A Meta-Analysis of School-Based Bullying Prevention Programs’ Effects on Bystander Intervention Behavior

Joshua R. Polanin
Loyola University Chicago

Dorothy L. Espelage
University of Illinois Urbana—Champaign

Therese D. Pigott
Loyola University Chicago

Abstract. This meta-analysis synthesized bullying prevention programs’ effectiveness at increasing bystander intervention in bullying situations. Evidence from 12 school-based programs, involving 12,874 students, indicated that overall the programs were successful (Hedges’s $g = .20$, 95% confidence interval [CI] = .11 to .29, $p < .001$), with larger effects for high school (HS) samples compared to kindergarten through eighth-grade (K-8) student samples (HS effect size [ES] = 0.43, K-8 ES = 0.14; $p < .05$). A secondary synthesis from eight of the studies that reported empathy for the victim revealed treatment effectiveness that was positive but not significantly different from zero ($g = .05$, 95% CI = -.07 to .17, $p = .45$). Nevertheless, this meta-analysis indicated that programs increased bystander intervention both on a practical and statistically significant level. These results suggest that researchers and school administrators should consider implementing programs that focus on bystander intervention behavior supplementary to bullying prevention programs.

Bullying perpetration often occurs when bystanders are present (Hawkins, Pepler, & Craig, 2001; Lagerspetz, Bjorkqvist, Bertz, & King, 1982). In fact, some research has indicated that more than 80% of the time an observer witnesses victimization (O’Connell, Pepler, & Craig, 1999). Despite the presence of witnesses and bystanders, nearly 1 in 3 children report victimization by a bully in the past 2 months (Frey, Hirschstein, Edstrom, & Snell, 2009; Nansel et al., 2001; Wang, Ian­notti, & Nansel, 2009). Consequentially, bullying occurs with an audience of members who play multiple roles (Salmavalli, Lagerspetz, Bjorkqvist, Osterman, & Kaukianen, 1996) and often fail to intervene on behalf of the victim with regularity. These bullying incidents have lasting negative effects on the bully, victim, and bystanders (Olweus, 2002; Swearer, Espelage, Villancourt, & Hymel, 2010; Sweeting, Young, West, & Der, 2006; Stevens, Oost, & Bourdeaudhuij, 2004).

Correspondence regarding this article should be addressed to Joshua R. Polanin, Loyola University Chicago, 820 N. Michigan Ave., Suite 1022C, Chicago, IL 60611; e-mail: jpolanin@luc.edu

Copyright 2012 by the National Association of School Psychologists, ISSN 0279-6015
The past 20 years have seen a burgeoning of bullying prevention programs (Ferguson, Miguel, Kilburn, & Sanchez, 2007; Ryan & Smith, 2009; Ttofi & Farrington, 2009, 2011). Researchers, school administrators, and teachers have used myriad designs, theories, and techniques in an attempt to mitigate the prevalence of bullying (Astor, Meyer, Benbenishty, Marachi, & Rosemond, 2005). Ttofi and Farrington’s (2011) recent large-scale meta-analysis of over 90 studies found that the majority of these programs have been successful at slowing the rate of bullying.

Although successful bullying programs remain important accomplishments, Ttofi and Farrington (2011) found that few programs specifically target the behavior of bystanders (i.e., an individual who witnesses bullying). As such, prevention programs de-emphasize a population that constitutes between 60% and 70% of primary or secondary school students (Glew, Fan, Katon, Rivara, & Kerntic, 2005; Rivers, Poteat, Noret, & Ashurst, 2009). This program oversight is unfortunate because observational research has found that when bystanders intervene on behalf of the victim, they successfully abate victimization more than 50% of the time (Craig, Pepler, & Atlas, 2000; O’Connell et al., 1999).

Supported by the knowledge that bystanders can successfully intervene on behalf of the victim, a small amount of literature has focused recently on increasing this behavior. These programs explicitly emphasize the importance of bystander intervention behavior and measure this construct. Given these conditions, the purpose of this meta-analysis is to synthesize school-based bullying prevention programs’ effectiveness to change bystander intervention behavior. We also aggregated the program’s influence on empathy for the victim as a secondary synthesis because it has received recent investigation (Gini, Albiero, Benelli, & Altoe, 2007). The following summarizes the relevant literature, provides a comprehensive examination of the synthesis process and quantitative analysis outcomes, and elucidates moderator analysis and publication bias results. Suggestions for future research and policy are also provided.

Bullying in the Schools

Definition

Olweus (1973) first described bullying as “mobbing” where a group or individual teases or harasses another individual. As such, early research focused solely on the physical aspects of school environment (e.g., teacher–student ratio), but found little connection to perpetration or victimization (Swearer et al., 2010). Recently, Frey et al. (2009) described bullying as a social construct that disrupts social connections among students. Ross and Horner (2009) summarized the plethora of definitions:

Common definitions of bullying involve repeated acts of aggression, intimidations, or coercion against a victim who is weaker in terms of physical size, psychological or social power, or other factors that result in a notable power differential. (p. 748)

The bully construct has received much research and hosts of definitions remain. Taken together, the bully generally involves an individual or group who incites physical or emotional abuse on another individual or group. Although other research exists on Internet and workplace bullying (Mishna, Cook, Saini, Wu, & MacFadden, 2010), this review focuses on school bullying.

Prevalence and Negative Effects

School bullying is not a problem local to the United States; rather, it is recognized worldwide (Espelage & Swearer, 2003). Reports from Europe to North America have indicated that anywhere from 1% to 50% of students had been bullied or victimized within the last 2 months (Wang et al., 2009). Some observations have shown that as many as 30% of students were involved in bullying as either the bully or victim (Frey et al., 2009). Hymel and Swearer (2010) recently reported that 35% of students indicated being bullied at least once in the last 2 months with as many as 11% of those sampled reported being bullied more than 2 or 3 times in the last 2 months. Moreover, some research has found that bullying roles remain relatively stable across time. In a
sample of 516 middle school students, Espelage, Bosworth, and Simon (2001) found that individuals who perpetrated bullying continued to across multiple years. This finding has been replicated from other settings and populations (McDougal, Hymel, & Vaillancourt, 2009; Scholte, Engels, Overbeek, Kemp, & Haselager, 2007; Sourander, Helstela, Hele- nius, & Piha, 2000).

Although bullying’s pervasiveness is cause for concern, the negative consequences related to bullying remain tantamount. Researchers have indicated that bullying has been linked to anger and misconduct (Bosworth, Espelage, & Simon, 1999), criminal or delinquent behavior (Olweus, 2002), and suicidal ideations (Kaltiala-Heino, Rimpela, Marttunen, Rimpela, & Rantanen, 1999). Victimization, on the other hand, has been linked to poor physical health (Rigby, 1999), low self-esteem (Rigby & Slee, 1993), depression (Sweeting et al., 2006), anxiety (Craig, 1994), and school avoidance (Kochenderfer & Ladd, 1996). Indeed, the negative effects of bullying and being bullied are persistent and problematic.

**Bystander**

**Definition, Roles, and Negative Effects**

A significant proportion of individuals within school systems are considered individuals who are bystanders of bullying (Glew et al., 2005). Twemlow, Fonagy, and Sacco (2004) defined a bystander as an individual who lacks participation in bullying scenarios as either the bully or victim. The bystander may actively intervene to stop the bully, encourage the bully to continue, or view bullying passively; bystanders can be either boys or girls (Cowie, 2000; Smith, Twemlow, & Hoover, 1998).

There are specific roles that the bystander can demonstrate. Some authors refer to the bystander as a passerby, observer, witness, or participant (Salmivalli, Kaukinainen, & Voeten, 2005; Twemlow et al., 2004); others described roles in relation to sustaining or preventing the bullying behavior such as reinforcer (e.g., laughing or seeing what is happening), assistant (e.g., follower of the bully), defender (e.g., being supportive of the victim), or outsider (e.g., remaining away from the bullying situation; Salmivalli et al., 1996). For the purposes of this review, we defined the bystander generally as any student who witnessed a bullying episode, with the operationalizing characteristic being the witnessing presence of bullying, regardless of other characteristics.

Despite these literature discordances, few disagree about the adverse effects witnessing bullying can have on the bystander. Bystanders felt significantly more uncomfortable in bullying situations compared to bullies (Stevens et al., 2004), and reported feelings of anxiety and insecurity (Rigby & Slee, 1993). This anxiety due to witnessing bullying has been linked to aggressive retaliation (Musher-Eizenman et al., 2004), and the fear of being bullied often prevented bystanders from seeking adult help (Unnever & Cornell, 2003). A recent large-scale study conducted in the United Kingdom found that compared to perpetrators, bystanders were at elevated risk for nonclinical outcomes (i.e., interpersonal sensitivity), and compared to victims bystanders were more likely to have elevated levels of substance abuse (Rivers et al., 2009).

**Evidence of the Bystanders’ Effects on Bullying**

Individuals who are bystanders remain present more than 80% of bullying situations (O’Connell et al., 1999), and therefore some research has focused on a social-ecological model of bullying prevention and intervention (Frey et al., 2009; Swearer & Espelage, 2004). The social presence and pervasiveness of the bystander fosters myriad opportunities to intervene. For example, bystanders supported victims by reporting bullies to adults when participating in a setting specifically designed to change bullying behavior patterns through bystanders (Sharp, Sellors, & Cowie, 1994). Ross and Horner (2009) recently implemented a school-wide bullying intervention program that resulted in a decrease in reinforcing bystander behavior and bullying perpetration.
overall. Moreover, interventions that focused on dealing with conflict through peers instead of direct interventions with adults led to positive effects (Cowie & Hutson, 2006), and an individual’s willingness to intervene in bullying situations was inversely related to the amount of peer-group bullying perpetration (Espelage, Green, & Polanin, 2011).

Bystander Intervention Program Characteristics

To date, best practice guidelines to promote effective bystander intervention behaviors remain undefined because research findings varied widely with regard to their implementation focuses and approaches. Several mediums for interventions have been studied to teach children about bystander behavior, including classroom-based drama (Merrell, 2004), media such as videotaped reenactments (McLaughlin, 2009; Schumacher, 2007), and individualized computer-adaptive software to track students’ progress within social scenarios and provided feedback on effective bystander behavior (Evers, Prochaska, Vinters, Johnson, & Prochaska, 2007). However, all of these programs focus on bystander behavior perhaps, because there seems to be some support for targeting peer-group behaviors to mitigate individual bullying (Salmivalli et al., 1996). Peer-group interventions often encourage bystander intervention (Andreou, Didaskalou, & Vlachou, 2008; Frey et al., 2009; Stevens et al., 2000) or enhance bystander empathy for the victim (Gini et al., 2007; Nickerson, Mele, & Princiotta, 2008). However, few studies have examined the effects of peer-group interventions on previctim empathy (Merrell, Gueldner, Ross, & Isaya, 2008) and there are resulting guidelines to promote this behavior.

Previous Meta-Analyses on Bullying Prevention Programs

A number of recent quantitative meta-analyses (Ferguson et al., 2007; Ttofi & Farrington, 2009, 2011; Merrell et al., 2008; Smith, Schneider, Smith, & Ananiaidou, 2004) and qualitative systematic reviews (Ryan & Smith, 2009) have been conducted regarding bullying and victimization intervention and prevention programs. However, none of these meta-analyses focused specifically on bystander intervention constructs. Merrell et al.’s (2008) review included three studies that measured “intervene to stop bullying behavior,” which resulted in a mean effect size of 0.17, but this was a secondary analysis of a small number of studies. Therefore, the goal of the present study was to conduct a meta-analysis that would directly address bystander intervention behavior and empathy attitudes.

Given this goal, two primary research questions are addressed:

1. What is the average treatment effect, across the current literature, of bullying prevention programs on bystander intervention behavior?
2. What study characteristics produced the largest treatment effect?

A secondary research question addressed bystander empathy for the victim:

3. What is the average treatment effect, across the current literature, of bullying prevention programs on bystander empathy for the victim?

Method

Search Strategy

We used a comprehensive search to retrieve articles from the international research literature within the last 30 years (1980–2010). We searched primarily five online databases: Dissertation Abstracts International, Education Resources Information Center (ERIC), PsycINFO, Medline, and Science Direct. Combinations of the following terms were used: “bystander or participant or defender or other,” “bully or victim,” “school, school program, or program,” “prevention or intervention,” “aggression,” and “not higher education or not cyber-bully.” To ensure that the identified studies focused on bystander behavior as the primary goal, these terms were searched in the abstract of the study. In addi-
tion, we searched the bibliographies of all articles selected for relevant studies.

The search retrieved 360 total articles, but only 83 were unique and compared to the criteria listed in the next section. Of those 83, 53 were deemed irrelevant, 13 did not address the intervene construct, 6 failed to include a control group, and 1 was a repeat of a previous study. Finally, we corresponded with a number of experts in the field to ensure inclusion of all relevant articles. This correspondence produced 1 relevant study. Hence, we reviewed and included 11 studies total.

Criteria for Considering Studies for Review

The present study focused on school-based interventions that emphasized changing the bystander’s intervention behavior. To assess the effects of these programs, we collected peer-reviewed studies published or conducted from 1980 to 2010, based solely within a school system and intended purposefully to modify bystander intervention behavior. Subsequently, we excluded studies that focused on changing bullying behaviors primarily and collected a bystander measure only as a secondary procedure. The review included interventions from the United States and Europe, but we limited inclusion to English-written studies.

We reviewed studies that included participants from the kindergarten through 12th-grade population, but interventions with school-aged children based outside the school setting were excluded. In addition, we attempted to collect studies that included “at-risk” students and the general population, but none of the studies distinguished between these populations. It should be mentioned that one study attempted to deconstruct the bystander into several types of bystanders to observe treatment effects (Evers et al., 2007). Although the deconstruction was informative, it was the only study to implement such a procedure. As such, for that study we used the average intervention effects across all bystander types.

Furthermore, we included only studies that used a treatment-control research design. These designs included true experimental randomly assigned groups, nonrandom quasi-experimental designs, and nonrandomly assigned matched group. We also included all control group types; these included wait list, treatment-as-usual, and “straw-man.” However, single-group pre/post-test (e.g., gain scores) and cohort designs were excluded.

Outcomes

Studies must have included a bystander intervention measure. We operationalized this outcome as a measure that assessed the contribution of the bystander to a bullying situation (Frey et al., 2005, 2009). Therefore, we included studies that measured intention to intervene, intention to stop bullying, direct intervention, or conversely, difficulty in responding assertively to a bullying situation. For example, Andreou et al. (2008) included items that assessed students’ intention to intervene on a 5-point Likert scale. We included items that concerned students’ intention to “seek teacher’s help,” “react against bullying,” and “support the victims of bullying” (p. 241). Table 1 provides the measures used, the study’s stated construct, and the number of items combined to create the measure.

In addition to the intervention outcome, we collected results of the program’s effects on changes in attitudes of empathy toward the victim. We operationalized the empathy outcome as a measure that indicated empathy for the victim. For instance, Stevens et al. (2000) used an empathy measure that included “feeling sad about students who are bullied” and “unpleasantness when another student is being bullied” (p. 26). Measures that included items similar to this construct constituted an empathy scale. We should also mention that this outcome constituted a secondary outcome measure and thus was not a criterion for synthesis inclusion. Studies that included a bystander intervention behavior measure but failed to include a measure of empathy were included.
Table 1
Characteristics of the Studies used in the Meta-Analysis

<table>
<thead>
<tr>
<th>Study (DoP)</th>
<th>Type</th>
<th>Location</th>
<th>N (%Male)</th>
<th>Grade Setting</th>
<th>Program Title</th>
<th>Program Characteristics</th>
<th>Facilitator; (Length)</th>
<th>Intervention Measurement</th>
<th>Design</th>
<th>IN ES (95% CI)</th>
<th>EM ES (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andreou (2008)</td>
<td>J</td>
<td>Greece</td>
<td>418 (60)</td>
<td>4&lt;sup&gt;th&lt;/sup&gt;-6&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Urban-Based Anti-Bullying</td>
<td>Awareness-building (Awareness); Self-reflection; Behavior modification (BehMod)</td>
<td>Teacher (1)</td>
<td>Bystander intention to intervene</td>
<td>Q</td>
<td>-.01</td>
<td>-.19</td>
</tr>
<tr>
<td>Evers (2007)</td>
<td>J</td>
<td>Multiple US States</td>
<td>710 (41)</td>
<td>9&lt;sup&gt;th&lt;/sup&gt;-12&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Mixed</td>
<td>Build Respect</td>
<td>BehMod; Individualized computer software; Parent component (Parent)</td>
<td>Computer Based (2)</td>
<td>Passive bystanding</td>
<td>Q</td>
<td>.46</td>
</tr>
<tr>
<td>Fonagy (2009)</td>
<td>J</td>
<td>Kansas</td>
<td>578 (46)</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;-5&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Rural</td>
<td>CAPSLE</td>
<td>Psychiatric consultation; Awareness; Parent</td>
<td>Teacher (12)</td>
<td>Helpful bystanding</td>
<td>E</td>
<td>.05</td>
</tr>
<tr>
<td>Frey (2005)</td>
<td>J</td>
<td>Washington</td>
<td>913 (51)</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;-6&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Suburban</td>
<td>Steps to Respect</td>
<td>Awareness; Social-emotional skill-building; Parent</td>
<td>Teacher (6)</td>
<td>Bystander responsibility</td>
<td>E</td>
<td>.11</td>
</tr>
<tr>
<td>Karna (2011)</td>
<td>J</td>
<td>Finland</td>
<td>8166 (50)</td>
<td>4&lt;sup&gt;th&lt;/sup&gt;-6&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Mixed</td>
<td>KiVa</td>
<td>Awareness; Role-playing; Modeling; Parent</td>
<td>Teacher (12)</td>
<td>Bystander defending</td>
<td>E</td>
<td>.14</td>
</tr>
<tr>
<td>McLaughlin (2009)</td>
<td>D</td>
<td>Ohio</td>
<td>41 (41)</td>
<td>6&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Urban</td>
<td>Effective Bully Prevention</td>
<td>Awareness; Modeling with media (ModMed)</td>
<td>Researcher (1)</td>
<td>Bystander traits</td>
<td>Q</td>
<td>.21</td>
</tr>
<tr>
<td>Menesini (2003)</td>
<td>J</td>
<td>Italy</td>
<td>293 (53)</td>
<td>6&lt;sup&gt;th&lt;/sup&gt;-8&lt;sup&gt;th&lt;/sup&gt;</td>
<td>Urban</td>
<td>Befriending intervention</td>
<td>Awareness; Enhanced capacity; Responsibility-training</td>
<td>Not Reported (12)</td>
<td>Defender traits</td>
<td>Q</td>
<td>.03</td>
</tr>
</tbody>
</table>

(Table 1 continues)
<table>
<thead>
<tr>
<th>Study (DoP)</th>
<th>Type</th>
<th>Location</th>
<th>N (% Male)</th>
<th>Grade Setting</th>
<th>Program Title</th>
<th>Program Characteristics</th>
<th>Facilitator; (Length)</th>
<th>Intervention Measurement</th>
<th>Design</th>
<th>IN ES (95% CI)</th>
<th>EM ES (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merrell</td>
<td>D</td>
<td>New York</td>
<td>56 (30)</td>
<td>9th Urban</td>
<td>5 W’s Approach to Bullying Awareness; ModMed</td>
<td>Researcher (2)</td>
<td>Bystander intention to intervene</td>
<td>E</td>
<td>.60</td>
<td>(−1.17, 2.36)</td>
<td>Not Measured</td>
</tr>
<tr>
<td>Schumacher</td>
<td>M</td>
<td>Pennsylvania</td>
<td>825 (42)</td>
<td>9th-12th Mixed</td>
<td>Bullying Video Program Awareness; ModMed</td>
<td>Researcher (1)</td>
<td>Bystander intention to intervene</td>
<td>E</td>
<td>.43</td>
<td>(.29, .57)</td>
<td>.29</td>
</tr>
<tr>
<td>Stevens “A”</td>
<td>J</td>
<td>Belgium</td>
<td>301 (50*)</td>
<td>4th-8th Not Reported</td>
<td>Anti-bullying Intervention Social-cognitive training; ModMed</td>
<td>Teacher (1)</td>
<td>Bystander intention to intervene</td>
<td>E</td>
<td>.06</td>
<td>(−.17, .29)</td>
<td>−.05</td>
</tr>
<tr>
<td>Stevens “B”</td>
<td>J</td>
<td>Belgium</td>
<td>401 (50*)</td>
<td>9th-11th Not Reported</td>
<td>Anti-bullying Intervention Social-cognitive training; ModMed</td>
<td>Teacher (1)</td>
<td>Bystander intention to intervene</td>
<td>E</td>
<td>.39</td>
<td>(.19, .59)</td>
<td>.14</td>
</tr>
<tr>
<td>Whitaker</td>
<td>B</td>
<td>Texas</td>
<td>1763 (50)</td>
<td>5th Mixed</td>
<td>Expect Respect Awareness; Psycho-education; Parent</td>
<td>Teacher (12)</td>
<td>Bystander intention to intervene</td>
<td>E</td>
<td>.25</td>
<td>(.15, .34)</td>
<td>Not Measured</td>
</tr>
</tbody>
</table>

Notes: DoP = Date of Publication; J = Journal Article; D = Dissertation; M = Master’s Thesis; B = Book Chapter; * = imputed percent male; Parent = Parent Training; Length = Time to Posttest in months; Q = Quasi-experimental; E = Experimental; T = Treatment groups; C = Control groups; IN ES = Intervention Effect Size; EM ES = Empathy Effect Size.
As shown in Table 1, 10 of the 83 unique articles met the inclusion criteria and were included in the meta-analysis. We discontinued the literature search on May 20, 2010, but added 1 article brought to our attention 3 months later, which brought the total number to 11 studies.

Coding

Study details, appropriate program, and sample information were coded directly into an EXCEL (2010) database. This included publication year, publication type, funding provided, country of origin and publication, program location, treatment and control sample characteristics (e.g., age, gender, race, SES, disabilities), program characteristics (e.g., length of time, intervention details), and program facilitator. In addition, intervention and empathy outcome measures were transcribed. By coding directly into the EXCEL database, we eliminated errors that might have occurred during the normal transcription phase (Lipsey & Wilson, 2001).

The first author coded all 12 studies, but one independent rater coded a randomly selected portion of studies (5) for reliability purposes. The two raters agreed 92% of the time. The coders came to an agreement for all discrepancies prior to completion.

Analysis

Quantitative synthesis, or meta-analysis, is a statistical technique that combines related research studies to estimate an overall treatment effect (Cooper, Hedges, & Valentine, 2009; Glass, 1976; Hedges & Olkin, 1985). Often, and in the case of the present review, meta-analysis aggregates treatment effect sizes to assess an intervention’s effectiveness. The purpose of a meta-analysis, then, is to generalize findings across multiple treatment and setting types, participants, and times (Matt & Cook, 2009). We conducted analyses using SPSS (2010) and Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein, 2005) software.

Independent findings. A paramount meta-analytic assumption is independence of findings. Cooper (2010) discussed several occurrences that constitute nonindependence and their effects on subsequent findings. Thus, we conducted several common procedures to ensure independent findings. To ensure that only one effect size was derived from each intervention, we used only the first treatment outcome reported for studies that reported multiple post-treatment outcomes (see Andreou et al., 2008). If studies implemented interventions with two groups but only one control group, we synthesized the treatment effects prior to calculating the study effect size (see McLaughlin, 2009). Finally, if one author implemented an intervention and published multiple articles on the same sample, then we reviewed only the first article published (see Frey et al., 2005, 2009).

Effect size metrics. The majority of effect sizes calculated used a continuous scale. As such, the appropriate effect size metric was the standardized mean difference (Equation 1):

\[ d = \frac{X_{G1} - X_{G2}}{S_p} \]  

(1)

where the numerator is the mean difference between treatment and control group posttests, and the denominator is the pooled standard deviation for the intervention and comparison groups. Further, all \( d \) metrics were bias corrected using Hedges’s (1981) small sample correction (\( g \)). This correction as well as the sampling variance is represented by Equations 2 and 3:

\[ g = \left[ 1 - \left( \frac{3}{4N - 9} \right) \right] \cdot d \]  

(2)

\[ \text{Var}_g = \frac{n_{G1} + n_{G2}}{n_{G1}n_{G2}} + \frac{g^2}{2(n_{G1} + n_{G2})} \]  

(3)

where \( N \) is the total sample, \( d \) is the original standardized mean difference, and \( n_{G1} \)
and \( n_{g2} \) represent the treatment and control group sample sizes, respectively.

In addition, we calculated logged odds ratio effect sizes for two studies that used a categorical outcome measure (Evers et al., 2007; Merrell, 2004). Both measures were observations of treatment and control participants during free period times. The authors observed how many times treatment children intervened (or intended to intervene) in a bullying situation compared to children in the control group. Standard odds ratio calculations were first used (Sanchez-Meca et al., 2003). We then converted the logged odds ratio into a standardized mean difference as outlined by Lipsey & Wilson (2001) (Equations 4 and 5):

\[
d = \frac{\sqrt{3} \cdot \ln(OR)}{\pi} \quad (4)
\]

\[
SE = \frac{\sqrt{3} \cdot SE_{\ln(OR)}}{\pi^3} \quad (5)
\]

where \( \ln(OR) \) represents the original logged odds ratio and \( SE_{\ln(OR)} \) represents the original sampling variance.

Missing data. Only one of the studies failed to provide appropriate descriptive statistics. Schumacher (2007) provided only a \( t \) statistic, as well as sample sizes, that compared the treatment and control groups. Lipsey and Wilson (2001) provided an appropriate conversion (Equation 6):

\[
ES_{sm} = t \sqrt{\frac{n_1 + n_2}{n_1n_2}} \quad (6)
\]

where \( t \) represents the \( t \) statistic the study provided, \( n_1 \) represents the sample size of the treatment group, and \( n_2 \) represents the sample size of the control group.

Random effects model. We assumed that the treatments were derived from a random sample of the literature but lacked a common effect size. Given this assumption, Borenstein, Hedges, Higgins, and Rothstein (2010) posited that the random effects model was most appropriate. We further assumed that an underlying distribution of effect sizes was plausible; thus our goal, given the random effects framework, was to estimate the distribution’s mean and confidence interval.

To estimate a random effects mean and confidence interval, we calculated the weighted treatment effect of each study. The weighted effect estimation synthesized both within-study error variance and common between-study variance. We represented this weight calculation as Equation 7:

\[
W_i = \frac{1}{V_i + T^2} \quad (7)
\]

where \( W \) represents the \( i \)th study weight, \( V_i \) indicates the within-study error variance, and \( T^2 \) represents the between-study variance. We used these weights then to estimate the combined treatment effect. This can be represented by Equation 8:

\[
M = \frac{\sum W_i \cdot g_i}{\sum W_i} \quad (8)
\]

where \( M \) represents the combined effect, \( W_i \) represents the \( i \)th study weight, and \( g_i \) indicates Hedges’s effect size \( g \) for the \( i \)th study. Further, we calculated confidence intervals and \( p \) values by taking the square root of the inverse of the sum of the weights.

Moderator analysis. A critical next step to the investigation of effect size distribution is moderator analysis. We started by calculating the homogeneity statistic \( Q \). This statistic provided information about the distribution of effect sizes, and a large test statistic (i.e., rejecting the null hypothesis of study homogeneity) indicated that moderator analyses were appropriate (Raudenbush, 2009). Given this statistical confirmation, we used procedures analogous to analysis of variance (ANOVA), where one attempts to model effect-size heterogeneity associated with categorical study-level variables. Further, because of the small number of studies that constituted the review, we calculated the variance compo-
ment across all groups rather than within groups as is generally conducted (Hedges & Vevea, 1998).

Lipsey (2009) discussed three types of independent variables common to meta-analytic practice: extrinsic variables, method variables, and substantive variables. Extrinsic and method variables relate to the study’s dissemination (i.e., published or unpublished) or methodological constraints (i.e., randomized or nonrandomized). Substantive variables, on the other hand, should be regarded as variables of interest and generally include characteristics of the population or treatment. For this review, we coded substantive independent variables to reflect the participant’s age, length of treatment, and treatment type (e.g., individual or group).

Finally, we conducted a moderator analysis procedure analogous to regression, weighted meta-regression. This statistical procedure allows for the simultaneous estimation of study-level effects, but shares the problems of typical regression (Cooper, 2010). For this review, we modeled two independent variables simultaneously, treatment population (categorical variable) and the percent of males in the treatment group.

**Sensitivity analysis.** Two of the reviewed studies contributed a measure that used an odds ratio. To ensure that study findings were not biased by including these measures, we conducted a sensitivity analysis. The analysis consisted of removing the 2 studies (Evers et al., 2007; Merrell, 2004) and recalculating the weighted effect size. We hypothesized that no difference would be found between the two types of measurements.

**Publication bias.** Publication bias remained an important consideration during the literature search and analyses. Rosenthal (1979) introduced the “file-drawer problem,” which stated that studies with small or nonsignificant effect sizes tended to remain unpublished. To combat this problem, we included unpublished works from three dissertations and one master’s thesis (Rothstein, Sutton, & Borenstein, 2005).

To avoid and interpret the overestimation of the random-effects estimate, we used the nonparametric trim and fill procedure (Duval & Tweedie, 2000) to assess the sensitivity of results to publication bias. This procedure estimates the number of publications theoretically missing because of funnel plot asymmetry, and then recalculates the random-effects mean and confidence interval to include the imputed missing studies. We also reported Rosenthal’s fail-safe N (1979) and Egger, Davey Smith, Schnieder, & Minder’s (1997) regression coefficient.

**Results**

**Meta-Analysis Literature**

Table 1 provides characteristics for each study included in the review. The studies reviewed include 7 published journal articles, 1 book chapter, and 3 unpublished papers (two dissertations and one master’s thesis). Seven of the 11 studies were conducted within the United States and the other 4 were conducted in Belgium, Finland, Greece, and Italy, respectively. All studies were completed between 2000 and 2010.

One article contributed two effect sizes because it included two mutually exclusive interventions from two separate populations (Stevens et al., 2000). Therefore, we synthesized a total of 12 interventions. Each of the 12 interventions included a treatment and control group; 4 of the 12 programs used quasi-experimental design and the other 8 used a randomized experimental design. A total of 12,874 students participated in the 12 interventions.

**Outcome Effect Sizes**

**Bystander intervention outcome.** As delineated in Lipsey and Wilson (2001), we estimated the random effects weighted mean by using Equation 8. The results revealed a statistically significant positive weighted average ($g = .20, p < .001, 95\% CI = .11$ to $.29$). In other words, the treatment increased bystander intervention behavior 20% of one standard deviation more than individuals in the
control group. Table 2 provided a forest plot of the random-effects model’s relevant statistics.

**Empathy outcome.** Of the 12 interventions used to calculate the bystander intervention outcome weighted average, 8 included a measure on victim empathy. As previously conducted, we used a random effects model to estimate the weighted treatment mean. The results revealed a very small, nonstatistically significant result ($g = .05, p = .38$) with a confidence interval that included zero (95% CI = −.07 to .17). The small number of studies that included a measure of empathy may not provide enough power to detect a small effect. Therefore, the results of this analysis should be considered inconclusive (see Table 3).

**Moderator Analysis**

We categorized the effect sizes into several relevant groups and conducted the random-effects ANOVA-like analysis (Table 4). A statistically significant $Q$ value indicated appropriate heterogeneity between studies that

**Table 2**

Summary Statistics for Bystander Intervention Effect Sizes

<table>
<thead>
<tr>
<th>Study Name</th>
<th>Hedges’ g</th>
<th>Standard Error</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>Z-Value</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andreou</td>
<td>−.01</td>
<td>.10</td>
<td>−.20</td>
<td>.19</td>
<td>−.04</td>
<td>.97</td>
</tr>
<tr>
<td>Evers</td>
<td>.46</td>
<td>.09</td>
<td>.27</td>
<td>.64</td>
<td>4.88</td>
<td>.01</td>
</tr>
<tr>
<td>Fonagy</td>
<td>.05</td>
<td>.08</td>
<td>−.11</td>
<td>.22</td>
<td>.62</td>
<td>.54</td>
</tr>
<tr>
<td>Frey</td>
<td>.11</td>
<td>.07</td>
<td>−.02</td>
<td>.24</td>
<td>1.68</td>
<td>.09</td>
</tr>
<tr>
<td>Karna</td>
<td>.14</td>
<td>.02</td>
<td>.10</td>
<td>.19</td>
<td>6.51</td>
<td>.01</td>
</tr>
<tr>
<td>McLaughlin</td>
<td>.21</td>
<td>.32</td>
<td>−.42</td>
<td>.83</td>
<td>.65</td>
<td>.52</td>
</tr>
<tr>
<td>Mensini</td>
<td>.03</td>
<td>.12</td>
<td>−.21</td>
<td>.26</td>
<td>.22</td>
<td>.82</td>
</tr>
<tr>
<td>Merrell</td>
<td>.60</td>
<td>.90</td>
<td>−1.17</td>
<td>2.24</td>
<td>.66</td>
<td>.51</td>
</tr>
<tr>
<td>Schumacher</td>
<td>.43</td>
<td>.07</td>
<td>.29</td>
<td>.57</td>
<td>6.08</td>
<td>.01</td>
</tr>
<tr>
<td>Stevens “A”</td>
<td>.06</td>
<td>.12</td>
<td>−.17</td>
<td>.29</td>
<td>.53</td>
<td>.60</td>
</tr>
<tr>
<td>Stevens “B”</td>
<td>.39</td>
<td>.10</td>
<td>.19</td>
<td>.59</td>
<td>3.89</td>
<td>.01</td>
</tr>
<tr>
<td>Whitaker</td>
<td>.25</td>
<td>.05</td>
<td>.15</td>
<td>.34</td>
<td>5.16</td>
<td>.01</td>
</tr>
<tr>
<td>Overall</td>
<td>.20</td>
<td>.04</td>
<td>.11</td>
<td>.28</td>
<td>4.54</td>
<td>.01</td>
</tr>
</tbody>
</table>

**Table 3**

Summary Statistics for Empathy Effect Sizes

<table>
<thead>
<tr>
<th>Study Name</th>
<th>Hedges’ g</th>
<th>Standard Error</th>
<th>Lower Limit</th>
<th>Upper Limit</th>
<th>Z-Value</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andreou</td>
<td>−.19</td>
<td>.10</td>
<td>−.38</td>
<td>.01</td>
<td>−1.90</td>
<td>.06</td>
</tr>
<tr>
<td>Fonagy</td>
<td>−.23</td>
<td>.09</td>
<td>−.40</td>
<td>−.07</td>
<td>−2.75</td>
<td>.01</td>
</tr>
<tr>
<td>Frey</td>
<td>.18</td>
<td>.07</td>
<td>.05</td>
<td>.31</td>
<td>2.72</td>
<td>.01</td>
</tr>
<tr>
<td>Karna</td>
<td>.15</td>
<td>.02</td>
<td>.10</td>
<td>.19</td>
<td>6.62</td>
<td>.01</td>
</tr>
<tr>
<td>McLaughlin</td>
<td>−.17</td>
<td>.32</td>
<td>−.80</td>
<td>.45</td>
<td>−.05</td>
<td>.59</td>
</tr>
<tr>
<td>Schumacher</td>
<td>.29</td>
<td>.07</td>
<td>.15</td>
<td>.43</td>
<td>4.13</td>
<td>.01</td>
</tr>
<tr>
<td>Stevens “A”</td>
<td>−.05</td>
<td>.12</td>
<td>−.28</td>
<td>.18</td>
<td>−.40</td>
<td>.69</td>
</tr>
<tr>
<td>Stevens “B”</td>
<td>.14</td>
<td>.10</td>
<td>−.06</td>
<td>.33</td>
<td>1.35</td>
<td>.18</td>
</tr>
<tr>
<td>Overall</td>
<td>.05</td>
<td>.06</td>
<td>−.07</td>
<td>.17</td>
<td>.74</td>
<td>.46</td>
</tr>
</tbody>
</table>
measured the bystander intervention construct ($Q = 39.81$, $df = 11$, $p < .001$; $I^2 = 72.36$). However, because of the relatively small number of studies that included an empathy measure, we chose not to conduct moderator analyses to protect against findings of chance.

The results of these analyses were decomposed into substantive and methodological characteristics. Samples that consisted of high school students only generated a significantly greater treatment effect ($ES = 0.43$, CI = .33 to .52) compared to samples of primary schools only ($ES = 0.14$, CI = .11 to .18). We further assessed sample differences by conducting an analysis to evaluate location differences. The results revealed that programs located in the United States (US) did not differ significantly from those located in Europe (EU; US $ES = 0.26$, EU $ES = 0.13$, $p = .17$).

Another substantive moderator, treatment length, failed to produce significantly greater treatment effects ($1–2$ months $ES = 0.31$, $6–12$ months $ES = 0.16$; $p = .09$). Similarly, treatments that included a parental component (e.g., parent guides, parent training sessions) failed to influence the treatment significantly compared to programs without the component (parent included $ES = 0.19$, parent excluded $ES = 0.20$, $p = .92$).

We also grouped the studies by who facilitated the treatment programs. The teach-

<p>| Table 4 |</p>
<table>
<thead>
<tr>
<th>Moderator &amp; Meta-regression Analysis of Bystander Intervention Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moderator</td>
</tr>
<tr>
<td>Population</td>
</tr>
<tr>
<td>3rd–8th Grade</td>
</tr>
<tr>
<td>9th–12th Grade</td>
</tr>
<tr>
<td>Location</td>
</tr>
<tr>
<td>United States</td>
</tr>
<tr>
<td>Europe</td>
</tr>
<tr>
<td>Treatment length</td>
</tr>
<tr>
<td>1–2 months</td>
</tr>
<tr>
<td>6–12 months</td>
</tr>
<tr>
<td>Parent Component</td>
</tr>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Facilitator$^a$</td>
</tr>
<tr>
<td>Teacher</td>
</tr>
<tr>
<td>Other</td>
</tr>
<tr>
<td>Assignment</td>
</tr>
<tr>
<td>Non-Random</td>
</tr>
<tr>
<td>Random</td>
</tr>
<tr>
<td>Publication type</td>
</tr>
<tr>
<td>Peer-review</td>
</tr>
<tr>
<td>Non-peer review</td>
</tr>
<tr>
<td>Meta-Regression</td>
</tr>
<tr>
<td>Percent Male</td>
</tr>
<tr>
<td>HS (1 = Yes)</td>
</tr>
</tbody>
</table>

* $p < .05$; $a$-Menesini et al. (2003) did not indicate the facilitator; HS = High School sample.
ers implemented a significant portion of the programs for 7 of the 12 programs. Four other programs were facilitated by the researcher, a counselor, or in one case (Evers et al., 2007), with computer software. One study failed to indicate who implemented the program (Menesini, Codecasa, Benelli, & Cowie, 2003). The results of this moderator analysis revealed significantly greater treatment effects for programs that implemented the program with facilitators other than the teacher (teacher ES = 0.15, other ES = 0.43, p < .01). However, serious caution should be given to this finding because two of the four programs that used researchers as facilitators had the smallest sample sizes, and therefore this could be a reflection of imprecision or biased effects because of small samples (Levine, Asada, & Carpenter, 2009).

Finally, we conducted moderator analyses with two methodological groupings. We first observed mean group differences between randomly assigned (RA) and nonrandomly assigned treatments (NRA) groups. The results indicated that there were no statistically significant differences (RA ES = 0.21, NRA ES = 0.17, p = .74). We also estimated group differences between peer-reviewed (PR) and nonpeer-reviewed studies (NPR). The results of this calculation revealed that nonpeer-reviewed studies did not produce a greater treatment effect (PR ES = 0.16, NPR ES = 0.32, p = .07).

In addition to the ANOVA-like modeling, we conducted a weighted regression analyses. This analysis allowed us to estimate the effects of several predictors simultaneously. We hypothesized that the percentage of males in the treatment interventions and programs conducted in a high school would be significantly related to the treatment effect (i.e., Hedges’s g). The results of this analysis again revealed that, after controlling for the percentage of males in the treatment group, high school samples produced a greater treatment effect compared to middle or elementary school interventions (B = 0.25, Z = 1.98, p < .05). These results bolstered the previous findings.

To ensure that the overall effect size was not upwardly biased by including different measurements, we conducted a sensitivity analysis removing studies that used a dichotomous outcome. The results of the analysis found that the overall weighted effect size decreased slightly (original ES = 0.20, modified ES = 0.18). The overall effect remained statistically and practically significant.

**Publication Bias**

We applied Duval and Tweedie’s (2000) trim and fill procedure to address publication bias. This procedure revealed that one negative result was missing from the bystander intervention outcomes. However, the imputed missing values would only slightly change the overall fixed effect size (ES = 0.20); more important, it remained practically and statistically significant (see Figure 1). We also used Rosenthal’s fail-safe N procedure; the results of this calculations indicated that 236 null studies would be required to result in a non-significant finding. Egger’s regression intercept coefficient calculation also produced non-significant results (β0 = 0.57, p = .26). Taken together, we concluded that the review’s results were not affected significantly by publication bias.

**Discussion**

The purpose of this quantitative synthesis was to examine the treatment effects of bullying prevention programs on bystander intervention behavior. Empathy for the victim was also synthesized as a secondary outcome, but was not of primary purpose for the current review. In total, we reviewed 11 studies (12 effect sizes) from the United States and Europe that included 12,874 children.

Using meta-analytic techniques, the results revealed that the intervention behavior of bystanders increased (i.e., bystanders indicated greater intervention behavior in bullying situations) compared to control groups (g = .20). The results of a secondary analysis revealed that intervention programs did not have a similar effect on empathy for the victim (g = .05), but this finding should be viewed as
inconclusive because of the small number of studies that reported this outcome and its secondary nature.

These overall results mirrored the findings of a previously conducted small synthesis. Merrell et al. (2008) synthesized bullying prevention programs to investigate the effects of the programs on bullying perpetration, but also included several secondary measures (e.g., bystander intervention behavior, empathy). The authors reported small but significant treatment effects for bystander intervention behavior \((k = 3, g = .17)\) and nonsignificant negative effects with regard to empathy \((k = 3, g = -.10)\). Taken together, these replications results provided evidence against mono-operation bias and thus greater validity (Shadish, Cook, & Campbell, 2002).

Moderator analyses also revealed several findings of interest. Results of both ANOVA-like and weighted-regression analyses revealed that the treatment effects were greater for high school only samples. This is somewhat surprising because some scholars have postulated that bullying prevention programs are more effective for middle school-aged children (Williford et al., 2011). These results may indicate that bystander intervention behavior is a developmental process and programs may not influence younger students as intended.

The purpose of meta-analytic research is to generalize findings across populations, treatments, outcomes, and designs (Matt & Cook, 2009). Although this synthesis aggregated a smaller number of studies, its findings rendered generalizability of bullying prevention program’s effects on the bystander intervention construct. A few factors bolster this belief.

The populations assessed varied across ages, locations, and treatments. The largest effects were found for high school only samples; however, an overall significant treatment effect was also found for the total sample. With regard to location, no significant differences were found between U.S. and European samples. In addition, the studies employed a wide variety of treatment programs that proved efficacious. Moderator analyses also suggested that study design, publication type, and parental components produced similar
findings across studies. These results constituted a test of effects holding “across presumed irrelevancies” (Shadish et al., 2002, p. 455) and increases the findings’ external validity. Further investigation and evaluation is certainly required, but these results should cautiously suggest program generalizability.

**Limitations**

Several limitations should be noted. First, this meta-analysis included only 11 studies and 12 effect sizes. Although we took precautions to ensure unbiased effect sizes and findings, a great deal of caution should be used when interpreting the findings. The findings from a small collection of studies, no matter the statistical technique or number of students surveyed, should not enact immediate policy and practical changes. This becomes especially clear when one considers the effect of studies across time. Recent research in the field of meta-analysis publication bias has indicated that as programs increase in size and fidelity, effect sizes tend to decrease (Trikalinos & Ioannidis, 2005). Therefore, we plan to update these results periodically to observe the effect of time.

Second, because of the nature of meta-analyses, causal inferences should be stated cautiously. Quantitative meta-analysis, although statistically sophisticated and important, remains essentially an observational study (Lipsey & Wilson, 2001; Cooper & Hedges, 2009). On the other hand, the sample of studies we synthesized contained only those that used a treatment and control group, and these research designs constitute the most efficient measure of treatment effect (Shadish et al., 2002). Therefore, the synthesis of these results should partially reflect the nature of the primary studies.

Third, because of the low number of studies per factor, the ANOVA-like moderator analyses generated relatively low statistical power and the results should be interpreted with caution (Hedges & Pigott, 2004). Moreover, the low number of studies per group could easily capitalize on chance, or the groupings may reflect some other unforeseen variable. This becomes especially clear with regard to the studies grouped by treatment length. As alluded to previously, a majority of the studies grouped in the 1–2 month category were the smallest of studies. Smaller studies tend to produce larger and more unstable effect sizes (Levine et al., 2009), and therefore the grouping could reflect this phenomenon.

Fourth, although we made efforts to collect all primary studies that focused on bystander intervention behavior, it is quite possible that studies failed to be included. New material published postsearch, misspecified search terms, or simple human error could all cause inadvertent omission of extant literature. As such, we must temper our inferences with regard to the extrapolation of this information.

**Implications for Future Policy and Practice**

This meta-analysis should help cautiously to shift the emphasis of policy and practice. The results of this meta-analysis revealed two implications for policy. First, state and national bullying legislation should implement and evaluate programs that address bullying behaviors as a group process. Prevention frameworks and programs that attempt to abate bullying within schools are increasingly emphasizing changes in school climate that desist reinforcing bystander behavior or bullying perpetration (Cohen, 2006). The results of this study support these efforts to raise awareness about the participant roles, to encourage active and prosocial behavior, and to provide opportunities to role-play and practice bystander intervention in vivo.

Second, the results of this meta-analysis revealed that bullying prevention programs might be effective at encouraging prosocial bystander intervention when the framework, program, and/or curriculum explicitly target bystander attitudes and behaviors. It is simply not sufficient to only define prosocial bystander behaviors, such as walk away, get help, or stand up to those engaged in bullying. Policy must encourage the adoption of programs and interventions that shift attitudes
supportive of intervention (willingness to intervene) and behaviors through a consistent message about intervention and ample support from adults and administrators.

Future Research

This research suggests important future projects. First, primary research should focus on designing programs, implementing change, and measuring the bystander construct. As mentioned previously, researchers should focus on changing the behavior of the bystander. Second, further work is required to evaluate the effects of bystander behavior on bullying. This meta-analysis merely demonstrated that explicitly stated bystander programs have the ability to increase bystander intervention behavior. However, future research must continue to assess how bystanders implement these processes and the direct effects on active bullying. Third, the results from the empathy review revealed inconclusive findings. Future research is required to elucidate the effects of bystander behavior on bullying.

This meta-analysis merely demonstrated that explicitly stated bystander programs have the ability to increase bystander intervention behavior. However, future research must continue to assess how bystanders implement these processes and the direct effects on active bullying. Third, the results from the empathy review revealed inconclusive findings. Future research is required to elucidate the effects of bystander behavior on bullying.

Footnote

*Article used in meta-analysis.

References


Rigby, K., & Slee, P. T. (1993). Dimensions of interpersonal relation among Australian children and implica-


Date Received: February 1, 2011

Date Accepted: September 7, 2011

Action Editor: Joseph Betts

64
Joshua R. Polanin is a research methodology doctoral student at Loyola University Chicago. His interests include methodological improvements to meta-analysis and hierarchical linear modeling. He currently serves as the methodologist for a large, school-based bullying prevention program and as the managing editor of the Campbell Collaboration’s Methods Group.

Dorothy L. Espelage is a professor in the Department of Educational Psychology at the University of Illinois, Urbana-Champaign. Her research programs include investigations of bullying, sexual harassment, and dating violence among adolescents for almost two decades. She is engaged in a large randomized clinical trial of a school-based bullying prevention program.

Therese D. Pigott is a professor at the School of Education, Loyola University Chicago. Her research interests are statistical methods for meta-analysis and methods for handling missing data in statistical analysis. She is a co-chair and editor of the Methods Group of the Campbell Collaboration, an international organization that produces systematic reviews of social interventions.