression-related problems: Student and staff report** • A randomized trial[31] found no statistically-significant effect of the program on students' evaluation of aggression-related problems in the school after 1 year of implementation. However, staff reports showed a small (*d*=0.35) positive effect, consisting of a relative reduction of a third of a point on a 4-point scale. However, teachers reported higher rates of aggression-related problems than did administrative and non-academic staff. Comments: • The program showed no effects on self-reported aggression or on playground observations of aggression perpetration or bullying victimization; however, there was a small effect on bullying perpetration. • The program had no effect on teacher-reported nonphysical aggression, but a small reduction in teacher-reported physical aggression. • The program had no effect on student-reports of aggression-related problems in the school, although staff perceived fewer problems. Administrators who were most removed from the classroom perceived the greatest benefit of the program. • An RCT found no effect of program implementation on observed behaviour, and adherence to lessons was not associated with change in self-reported experiences. However, higher-quality lesson delivery resulted in greater self-reported victimization and more difficulty responding to bullying assertively[33]. • A second, larger randomized trial found no effect of program implementation on self-reported perpetration of aggression; however, lesson adherence was related to lower victimization[34]. • Claims such as "teachers will be amazed at how much more time they have to actually teach when bullying conflicts become a thing of the past" are not supported by the observed strength of program effects.

Resource implications: Benefits worth the costs?	Yes ☐ No ☑	• The cost of program implementation is low compared to other anti-bullying programs. • The program can be incorporated into daily lessons and most material can be re-used annually with minimal teacher training costs. The program is teacher-led and does not require external consultants. • Small benefits observed by staff or researchers were not corroborated by student reports. • Program implementation had mixed or no effects on primary outcomes.
Summary and Recommendation:	**WEAK** recommendation **AGAINST** using Steps to Respect as a school based bullying prevention program. Although the program offers relatively stronger effects at a lower price than other programs, the small benefits observed by staff are not corroborated by student reports. The program is only available while supplies last, and has been replaced by a condensed 'Bullying Prevention Unit' (K-5), which has not been evaluated.	

Second Step: Student Success through Prevention (SS-SSTP)

http://www.cfchildren.org/second-step/middle-school.aspx

<table>
<tr><th>General Description & Outcomes</th><th>Program Resources</th><th>Program Specifics</th></tr>
<tr>
<td rowspan="2"><u>General Descriptions:</u>
• A universal curricular classroom-based intervention delivered by classroom teachers.
• The program focuses on 5 themes: Developing empathy and communication; bullying prevention; emotion management; substance abuse prevention; and action steps for problem solving, decision making, and goal setting.
• Grade 6 bullying prevention lessons (2) focus on awareness of bullying and the role of bystanders; grade 7 lessons (3) focus on bullying and bystanders, cyberbullying and sexual harassment; grade 8 lessons (3) focus on bullying in friendships, stereotypes and prejudice, and bullying in romantic relationships.
• Lessons are highly interactive, incorporating small group discussions, and activities, dyadic exercises, whole class instruction and individual work. Teaching strategies are media-rich including interactive videos.

<u>Primary Outcomes:</u>
• Aggression: Self-reported verbal perpetration
• Aggression: Self-reported relational perpetration
• Aggression: Self-reported physical perpetration
• Aggression: Self-reported victimization
• Aggression: Self-reported homophobic name-calling victimization & perpetration
• Aggression: Self-reported sexual harassment/violence victimization & perpetration</td>
<td rowspan="2"><u>Duration of Program:</u>
• 13-15 weeks
• Weekly lessons are 50 minutes long (or 25 minutes over 2 sessions) 30 to 45 minute and a 15-minute “booster” lesson taught within the same week.
• 2-3 lessons on bullying prevention

<u>Financial Resources:</u>
• $999 for Grade 6-8 package; includes lessons, resources, and video-based training

<u>Instructor:</u>
• Teacher-led using available materials and manuals (available online)</td>
<td><u>Grade Range</u>
Grades 6-9

<u>Language</u>
English ☑
French ☐
Spanish ☐</td>
</tr>
<tr>
<td><u>Target Population</u>
Universal ☑
Targeted ☐
Indicated ☐

<u>Program Components</u>
Whole school ☑
Classroom based ☑
Parent component ☐</td>
</tr>
</table>

Second Step: Student Success Through Prevention Recommendation

Factors	Decision	Explanation
High or moderate quality evidence?	Yes ☑ No ☐	Research: • One well-designed randomized control trial[35] conducted in Midwestern USA Comments: • Only the grade 6 version of the program has been evaluated • Nested cohort design with randomization at the school level • Intervention and control schools were matched to assure similar baseline characteristics • Measures were: aggression occurring in the past 30 days (1 year for sexual aggression) as indicated by 'yes' responses to 2 or more items (1 item for sexual aggression).
Certainty: Benefits outweigh the downsides?	Yes ☑ No ☐	Findings: **Self-reported aggression: Victimization** • A randomized control trial[35] showed no statistically-significant effect of the intervention after one year of implementation. **Self-reported verbal aggression: Perpetration** • A randomized control trial[35] showed no statistically-significant effect of the intervention after one year of implementation. **Self-reported relational aggression: Perpetration** • A randomized control trial[35] showed no statistically-significant effect of the intervention after one year of implementation. **Self-reported physical aggression: Perpetration** • A randomized control trial[35] showed a statistically-significant but very small effect (OR=0.70, *d*= 0.197), consisting of a third of a point as measured on a 5-point scale. The raw data showed that physical aggression increased from 41% to 46% in the control group (absolute difference= 5%) and from 35% to 37% in the treatment group (absolute difference = 2%). The absolute reduction in perpetration that could be attributed to the intervention was therefore only 3%. **Self-reported homophobic name calling: Victimization & perpetration** • A randomized control trial[35] showed no statistically-significant effect of the intervention on victimization or perpetration after one year of implementation. **Self-reported sexual harassment/violence: Victimization & perpetration** • A randomized control trial[35] showed no statistically-significant effect of the intervention on victimization or perpetration after one year of implementation.

		Comments: • All measured outcomes (problem behaviours) showed increases from pre-test to post-test in both the intervention and control groups (e.g., sexual harassment perpetration increased by 175% in the intervention group and 155% in the control group). However, after subtracting the differences between the intervention and control groups, the intervention had no effect on most outcomes. • The intervention group showed a smaller increase in physical aggression than the control group • No reported downsides.
Resource implications: Benefits worth the costs?	Yes ☐ No ☑	• The cost of program implementation is low compared to other anti-bullying programs. • The program can be incorporated into daily lessons and most material can be re-used annually with minimal teacher training costs. The program is teacher led and does not require external consultants. • Only the grade 6 version of the program has been tested in a single study, which showed that the program had no effect on most tested outcomes but a small positive effect on physical aggression.
Summary and Recommendation:	**WEAK** recommendation **AGAINST** using SS-SSTP as a school-based bullying prevention program. The program has only been tested in one study with one grade level, which showed little to no effects.	

WITS Primary Program

http://www.witsprogram.ca/

General Description & Outcomes of Interest	Program Resources	Program Specifics
General Descriptions: • Focus on preventing peer victimization and increasing social competency (e.g., help-seeking, internalizing) • Teachers integrate one WITS book a month into language arts curricula. Each storybook focuses on a form of interpersonal conflict and how it would be resolved by the WITS response. Questions and activities complement the storybooks. • Students are 'sworn in' as 'WITS special constables' by a police officer at a school assembly. Police officers, emergency service personnel, and university athletes make monthly classroom visits to reinforce that WITS strategies are important beyond the school community. • Parents are encouraged to read WITS books with their children and use WITS principles and language in the home. *Primary Outcomes:* • Aggression: Self-reported physical and relational victimization • Aggression: Teacher-reported physical perpetration	*Duration of Program:* • Minimum of one lesson per month across the school year *Financial Resources:* • $5 - $20 per book. Eight books recommended minimum. • Questions and activities to complement the storybooks are available for free online • Program manual: $15 • A basic complement[xviii] of reminder gifts and classroom signs (not including the stuffed walrus) would cost $500 *Instructor:* • Teacher-led based on available materials and manuals. • Additional visits from community police officers, emergency service personnel, and university athletes	*Grade Range* Grades 1 to 3 Language English ☑ French ☑ Spanish ☐ *Target Population* Universal ☑ Targeted ☐ Indicated ☐ *Program Components* Whole school ☑ Classroom based ☑ Parent component ☑

[xviii] 150 students across 6 classrooms (25 students in 2 classes in each of 3 grades)

WITS Primary Program Recommendation

Factors	Decision	Explanation
High or moderate quality evidence?	Yes ☑ No ☐	Research: • Moderate quality of evidence • Four studies: A 30-month quasi-experiment[36] with a cohort of grade 1 students followed until the end of grade 2, with a secondary analysis[37] and a follow-up study[38] observing students until grade 6, and an 18-month quasi-experiment[39] that followed grade 1-3 students for 18 months. • All studies evaluated WITS in low- to high- income urban British Columbian elementary students. Comments: • All studies compared schools that had implemented the program for at least one year *before* baseline to matched control schools that had never implemented the program; control schools used other social skills programs such as Second Step. • Measures assessed peer victimization and perpetration rather than bullying per se. • The measures had no temporal reference period; i.e., they measured lifetime prevalence of aggression, which can make it more difficult to show the effects of a short-term program. • The analyses accounted for several controls, including baseline aggression, income, and gender. • Program effects could be underestimated by studies that faced low implementation and high attrition.
Certainty: Benefits outweigh the downsides?	Yes ☑ No ☐	Findings: **Self-reported relational aggression: Victimization** • The 30-month quasi-experiment[37] showed a very small (*b*=-0.06)[36] effect on relational victimization, consisting of a third of a point on a 3-point scale. • There was a small effect on classroom-level victimization; raw scores in the treatment group decreased from a mean of 2.92 at the beginning of grade 1 to 2.19 at the end of grade 2. In the control group, average scores increased from 2.38 to 2.42. Thus the benefit of the program was ¾ of a point on a 3-point scale. The effect was larger in low income compared to high-income schools. • The 6-year follow up study[38] showed that that the program continued to show beneficial effects until the end of implementation in grade 3, but mean scores increased after many students were replaced with new students (who had not been exposed to the program) in grades 5 and 6. The program showed a small beneficial (*d*=0.20, *b*= -0.130[xix]) effect. Raw scores in the treatment group decreased from a mean of 0.57 at the beginning of grade 1 to 0.47 at the end of grade 6; in the control group,

[xix] A standardized slope (*beta*) can be interpreted as a correlation; in this case, a correlation of -0.1 between program use and relational victimization is small.

		average scores increased from 0.47 to 0.49. Thus the benefit of the program was 1/8 of a point on a 3-point scale. • The 18-month study[39] showed that relational victimization declined by 28% in students who participated in the WITS program, whereas it remained stable for control students. **Self-reported physical aggression: Victimization** • The original 30-month quasi-experiment[36] showed a small statistically-significant effect on individual levels of physical victimization (*b*=-0.12), consisting of a third of a point on the 3-point scale[37]; there was also a small effect on classroom-level victimization; raw scores in the treatment group decreased from a mean of 2.92 at the beginning of grade 1 to 2.06 at the end of grade 2; in the control group, average scores increased from 2.58 to 2.49. Thus the benefit of the program was ¾ of a point on a 3-point scale. The effect was larger in low income (*moderate effect*) compared to high income schools (*small effect*). • The 6-year follow up study[38] showed that that beneficial effects continued until the end of implementation in grade 3, but mean scores increased possibly because of high attrition in grades 5 and 6. Even including these data, the program showed a very small beneficial (*d*=0.17, *b*= -0.102) effect. Raw scores in the treatment group decreased from a mean of 0.57 at the beginning of grade 1 to 0.37 at the end of grade 6; in the control group, average scores increased from 0.50 to 0.42. Thus the benefit of the program was 1/7 of a point on a 3-point scale. • The 18-month study[39] showed a small effect (r^2=0.01) on physical victimization, which declined by half a point on a 3-point scale. An average child in the WITS program saw a 31% decrease in physical victimization, while this remained stable in the control group. **Teacher-reported physical aggression: Perpetration** • The 6-year study[38] showed no statistically significant effect[xx] on physical aggression. However, by the end of grade 3, at program completion but before the study was affected by high attrition, there was a statistically significant difference between WITS and control students. Raw scores in the treatment group increased from a mean of 0.19 on a 4-point scale at the beginning of grade 1 to 0.22 at the end of grade 3; in the control group, average scores increased from 0.23 to 0.36. Thus the benefit of the program was very small (1/10) of a point on a 4-point scale.

[xx] Marginally significant benefit: p<0.10; *b*=-0.05

		Comments: • WITS had a small effects on relational and physical victimization, and a marginal effect on physical aggression perpetration. • Program effects in the 30-month and 6-year studies[36–38] are in comparison to other social skills programs used by control schools, rather than to no treatment. • The six year study[38] showed that the program had its greatest effect in grade 3 (i.e., after 2 years of implementation). Program effects were quickly lost in the last two years of the study, when about half of the WITS students were replaced with new students who had not received the program. • Implementation fidelity was moderate in the 6-year study[38]. Implementation was lower in the 18-month study[39]; e.g., 65% had the swearing-in ceremony, 69% used WITS language with students, 56% held a police visit, 32% used the posters, 33% recognized a student for using their WITS skills, 24% read from a WITS book 3-4 times, and 4% held a class visit with a varsity athletes.
Resource implications: Benefits worth the costs?	Yes ☑ No ☐	• The cost of the program is low. Reminder gifts such as stickers and pencils (a recurring cost) are the most expensive part of the program.
Summary and Recommendation:	**WEAK** recommendation **FOR** using the WITS Primary program as a school-based anti-bullying program. The program has the strongest effects at the lowest cost. Nevertheless, program effects are small.	

DISCUSSION

Using the GRADE approach to analyzing study results, we reviewed published peer-reviewed studies of bullying prevention programs where students who received the BP prevention were compared to students who did not. We reviewed the following seven programs: Dare to Care, Friendly Schools & Families, KiVa, the Olweus Bullying Prevention Program, Steps to Respect, Second Step: Student Success, and the WITS primary program.

All the studies we reviewed earned a 'weak' recommendation for or against implementation. Most studies of bullying interventions are conducted in a way that casts doubt on their conclusions, so no program received a strong recommendation. Take for example the WITS program; in the evaluations of this program, some schools had already decided to use the program and researchers had to find comparable schools that were not using the program. The challenge is that schools choosing to do an intervention may not be comparable to schools that do not. Even if schools appear to be comparable, they may differ on important but difficult-to-measure attributes such as the motivation of the school staff to implement any change that might reduce bullying, their dedication to their students' wellbeing, the health of the school climate or neighbourhood influences on the school. It may therefore be these factors by themselves or in addition to the BP program that account for the change. Following a strict randomization method will remove the biases that such factors can introduce so that one can be more confident that any changes that are found are due to the program and not to other differences between schools. Unfortunately, randomization in school-based studies is less common than selection by convenience or desire of decision makers. The problem of identifying true program effectiveness is even more difficult when program effects are small because even minor differences in motivation, student, school or neighbourhood characteristics may make the result (e.g., how bullying rates differ between schools that received or did not receive the intervention) smaller or larger than it actually is.

A weak recommendation does not mean that a program should not be implemented; rather, it indicates that it is particularly important to evaluate the program in its new setting. This helps ensure that precious time and money are not being wasted. Note that one may need to implement a program for several years before significant benefits accrue, because it could take time for teachers and administrators to become skilled and comfortable in delivering the program; also, some programs (such as WITS) show increased benefits as the same students are exposed to the program year after year.

Of all the programs we reviewed, only one received a recommendation in favour of implementation: the WITS program from the University of Victoria, British Columbia. WITS delivered relatively strong effects at low cost and showed long-term benefits. It also has components that promote healthy relationships, such as stories and classroom discussion about how to solve interpersonal conflict.

All other programs received weak recommendations against implementation, largely because they were resource-intensive yet delivered little or no reduction in bullying or victimization, or were not feasible in a North American environment.

The Dare to Care program, which focuses on training bystanders to respond to bullying, is supported by one low-quality

study[21], whose treatment and control groups were very different at baseline (the intervention school had more bullying at baseline). The authors reported that the two groups were too different to compare, but compared them anyway. There was no effect on 'bullying experienced' in either the treatment or the control group. On the other hand, the level of 'bullying witnessed' decreased by 2.7 points on a 30-point measure in the treatment group (a 21% relative reduction) and by 0.94 points in the control group (a 15% relative reduction). The change was statistically significant in the treatment group but not in the control group, leading the authors to interpret that the program was effective. However, the treatment effect needs to be adjusted for the improvement in the control group. Another difficulty with comparison is that the score in the treatment group had more 'room to move' because it began with twice the level of bullying (it is easier to drop from a high than a low score, whether by chance or through an intervention).

We also note that the study measured bullying perpetration but did not report results for this measure. We expect that studies should report all outcomes, so users can get the 'big picture' of its effects, rather than seeing only those the authors chose to show. Overall, the costs of this program are high and recurring, and evidence for benefits is weak. Notably, the curriculum-based program that was evaluated in the study has been replaced by a series of workshops, whose sole effect could be even less than what was observed with a more in-depth curriculum.

The Friendly Schools & Families program takes a strengths-based, whole-school approach to creating a cohesive, supportive community. The program showed promise in two well-designed longitudinal randomized control trials based in Australia [23,24]. The evaluations was beset by low compliance, particularly for the family activities designed to be implemented by parents or caregivers. Implementation at the classroom level was less than ideal, with 63% to 75% of program components implemented. The program showed no effect on perpetration, yet there were small effects on victimization, consistent effects on witnessing bullying, and evidence that benefits accrued over time. However, the full cost of the program is prohibitive (requiring travel to Australia) and issues with implementation also suggest that the program is too resource-intensive overall.

The KiVa program, from Finland, also teaches witnesses to bullying (bystanders or witnesses) to intervene when bullying occurs. The program has shown some promise with very small or small effects in elementary school students, and was the only program to document a decrease in cyberbullying victimization. However, it was not effective in middle school students. Versions in other languages including English have been developed but do not yet have published studies (at least in English) evaluating their effectiveness. KiVa is certainly worth considering in an anglophone setting once there is evidence that it works there.

The Olweus Bullying Prevention Program (OBPP) focuses on adult-centred activities including displaying warmth and a positive interest in students while functioning as positive role models, and setting boundaries for acceptable behaviour while consistently enforcing consequences when rules are broken. The results for the OBPP gave us cause for concern, given the program's high profile. Costs for program materials and training are high, but the quality of published evidence is low and mainly null or very small results were reported including one detrimental effect (see p. 26). The report of the Pennsylvania statewide implementation of the OBPP[13] reported some very small positive effects that were statistically significant because of the number of students involved

(about 10,000) but not meaningful to school personnel. See for example the 0.3% reduction in bullying discussed previously (p. 26).

The study[13] also did not report reductions in rates of perpetration and victimization that were collected for a number of groups of students, presumably because the changes were even smaller than those reported. A GRADE criterion of quality is that all relevant outcomes should be reported, not just those chosen by the authors to make a particular point. This helps the reader better assess whether or not a program has meaningful consistent effects that did not arise by chance.

The Steps to Respect program is no longer published but is still used in some schools as of 2015. It aims to change attitudes about bullying, increase empathy for victims, and empower witnesses or bystanders to prevent bullying. Two randomized control trials[30,31] found at most very small benefits. Playground observation showed that bullying perpetration increased in both the treatment and control groups, but the increase was smaller in students who had received the program. Program claims such as "teachers will be amazed at how much more time they have to actually teach when bullying conflicts become a thing of the past" are clearly overstated. The program has now been replaced by a condensed 'Bullying Prevention Unit' for grades K – 5, which has not been evaluated.

Second Step: Student Success through Prevention focuses on a number of outcomes, including teaching social and emotional skills, preventing substance abuse, and problem-solving and goal setting, in addition to preventing bullying. Over the three program years, students receive a total of 8 interactive lessons on bullying. The program was evaluated in one well-designed randomized control trial[35], which examined 6 types of aggression as outcomes in grade 6 students only. The grade 6 students would have only been exposed to two bullying prevention lessons. The program had no effect on 5 of the 6 types of aggression measured. It showed a very small positive effect on perpetration of physical aggression, which increased in both the treatment and control group but showed a smaller increase in the treatment group. While Second Step is a useful program for improving social and emotional learning, it seems to have little effect on reducing aggression; at best, physical aggression in the intervention group increased at a slower rate than in the control group, but no absolute reductions in aggression were observed. We note that the program was only tested in grade 6 students, who finished the first year of a three year program, and had only been exposed to a quarter of the total lessons on bullying. It is possible that the program could be effective after all three years have been implemented, but this requires an evaluation study.

Overall, there are many possible reasons why these programs showed limited effectiveness, which are detailed below.

Research studies typically only examine a small number of students across one or a few schools. The statistical tests that researchers use to determine whether program effects exist are very sensitive to sample size and especially the small number of schools or classrooms. In small studies, true and possibly important effects may not statistically significant. In large studies, effects may be found to be statistically significant but are so small as to be meaningless to those who want to start a program that will reduce bullying. These larger studies with statistically significant but trivial or small impacts will be promoted as 'evidence-based' despite their negligible value in the real world.

Bullying is an intentional act where someone consciously decides to use their power to dominate another person. Structured programs applied to the school as a whole may not address the internal and external factors leading some people to bully. For example, some youth may bully because they are regularly exposed to abuse at home. Such youth may bully because this is what is modelled to them at home as the way to solve conflicts. Youth who have difficulty with emotional regulation can also behave aggressively. A universal program designed to build social competency in typical student is likely not enough for students with serious disruptive behaviour, anger or other issues that increase their risk of mental or social difficulties. It is important to note that implementing a universal bullying prevention program does not replace more in-depth interventions for groups of high risk students or individual students.

A successful evaluation of a bullying prevention program shows that it worked when implemented in a particular way in a particular setting. Bullying prevention interventions are complex with different roles for students, teachers, other school staff, families and communities. It is difficult for complex and dynamic organizations like schools to implement programs in the exact manner intended by program creators. This is the concept of 'fidelity'. It is possible that some of the programs we reviewed as providing insufficient benefits for the cost did not work because they were not implemented fully or properly. For example, in one evaluation of the KiVa program, teachers had delivered less than half of the lessons and in one evaluation of the Friendly Schools and Families program, parents only implemented 17% of the family activities. Yet, while program developers might argue that a program did not show effects because of insufficient implementation, we recognize that schools work with limited resources. If school administrators cannot properly implement a program during an evaluation study (when stakeholders such as researchers and program developers are no doubt keen to see the program implemented as designed) it may simply indicate that the program is not feasible in that school setting, though it might be in others. Full fidelity to implement exactly as designed is also in conflict with the role of teachers as professionals trained to design and modify their own curricula to meet specific goals.

There are many factors that affect bullying, for example, school climate, quality of relationships within the school, and presence or absence of a dedicated anti-bullying champion such as a teacher or principal. Some programs focus on bystander involvement, some on changing attitudes, and some on adult behaviours such as modelling and setting boundaries. A given bullying prevention program may not address the most relevant cause (or all the relevant causes) in a given setting.

There are some limitations to our review. We found the programs we reviewed through an environmental scan of programs used in Nova Scotia schools in 2011, and also included other popular programs that stakeholders expressed interest in. We did not conduct a comprehensive review of all the bullying prevention programs that exist, and there may be effective programs that we are not aware of. We only reviewed programs that were evaluated in at least one controlled study published in a peer-reviewed scientific journal. There may be effective programs that have yet to be formally tested, and there may be effective programs that received high quality evaluations that were published through different venues (e.g., PhD theses, government reports).

While there may be programs that work but only have not been tested, we believe that any program that is going to be

rolled out to a large number of students (likely at a significant cost) should first be tested or piloted, and that even when a tested program is implemented, it should be continuously monitored in its new setting.

Given the little impact of universal bullying prevention programs for the amount invested, it may be worthwhile for organizations to look at other universal interventions that are more focussed on developing healthy relationships, social and emotional learning and school climate. We know there is value in these projects from the point of view of students being ready to learn and becoming healthy and productive adolescents and adults. There *may* be a place for a universal program focused primarily on bullying prevention once those priorities are addressed. Note that there will always be a place for strategies that target high risk individuals or groups as no program applied to an entire school or organization will prevent all individuals from bullying others.

Glossary

Terms about Behaviour

Aggression: acts that inflict harm on others

Bullying: Aggression characterized by intent to harm, repetition or persistence over time, and a power differential favouring the perpetrator[1].

Perpetration: Those who bully or aggress against others are perpetrators of bullying or aggression.

Victimization: Those who are bullied or aggressed against by others are victims of bullying or aggression.

Terms about Interventions

Indicated (Tier 3): The intervention targets an *individual (and possibly their family)*, e.g., students who bully, who are victims, or may have serious mental or emotional issues that are unlikely to respond to a group-based intervention.

Targeted (Tier 2): The intervention targets a *group* such as children with self-regulation or anger management issues or with poor social skills.

Universal (Tier 1): The intervention targets the entire school population and ideally families of students in the school. This includes students and all staff who have contact with students.

Terms about Research Design

Attrition: Attrition occurs when participants leave a study before it is completed, so that there are fewer students to assess at the end than at the beginning of a study. This can bias (i.e., distort) the results. For example, if the most disruptive or aggressive students dropped out of the study (e.g., by leaving school or being suspended), rates of those behaviours would decrease for this reason alone. Attrition is mostly a concern when it differs between the control and treatment group (*differential attrition*), which can make a program look more effective (or less) than it really is.

Age-cohort design: A study design that compares post-intervention scores to the previous year's students of the same age that did not receive the intervention (e.g., comparing this year's grade 4 students at post-test after 1 year of intervention to the pre-test results of this year's grade 5 students, which were obtained back when these students were in grade 4). This design controls for the effects of age that threaten the interpretation of simple pre-post studies (e.g., comparing grade 4 students at pre-test to grade 5 students after 1 year of intervention at post-test); however, it is threatened by historical effects (e.g., there may be less bullying in one year because of other events, such as intensive anti-bullying campaigns) that might have taken place in one year only.

Cohort or observational study: Cohorts (groups) of people who share a certain exposure (e.g., program vs. no program) are followed over time. The researcher measures characteristics of the cohorts before the intervention, possibly during the intervention, and either at the end of the intervention or sometime afterward and then observes the outcomes (e.g., victim vs. not victim). The researchers have no say in how the cohorts were formed. They observe and do not intervene. Observational studies provide a lower-quality of evidence for program effects, because we don't know if outcomes are different because a) the program worked, b) the more committed schools chose to use the program, and/or c) schools using and not using the program differed at baseline.

Experimental study: The researcher decides which people or schools receive the program and which do not then observes what happens. The highest quality experimental study design is the *randomized controlled trial* (see below).

Longitudinal study: Typically, cohort studies that last for many years are called longitudinal studies.

Quasi-experimental study: The researcher has some but not complete control in deciding which research participants receive the intervention. This sometimes happens in educational settings because schools or school boards decide to implement a program before considering how to compare it to others. The evaluation team therefore cannot alter the decisions made about who will implement an intervention but they can have input in selecting comparison sites that have characteristics that are as close as possible to the intervention sites. Therefore, these studies have potentially more ability to remove bias than purely observational studies but are cannot guarantee lack of bias in how an intervention is allocated to a research participant the way that a randomized controlled trial can. Therefore, it is possible that differences between intervention and control sites might be due to other factors rather than the program.

Randomized controlled trial (RCT): A type of study in which research participants (e.g., students, classrooms or schools) are assigned to the program or control group using a recognized valid method of random assignment. This is the highest level of research evidence: groups may still differ at baseline but differences are due to chance and can be adjusted for. In observational studies, it is

difficult to disentangle baseline differences between groups (e.g., school climate) are more difficult to account for in observational studies.

Reference period: This is the duration of a period prevalence estimate; for example, one could examine the prevalence of bullying over one month, one semester, one year, or even the entire lifetime. Shorter reference periods are more valid because they are less influenced by recall problems. Very long reference periods (e.g., have you ever been bullied?) are less appropriate for intervention studies, because bullying that happened in the distant past cannot be prevented and may be quite different in nature from recent bullying.

Prevalence: The proportion of people with a certain condition sometime during a defined period of time (the *reference period*), divided by total number of people in a given sample. For example, if 5 students in a class of 50 were bullied in the past month, the one-month prevalence rate is 5/50 = 10%. Typically, prevalence refers to the rate of a single point in time and is called "point prevalence." In studies of human behaviour, such as bullying and aggression, typically measures the proportion of people who had an event in a period of time such as a month. This is called "period prevalence."

Secondary analysis: A study that re-analyzes data from a previous study instead of collecting and analyzing new data.

Terms about Program Effects

Large effect: A difference that is 'grossly perceptible' or very obvious, such as the difference in the average heights of men and women.

Moderate effect: A difference that is clearly present to a careful observer, such as the difference between the average heights of 14 and 18 year old girls.

Small effect: An effect that is difficult to see, such as the difference between the average heights of 15 and 16 year old girls. There is a lot of overlap between the scores for the two groups.

Very small effect: An effect that is so small that it might be considered 'trivial'.

Appendix

More on experimental design and the GRADE Approach

The GRADE approach was designed to help decision-makers choose specific interventions or treatments in a health care setting, and relies heavily on study quality to arrive at recommendations. It gives greater weight to higher quality evidence (e.g., a randomized controlled trials versus observational studies). Expert opinion, anecdotes, and even theory-grounded rationalizations do not count as evidence[19].

Why are randomized trials so important? Studies that do not randomize participants to groups (e.g., observational or quasi-experimental studies) can make false interpretations about program effects if the two groups are not comparable at baseline or if some other factors, unknown to the researchers, are really driving the difference between the intervention and control groups. A good example is the belief, held by many people in the health professions and in the general public, that a moderate amount of alcohol protects the heart and blood vessels[40]. This belief is based on a large number of epidemiological studies showing that people with light or moderate alcohol use have better cardiovascular health than abstainers[41]. These observational studies compared people who abstained because they were too old or ill to drink (i.e., because of interactions with prescription drugs) to a much healthier group of drinkers. The difference in heart health between the two groups cannot be attributed to alcohol alone because there were other differences between the groups (age and illness) that also predicted heart disease. Once these factors are controlled for (which can often be achieved through statistical adjustments) it is no longer apparent that light alcohol consumption is healthful for the heart[42,43].

The best way to prevent systematic differences between groups at baseline is to randomly assign participants to each group. The highest-quality program evaluations will compare outcomes between schools that are randomly assigned to the program to those who are not. However, these studies are not always acceptable to schools and communities, even when a 'wait-list' strategy is employed, i.e., the control group is guaranteed to receive the intervention but at a later time. Because of this, quasi-experimental designs are often used in school-based research, i.e., studies where the evaluators had some but not complete control over assignment of the intervention[44].

A quasi-experiment may compare a treatment group consisting of schools that chose to implement the program to a control group consisting of schools that were not interested in implementing the program. In such cases it is important that researchers are aware of, and can adjust for, baseline differences between the two groups. For example, imagine that a treatment group consisted of schools that were wealthier and more able to focus on student welfare, whereas the control schools had fewer resources and more stressors. After one year, the control group might come out looking worse, and the treatment group might come out looking better, for other reasons besides the presence or absence of the program.

The ‘age cohort’ quasi-experimental design has been used several times to evaluate bullying prevention programs. In this design, students at post-test in one year are compared to last year’s students of the same age, who had not yet received the

program [45]. For example, imagine that we measure bullying before implementing the program (June 2011) and then again after one year of implementation (June 2012). Bullying would be measured in grade 6 students in June 2011 before the implementation of the program (pre-test). The program would start in September 2011 and results measured in grade 6 students in June 2012. The June 2012 results would be compared to the June 2011 pre-intervention results. This design effectively controls for age- and season-related effects that threaten the interpretation of uncontrolled pre-test vs. post-test designs, and it also benefits from the fact that students in subsequent grades in a school are more fundamentally similar to each other (e.g., demographic characteristics, family income) than to students in different schools or school districts. However, this design is threatened by 'historical' effects; for example, June 2012 may have shown better bullying outcomes than June 2011, not because of any effect of the program, but rather because of Justin Bieber's and Lady Gaga's vocal anti-bullying campaigns in response to 2012's high-profile bullying-related suicides.

Remember that an evidence-based program is one that was supported to work in a certain setting (or settings) at a certain time. There is no guarantee that a program supported to work in one context (e.g., Norwegian students in the 1980s) will automatically work in *your* setting. It is important to regularly evaluate any program by collecting data (e.g., from school surveys or analyzing routinely-collected data) in your setting. For a quick guide to program effects and recommendations of bullying prevention programs reviewed, please take a look at Table 3 on p. 51.

Why control groups are essential

It is essential that formal evaluations of bullying prevention programs include a control group. A 'control group' consists of individuals or sites that are as similar as possible (e.g., socio-economic characteristics, size of school, school climate) to the intervention group *except* for the use of the program. This allows for a valid comparison that suggests differences between the intervention and control groups may due to the program and not to other factors such as school characteristics.

Control groups are particularly important for research on bullying and aggression, which are age-sensitive. For example, the prevalence of physical aggression *decreases* sharply with age during the pre-adolescent years[46] as children develop better social skills and learn to manage frustration. In contrast, verbal and relational forms of bullying *increase* around middle school, when students experience puberty, school changes, and become more aware of their abilities to influence social hierarchy[47]. To illustrate the importance of a control group, imagine a two-year evaluation of a program with 5 year old children, where physical bullying/aggression will be measured before the implementation of the program (age 5) and after the program (age 7). If we observe marked decreases in physical bullying/aggression over two years, is it because *a*) the program works very well, *b*) 7 year olds are naturally less aggressive than 5 year olds, or *c*) something else that happened over that period, such as a community-wide anti-bullying campaign? Consider as well marked increases in verbal/social victimization between grade 6 students pre-test and grade 8 students post-test. Is it because the program is ineffective or harmful, or because grade 8 students are naturally more involved in bullying? A control group serves as a comparison condition in which, *ideally*, everything else besides the use of the program is the same: grade 8 students would be compared to other grade 8 students, rather than to grade 6 students. This allows us to attribute the difference between the treatment group and the control group to the effect of the program.

Understanding statistical significance

Measuring bullying with a self-report survey isn't as clear cut as measuring temperature in degrees Celsius or measuring size in centimeters. Bullying is often measured as a yes/no question (e.g., have you been bullied in the past month?) and this type of categorical, subjective measurement has a considerable amount of natural fluctuation. In addition to what actually happened, responses will also include factors such as forgetting what happened in the last month, incorrectly including an event that actually happened a month and a half ago, and misunderstanding what 'bullying' means.

In order to address frequency, bullying is also measured on a 'Likert scale'. A typical Likert scale measures bullying on a 5-point scale, with responses ranging through 1 (never), 2 (rarely), 3 (sometimes), 4 (often), to 5 (always). Different people may have different standards for what words like 'sometimes' and 'often' mean; for example, would eating dinner at a restaurant once a month be 'rarely', 'sometimes', or 'often' for your family? The answer would depend on many factors, such as income, habits, and culture. Unlike the difference between 3 cm and 5 cm, the difference between concepts like 'rarely' and 'sometimes' is not clear.

It's easy for outcomes measured in such subjective ways to vary even if the underlying situation has not changed; this is a

kind of error that varies randomly. In the event that a difference in measurement is small enough to be *within* the expected variation of the scale, we would say that the result is not statistically significant. For example, a change in the average level of reported bullying from 4.1 to 4.0 out of 5 is so small that it could just be attributed to the imprecision of the tool we used to measure it. However, a change from 4.1 to 1.0 out of 5 would likely be larger than the expected variability in the scale, and the result would likely be statistically significant. Likewise, a 50% relative reduction in bullying (from 8% to 4%) might not be statistically significant because this change of 4% is within the normal fluctuation of the measure (which might be 5% above or below the true value), whereas another 50% relative reduction (from 90% to 45%) would be statistically significant, because in this case a difference of 45 points surpasses the expected fluctuation of the scale. In other words, statistical significance is a way of separating a true signal from the noise (random error) that surrounds it.

It is important to note that a statistically significant effect is not necessarily a meaningful or important effect. The calculation of statistical significance is based on the size of the sample as well as the magnitude of the difference, because large samples more accurately represent what is going on in a population than small samples; in other words, large samples make it much easier to separate 'signals' (true effects) from the 'noise' (random error) that surrounds them. For example, if you wanted to determine the effect of a program in a school, surveying 200 randomly-selected students will provide far more reliable results compared to asking only 20 students. Accordingly, studies with very large samples can report 'statistically significant' effects that are negligible in practical terms, though they are 'real' differences. A good example is the large-scale evaluation of the Olweus Bullying Prevention Program (OBPP) in Pennsylvania[13], which reported a 'statistically significant' benefit in elementary students that works out to a trivial reduction in the absolute rate of bullying of 0.3% (see page 27).

Understanding standard measures of program effects: A deeper discussion

There are other measures of effect sizes In addition to Cohen's d and the odds ratio, which were described in the introduction. Effect sizes can also be described in terms of how well the independent variable (e.g., exposure to the program) predicts the dependent variable (e.g., bullying). R^2 is a measure of this relationship. A program with a moderate effect might predict almost half of the variance in bullying scores among participants. However, if only 1% of the variance in bullying scores could be attributed to the effect of the program (the lowest threshold for a small effect), we would do well to wonder what factor(s) accounted for the other 99%. Please see Table 3 (p. 51) for an expanded accounting of effect sizes.

Table 3. An expanded guide to interpreting and converting effect sizes.

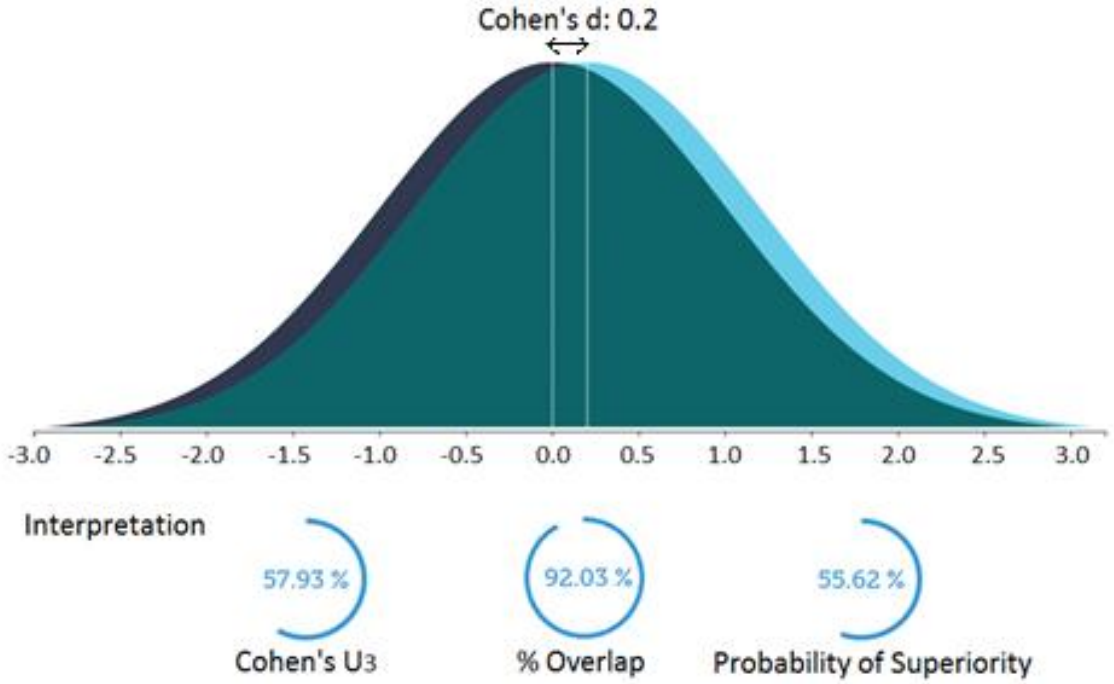
Effect Size	Very Small	Small	Moderate	Large
Description	Little or no noticeable difference; e.g., difference in average heights of girls aged 15 & 15½, reducing bullying from 60% to 55%.	The difference is noticeable but small; e.g., average heights of girls aged 15 and 16, or reducing bullying from 60% to 50%.	The difference is 'visible to the naked eye of a careful observer'; e.g., average heights of girls aged 14 and 18, or reducing bullying from 60% to 30%.	The difference is obvious and 'grossly perceptible'; e.g., average heights of 10 vs. 18 year old girls, or reducing bullying from 60% to 10%
Frequency Difference *Absolute*	<10%	10%	30%	≥50%
Odds Ratio	<1.5	1.5	3.5	≥9
Correlation (r)	<0.1	0.1	0.3	0.5
Percentage of Variance Explained (r^2)	<1%	1%	9%	≥25%
Eta squared (η^2)[xxi] *Multiple regression*	<0.02	0.02	0.13	≥0.26
Standardized Difference *Cohen's* d	<0.2	0.2	0.5	≥0.8

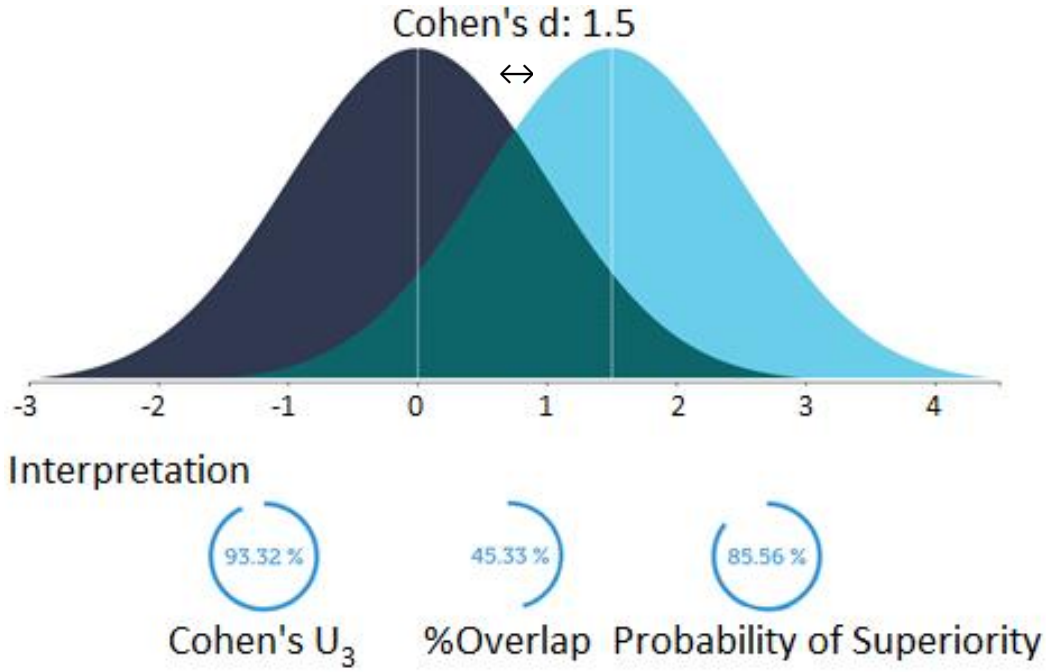
[xxi] Analogous to r^2

Cohen's d

Cohen's *d* is the most common estimate of effect size. It describes the amount of overlap between two distributions of scores. The less overlap, the bigger the difference between the groups. The standard thresholds for Cohen's d are 0.2 for a small effect, 0.5 for a moderate effect, and 0.8 for a large effect[48]. Let's take a closer look at a small effect (0.2) and a large effect (1.5). These two distributions represent bullying scores in two groups that have been standardized so we can compare them. The black distribution shows bullying scores in the participants who have been exposed to the program, the light blue distribution denotes scores in the control group, and the large green area represents the area of overlap between the two distributions.

Figure 1. Understanding Cohen's D. Adapted from Magnusson[49]





In the case of the small effect size, *d*= 0.2, only 58% of students in the control group will have bullying scores that are higher than the mean of the treatment group (*Cohen's U_3*); we would expect that value to be 50% if there was no effect and the distributions overlapped completely. The distributions are very close, with 92% overlap, and there is a 56% chance that a student picked at random from the control group would have a higher bullying score than a student picked at random from the program group (*probability of superiority*). Overall the two distributions are looking pretty similar, and it would certainly be difficult to justify effects even smaller than 0.2, which is the lowest threshold for a 'small' effect. In the case of the large effect size, *d*= 1.5, there is only 45% overlap between the two distributions, 93% of students in the control group will score higher on bullying than the average for the treatment group, and there is an 86% chance that a someone picked at random from the control group will have a higher bullying score than someone picked at random from the treatment group. This is a difference that you would notice, and with high consistency. In real life, a *d* of 1.5 is represented by the difference in average heights of men and women. A *d* of 0.2 is representative of the difference of average heights between 15 and 16 year old girls, and, unfortunately, the typical size of those relatively few statistically-significant effects of bullying prevention programs.

References

1. Olweus D. *Bullying at School: What We Know and What We Can Do*. New York: Blackwell; 1993.
2. Department of Education. Legal definitions for bullying, cyberbullying ensures consistency. http://novascotia.ca/news/release/?id=20130208006. Published 2013.
3. Klomek AB, Sourander A, Kumpulainen K, et al. Childhood bullying as a risk for later depression and suicidal ideation among Finnish males. *J Affect Disord*. 2008;109(1-2):47-55. doi:10.1016/j.jad.2007.12.226.
4. Sourander A, Ronning J, Brunstein-Klomek A, et al. Childhood bullying behavior and later psychiatric hospital and psychopharmacologic treatment: findings from the Finnish 1981 birth cohort study. *Arch Gen Psychiatry*. 2009;66(9):1005-1012. doi:10.1001/archgenpsychiatry.2009.122.
5. Klomek AB, Marrocco F, Kleinman M, Schonfeld IS, Gould MS. Peer victimization, depression, and suicidiality in adolescents. *Suicide Life Threat Behav*. 2008;38(2):166-180. doi:10.1521/suli.2008.38.2.166.
6. Sweeting H, Young R, West P, Der G. Peer victimization and depression in early-mid adolescence: a longitudinal study. *Br J Educ Psychol*. 2006;76(Pt 3):577-594. doi:10.1348/000709905X49890.
7. Rivers I, Poteat VP, Noret N, Ashurst N. Observing bullying at school: The mental health implications of witness status. *Sch Psychol Q*. 2009;24(4):211-223. doi:10.1037/a0018164.
8. LeBlanc JC, Parkington KP, Varasarathan N, Donato A, Bilsbury T. *Social and Emotional Learning Programs for Schools*. Halifax, Nova Scotia; 2013.
9. Mishara BL, Ystgaard M. Effectiveness of a mental health promotion program to improve coping skills in young children: Zippy's Friends. *Early Child Res Q*. 2006;21:110-123. doi:10.1016/j.ecresq.2006.01.002.
10. Cornell DG, Brockenbrough K. Identification of bullies and victims. *J Sch Violence*. 2004;3:63-87. doi:10.1300/J202v03n02_05.
11. Branson CE, Cornell DG. A Comparison of Self and Peer Reports in the Assessment of Middle School Bullying. *J Appl Sch Psychol*. 2009;25:5-27. doi:10.1080/15377900802484133.
12. Juvonen J, Nishina A, Graham S. Self-views versus peer perceptions of victim status among early adolescents. In: Juvonen J, Graham S, eds. *Peer Harassment in School*. New York: The Guilford Press; 2001:105-124.
13. Schroeder B, Messina A, Schroeder D, et al. The implementation of a statewide bullying prevention program: Preliminary findings from the field and the importance of coalitions. *Health Promot Pract*. 2012;13(4):489-495. doi:10.1177/1524839910386887.
14. Bauer NS, Lozano P, Rivara FP. The effectiveness of the Olweus Bullying Prevention Program in public middle schools: a controlled trial. *J*

Adolesc Health. 2007;40(3):266-274. doi:10.1016/j.jadohealth.2006.10.005.

15. Bowllan NM. Implementation and evaluation of a comprehensive, school-wide bullying prevention program in an urban/ suburban middle school. *J Sch Health*. 2011;81(4):167-173. doi:10.1111/j.1746-1561.2010.00576.x.

16. PREVNet, Substance Abuse and Mental Health Services Administration. *Bullying Prevention and Intervention*. http://www.prevnet.ca/research/fact-sheets/bullying-prevention-and-intervention.

17. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Vol 2nd.; 1988. doi:10.1234/12345678.

18. Watson P. Rules of thumb on magnitudes of effect sizes. MRC CBSU Wiki. http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize. Published 2014.

19. Brozek JL, Akl EA, Alonso-Coello P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions. *Allergy*. 2009;64(5):669-677. doi:10.1111/j.1398-9995.2009.01973.x.

20. Organizations that have endorsed or that are using GRADE. Grade Working Group. http://www.gradeworkinggroup.org/society/. Published 2014.

21. Beran TN, Tutty L, Steinrath G. An Evaluation of a Bullying Prevention Program for Elementary Schools. *Can J Sch Psychol*. 2004;19(1-2):99-116. doi:10.1177/082957350401900105.

22. Menard S, Grotpeter J, Gianola D, O'Neal M. *Evaluation of Bullyproofing Your School: Final Report*.; 2008. https://www.ncjrs.gov/pdffiles1/nij/grants/221087.pdf.

23. Cross D, Monks H, Hall M, et al. Three year results of the Friendly Schools whole of school intervention on children's bullying behaviour. *Br Educ Res J*. 2011;37(1):105-129. doi:10.1080/01411920903420024.

24. Cross D, Waters S, Pearce N, et al. The Friendly Schools Friendly Families programme: Three-year bullying behaviour outcomes in primary school children. *Int J Educ Res*. 2012;53:394-406. doi:10.1016/j.ijer.2012.05.004.

25. Kärnä A, Voeten M, Little TD, Poskiparta E, Kaljonen A, Salmivalli C. A large-scale evaluation of the KiVa antibullying program: grades 4-6. *Child Dev*. 2011;82(1):311-330. doi:10.1111/j.1467-8624.2010.01557.x.

26. Kärnä A, Voeten M, Little TD, Alanen E, Poskiparta E, Salmivalli C. Effectiveness of the KiVa Antibullying Program: Grades 1–3 and 7–9. *J Educ Psychol*. 2013;105(2):535-551. doi:10.1037/a0030417.

27. Kärnä A, Voeten M, Little TD, Poskiparta E, Alanen E, Salmivalli C. Going to scale: A nonrandomized nationwide trial of the KiVa antibullying program for grades 1-9. *J Consult Clin Psychol*. 2011;79(6):796-805. doi:10.1037/a0025740.

28. Salmivalli C, Karna a., Poskiparta E. Counteracting bullying in Finland: The KiVa program and its effects on different forms of being bullied. *Int J Behav Dev*. 2011;35(5):405-411. doi:10.1177/0165025411407457.

29. Williford A, Elledge LC, Boulton AJ, DePaolis KJ, Little TD, Salmivalli C. Effects of the KiVa antibullying program on cyberbullying and cybervictimization frequency among Finnish youth. *J Clin Child Adolesc Psychol*. 2013;42(6):820-833. doi:10.1080/15374416.2013.787623.

30. Frey KS, Hirschstein MK, Snell JL, Edstrom LVS, MacKenzie EP, Broderick CJ. Reducing playground bullying and supporting beliefs: An experimental trial of the steps to respect program. *Dev Psychol*. 2005;41:479-490. doi:10.1037/0012-1649.41.3.479.

31. Brown EC, Low S, Smith BH, Haggerty KP. Outcomes from a school-randomized controlled trial of Steps to Respect: A Bullying Prevention Program. *School Psych Rev*. 2011;40(3):423-443.

32. Frey KS, Hirschstein MK, Edstrom L V., Snell JL. Observed reductions in school bullying, nonbullying aggression, and destructive bystander behavior: A longitudinal evaluation. *J Educ Psychol*. 2009;101(2):466-481. doi:10.1037/a0013839.

33. Hirschstein MK, Edstrom LVS, Frey KS, Snell JL, Mackenzie EP, Children C. Walking the talk in bullying prevention: Teacher implementation variables related to initial impact of the Steps to Respect Program. 2007;36:3-21.

34. Low S, Van Ryzin MJ, Brown EC, Smith BH, Haggerty KP. Engagement matters: Lessons from assessing classroom implementation of Steps to respect: A Bullying Prevention Program over a one-year period. *Prev Sci*. 2013;15(2):165-176. doi:10.1007/s11121-012-0359-1.

35. Espelage DL, Low S, Polanin JR, Brown EC. The impact of a middle school program to reduce aggression, victimization, and sexual violence. *J Adolesc Health*. 2013;53(2):180-186. doi:10.1016/j.jadohealth.2013.02.021.

36. Leadbeater B, Hoglund W, Woods T. Changing contexts? The effects of a primary prevention program on classroom levels of peer relational and physical victimization. *J Community Psychol*. 2003;31:397-418. doi:10.1002/jcop.10057.

37. Giesbrecht GF, Leadbeater BJ, Macdonald SWS. Child and context characteristics in trajectories of physical and relational victimization among early elementary school children. *Dev Psychopathol*. 2011;23(1):239-252. doi:10.1017/S0954579410000763.

38. Hoglund WLG, Hosan NE, Leadbeater BJ. Using your WITS: A 6-year follow-up of a peer victimization prevention program. *School Psych Rev*. 2012;41(2):193-214.

39. Leadbeater B, Sukhawathanakul P. Multicomponent programs for reducing peer victimization in early elementary school: A longitudinal evaluation of the WITS Primary Program. *J Community Psychol*. 2011;39(5):606-620. doi:10.1002/jcop.

40. Peele S. The truth we won't admit: Drinking is healthy. The Pacific Standard. http://www.psmag.com/navigation/health-and-behavior/truth-wont-admit-drinking-healthy-87891/. Published 2014.

41. Roerecke M, Rehm J. The cardioprotective association of average alcohol consumption and ischaemic heart disease: a systematic review and meta-analysis. *Addiction*. 2012;107:1246-1260. doi:10.1111/j.1360-0443.2012.03780.x.

42. Fillmore KM, Stockwell T, Chikritzhs T, Bostrom A, Kerr W. Moderate Alcohol Use and Reduced Mortality Risk: Systematic Error in Prospective Studies and New Hypotheses. *Ann Epidemiol*. 2007;17. doi:10.1016/j.annepidem.2007.01.005.

43. Hansel B, Kontush A, Bruckert E. Is a cardioprotective action of alcohol a myth? *Curr Opin Cardiol*. 2012;27:550-555. doi:10.1097/HCO.0b013e328356dc30.

44. Snow DL, Tebes JK. Experimental and quasi-experimental designs in prevention research. *NIDA Res Monogr*. 1991;107:140-158. doi:10.1016/0306-4573(84)90053-0.

45. Olweus D, Limber SP. The Olweus Bullying Prevention Program: Implementation and evaluation over two decades. In: *Handbook of Bullying in Schools: An International Perspective*. ; 2010:377-401.

46. Brame B, Nagin DS, Tremblay RE. Developmental trajectories of physical aggression from school entry to late adolescence. *J Child Psychol Psychiatry*. 2001;42:503-512. doi:http://dx.doi.org/10.1017/S0021963001007120.

47. PREVNet, US Department of Health and Human Services. *Age Trends in the Prevalence of Bullying*. Kingston, Ontario www.prevnet.ca/sites/prevnet.ca/files/fact-sheet/PREVNet-SAMHSA-Factsheet-Age-Trends-in-the-Prevalence-of-Bullying.pdf.

48. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Vol 2nd.; 1988. doi:10.1234/12345678.

49. Magnusson K. Interpreting Cohen's d effect size: An interactive visualization. R Psychologist. rpsychologist.com/d3/cohend. Published 2014.

50. Garrity C, Jens K, Porter W, Sager N, Short-Camilli C. *Bully-Proofing Your School: A Comprehensive Approach.* Vol 5Reclaimin. 2nd ed. Longmont: Sopris West; 2000.

51. Dare to Care. *Dare to Care Program Costs*.; 2014.

52. Low S, Frey KS, Brockman CJ. Gossip on the playground: Changes associated with universal intervention, retaliation beliefs, and supportive friends. 2010;39(4):536-551.