Declines in Efficacy of Anti-Bullying Programs Among Older Adolescents:

Theoretical Considerations and a Three-Level Meta-Analysis

David Scott Yeager

Carlton J. Fong

Hae Yeon Lee

University of Texas at Austin

Dorothy Espelage

University of Illinois at Urbana-Champaign

Author Note

For their help conducting additional analyses with raw data to contribute to the present meta-analysis, the authors would like to thank Alexander Vazsonyi, Arthur Horne, Pamela Orpina, Caroline Hunt, Christopher Cheng, Dennis Wong, Tanya Beran, Maila Koivisto, Donna Cross, Scott Menard, Karin Frey, Antti Karna, Anne Williford, Aaron Boulton, Mario Gollwitzer, and Sally Kuykendall. We also thank the following people for responding to our requests for data, even though they were unable to provide it: Katherine Raczynski, Chris Bell, Carol Metzler, Ilse De Bourdeaudhuji, Susan Limber, Anna Baldry, David Farrington, Allison Fahsl, Robin Heydenberk, Leila Rahey, Wendy Hoglund, and Annis Fung. Last we thank Carol Dweck for her comments on the manuscript and Mike Cheung for his assistance with statistical models.

Please address correspondence to David Yeager at dyeager@utexas.edu.
Abstract

Highly visible tragedies thought to involve bullying in high schools have directly contributed to public support for state-mandated K-12 anti-bullying programming. But are existing programs actually effective for older adolescents? This paper first outlines theoretical considerations that lead to the prediction of a discontinuity in program efficacy among older adolescents. This paper then reports a novel meta-analysis of intervention programs that administered the same program to multiple age groups ($k = 17$, with 68 effect sizes). By conducting a hierarchical meta-analysis of the within-study moderation of efficacy by age, more precise estimates of age-related trends were possible. Results were consistent with theory in that while bullying appears to be effectively prevented in 7th grade and below, in 8th grade there is a sharp drop to an average of zero. There was a seeming reversal in efficacy through the high school years, such that programs, if anything, cause harm. This finding contradicts past meta-analyses that used between-study tests of moderation. Implications of these findings for developmental theory and for the interface between program evaluation research and social policy are discussed. The findings also represent a cautionary tale for the consumption of meta-analyses.
Declines in Efficacy of Anti-Bullying Programs Among Older Adolescents:

Theoretical Considerations and a Three-Level Meta-Analysis

Bullying is an aggressive act in which one or more individuals with relatively higher social power systematically and intentionally cause harm to an individual with relatively lower-power (Olweus, 1993). By now, the data are quite clear that victims of bullying suffer in terms of their social, emotional and academic development (Cook, Williams, Guerra, Kim, & Sadek, 2010; Reijntjes, Kamphuis, Prinzie, & Telch, 2010; Reijntjes et al., 2011; Ttofi, Farrington, Losèl, & Loeber, 2011) and even physical health (Copeland et al., 2014). These trends alone might have been cause for public action against bullying. But in the past decade and a half, there have also been a number of high-profile shootings and suicides carried out by older adolescents in high schools. Popular media interpretations have emphasized the role of bullying in these events (Grossman, April, 2, 2009; Gibbs, April, 19, 2010). It is unclear to what extent bullying actually contributes to such rare, extreme tragedies, but it is clear that such events have galvanized public support for laws requiring school-wide K-12 anti-bullying programs (Bierman, May, 4, 2010; Bullying Prevention Plan Act, Mass. Stat. c71, §§ 370, 2010). However, do existing programs work for high school students? If not, then such state laws, without more specific guidance, could lead to wasted resources or even increased bullying—at least until improved interventions were developed.

Recent meta-analyses of past anti-bullying interventions have suggested that, although there is some notable variability across nations and types of programs, “on average, anti-bullying campaigns have had some modest success” (Smith, 2011, p. 419; for null effects, see Merrell, Gueldner, Ross, & Isava, 2008; for modest positive effects, see Smith, Schneider, Smith, & Ananiadou, 2004; Ttofi & Farrington, 2011). Because past meta-analyses aggregated many
studies conducted with many different grade levels, it has further been possible to conduct “meta-regression” analyses to test for moderation by age. In one of the most recent and comprehensive meta-analyses, a meta-regression produced a significant positive effect of grade level (Ttofi & Farrington, 2011), leading the authors to conclude that “programs should be targeted on children aged 11 years or older rather than on younger children” (p. 46). Thus, based on the published record, policymakers may have been justified in requiring older adolescents to receive anti-bullying programs.

However, these past meta-analyses may have suffered from a methodological limitation. They involved between-study comparisons of age-related effectiveness, rather than aggregating within-study tests of moderation (see Smith, Salmivalli, & Cowie, 2012; Ferguson, San Miguel, Kilburn, & Sanchez, 2007). The potential for between-study tests of moderation to mislead meta-analysts has been well-documented by statistical experts (Cooper, 2009; Cooper & Patall, 2009; also see Olkin & Sampson, 1998; Lambert, Sutton, Abrams, & Jones, 2002). Within-study analyses allow for a comparison of the same content delivered at different ages, reducing confounds and increasing precision of age-related estimates. Between-study effects however average the ages of all students in a study. For example, in a between-study moderation analysis, an intervention administered to students in 3rd grade and 10th grade would be treated as an effect size for 6th grade students, even if no such students were included in the study (see, e.g., Figure 1 and row 12 of Table 2). This can mask real age-related trends. Figure 1 shows a simplified comparison to illustrate this. It shows how even though all studies may show negative slopes across grade levels, in the between-study analysis a single study with a lower average age and lower average effect size could produce a misleading, positive age-related slope.

In parallel, theory and data from past research in developmental psychology supports the
prediction that anti-bullying programs might be less effective in older age groups compared to younger children. There are developmental changes in the form of bullying (United States Department of Justice, 2009; 2011), the characteristics of those who bully (Cook et al., 2010), and the underlying psychological causes of bullying (Cohen & Prinstein, 2006; Cook et al., 2010; Faris & Felmlee, 2011; Guerra, Williams, & Sadek, 2011). Furthermore, there may be a domain-general developmental decline in standard techniques for behavior change, perhaps due to increases in reactance against controlling adults among older adolescents (Brehm, 1966; Erikson, 1968; Hasebe, Nucci & Nucci, 2004; Nucci, 1996; also see Larson, Wilson, Brown, Furstenberg, & Verma, 2002; Lapsley & Yeager, 2012). As we will show, much available evidence in developmental psychology would predict the exact opposite of the age trend found in Ttofi and Farrington’s (2011) meta-analysis.

The present paper reviews the potential for developmental differences in the efficacy of anti-bullying programs. Part 1 reviews theoretical expectations that would lead to the prediction of declines in program effect sizes across grade levels. Next, Part 2 reports a novel meta-analysis that assesses whether the findings from past intervention studies—when analyzed using within-study moderation tests—show patterns across grade levels that conform to theoretical expectations. Finally, the paper discusses implications for future research.

PART 1: THEORETICAL EXPECTATIONS

Overview

What developmental differences could undermine the efficacy of school-based programs? There are at least three considerations. These could be relevant to any risk behavior, not only anti-bullying programs. We review them in the abstract before considering them in the context of bullying. See Table 1.
First, the base rate of the \textit{form} of the problematic behavior could change with development, creating a moving target for intervention. Take the example of attempting to prevent use of illicit substances in adolescence. Imagine that use of inhalants peaks in 7\textsuperscript{th} grade, while use of cocaine peaks in 11\textsuperscript{th} or 12\textsuperscript{th} grade. In that case, educational interventions about the health risks of inhalants—even if they were persuasive—might be less likely to reduce overall prevalence of substance abuse if delivered in high school as compared to middle school.

Second, the \textit{underlying causes} of a problematic behavior could change. Take the example of childhood obesity. Among young children, the primary channel factor leading to junk food consumption might be what the parents put in the pantry. Yet among older adolescents in high school, who have more pocket change and freedom, then the eating habits of peers who accompany them to corner stores may be more influential. Hence, in this hypothetical example, parent-education interventions may be more effective at younger versus older ages, while peer or social network interventions may be more effective for older adolescents.

Third, \textit{domain-general behavior-change techniques} might be differentially effective across development. Techniques such as explicit classroom instruction in interpersonal skills or whole-school assemblies have been used to address a number of social problems, including substance abuse, childhood obesity, and of course bullying. Yet features of adolescence—such as a growing concern for autonomy or the increasing influence of the peer group—may undermine such strategies to the extent that they rely on adult authority. While not unique to bullying programs, differential effectiveness of these domain-general techniques may nevertheless be one reason to expect moderation by age in bullying program effect sizes.

Are these three general considerations a cause for concern regarding bullying intervention? In what follows, we present evidence from past research suggesting that they are
Declines in Efficacy of Anti-Bullying Programs

(also see Table 1). We note that while our organization of these ideas is new, many of the arguments have appeared in some form in past writings by various experts (e.g., Björkqvist, Lagerspetz, & Kaukiainen, 1992; Ellis et al., 2012; Hawley et al., 2010; Heilbron & Prinstein, 2008; Volk, Camilleri, Dane, & Marini, 2012; Sutton, Smith, & Swettenham, 1999).

1. Changes in the Form of Bullying.

First, it is possible that high school anti-bullying interventions often seek to change behaviors that are less prevalent in that age range, which could attenuate treatment effect sizes in these older age groups. A number of past scholars have stated or demonstrated that, with development, direct, observable aggression such as hitting, slapping, pushing, or calling someone names declines as a proportion of bullying, while indirect, less observable aggression such as social exclusion or starting rumors increases as a proportion of bullying (e.g., Björkqvist et al., 1992; Coie & Dodge, 1998; Dodge, Coie, & Lynam, 2006; Trembley, 2010; also see qualitative research by Guerra et al., 2012). Illustrating this possible trend, one participant in pilot research conducted for Yeager, Trzesniewski, and Dweck’s (2013) study said:

“Our high school isn't like any school in the movies. Jocks don't throw freshmen into the trashcan, or dunk nerds’ heads into the toilet. Bullies aren't people who punish physically, but are mostly just people who ignore and exclude others.”

Guerra et al. (2012) conducted a more systematic qualitative inquiry by using focus groups in elementary school and high school. For younger children, bullying was seen by children as when a person is “wrecking your stuff” or “kicking” you, while older adolescents “focused more on getting and keeping power and not letting ‘perfect’ kids get too full of themselves” (p. 305).
These trends also appeared quantitatively in findings from Volk, Craig, Boyce, and King (2006), who analyzed large-scale representative sample data from Canada (the 1998 Health Behavior in School-Aged Children survey). They showed that direct, physical assault forms of aggression declined dramatically from age 13 to 16.

Building on this, we used a large representative sample to examine more directly the question of whether direct forms of bullying (insulting or hitting) decline as a proportion of bullying while indirect forms (exclusion or rumors) increase. We analyzed the United States federal government’s School Crime Supplement of the National Crime and Victimization Survey, which uses probability sampling methods and involves many thousands of students (United States Department of Justice, 2009; 2011). We aggregated data from the two most recent surveys, conducted in 2009 and 2011, to ensure robustness. For greater detail about the sample, survey questions, and analyses, see Appendix 1.

Looking at the sub-set of students who reported being the victim of bullying, Figure 2 shows that reports of being the victim of direct, observable aggression declined from roughly 80% of bullying in 6th grade to just over 50% by 12th grade (almost perfectly replicating analyses of a different dataset by Volk et al., 2006). In a logistic regression predicting direct bullying with age, this was a significant linear decrease, odds ratio = .78, Z = 10.12, p < .001.

Meanwhile, extending past research, reports of being the victim of indirect, less observable aggression increased from just over 60% of bullying in 6th grade to roughly 75% of bullying by 12th grade, a significant linear increase, odds ratio = 1.10, Z = 4.23, p < .001. These two developmental trends were significantly different from each other in a multivariate regression, Wald test of equality of coefficients, $F(1, 2824) = 90.63, p < .001$. Thus, in these analyses,
exclusion/rumors catches up to and then overtakes hitting/insults as the leading type of bullying in high school (Figure 2).

Furthermore, a large amount of research has documented that bullying and victimization related to sexual relationships or sexual orientation increases with development, spiking in 9th grade (e.g. Pepler et al., 2006). As adolescents age and complete puberty, they are increasingly likely to engage in aggression against individuals they deem competitors for sexual partners (Vaillancourt, Miller, & Sharma, 2010), or against individuals who do not conform to traditional gender roles (Hong, Espelage, & Kral, 2011).

Altogether, one might expect weaker effect sizes with age if an intervention focused on direct, observable forms of aggression and neglected either the more indirect, unobservable forms or sexuality-related victimization. Developmental trends in effect sizes might also appear if an intervention focused on the latter two forms of victimization but did so unsuccessfully.

2. Changes in the Underlying Causes of Bullying

Next, the causes of bullying could change across development. That is, the strength of a relation between a causal antecedent and bullying could become weaker or stronger as children become adolescents. Types of causes that could change are social competence, motives, and settings, and we emphasize the first two here (also see Table 1).

**Social competence.** Is bullying caused by a deficit of social competence or by a misguided surfeit of social intelligence? A number of scholars have questioned the accuracy of the popular stereotype of the “bully”—that is, a low-intelligence, socially-incompetent physically-aggressive person (Sutton et al., 1999). Instead, emphasis has been placed on the view that bullying is perpetrated by individuals who are effective at manipulating the feelings of others in order to attain social status—something that requires great social savvy to carry out
without getting caught (Garandeau & Cillessen, 2006; Sutton et al., 1999). Which view is more correct?

It is possible that both stereotypes have some truth to them, but to a greater or lesser extent at different ages.¹ Interestingly, prominent theories (Hawley et al., 2011; Sutton et al., 1999) have not sufficiently discussed this possibility, although data are accumulating in support of it. A large amount of research on young children has found a correlation between social and emotional deficiencies and a child’s tendency to bully others (Cook et al., 2010; Copeland, Wolke, Angold, & Costello, 2013; Espelage, Bosworth, & Simon, 2001). Meta-analyses of many past studies demonstrated that, among younger children in elementary school and sometimes earlier in middle school, bullying is predicted by lower social influence, poorer social perspective taking and social problem solving, poorer impulse control, low academic achievement, a poor home environment and other deficits (Cook et al., 2010; also see Caravita et al., 2009; Espelage, Low, Rao, Hong, & Little, 2014). And so designing interventions that seek to remediate deficits in these areas might be expected to benefit children’s ability to control or change their bullying behavior. And indeed, it is common for interventions remediating social skills to have positive effects on both aggressive behavior and adjustment more generally among younger age groups (Brown, Low, Smith, & Haggerty, 2011; Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011).

These same correlations, however, become markedly different at later ages. A meta-analytic synthesis by Cook et al. (2010) found that age moderated the relation between bullying

¹ Note that although here we focus on average trends, it is of course the case that at all age groups there will be individual students who do not fit the average description—i.e. socially-savvy young children who bully and physically bullying older adolescents (for the former, see, e.g., Farmer et al., 2010). Person-centered analyses can be helpful in identifying relatively-uncommon sub-groups across development (Hawley et al., 2010).
others and adjustment. Among children (age 3-11 years) popularity was a strong and significant *negative* predictor of bullying, but among older adolescents (age 12-18 years), including high school students, popularity was a significantly weaker, and non-significant, predictor of bullying others (Cook et al., 2010). Data from individual studies show that the negative relation between popularity and bullying reverses sign and becomes gradually more strongly positive (see a narrative review by Heilbron & Prinstein, 2008). Rose et al. (2004) found that popularity negatively predicted indirect, social bullying in 3rd grade (at $r = -.15$), but positively later: in 7th grade $r = .27$, in 9th grade $r = .39$ (also see Cillessen & Mayeaux, 2004).

Some compelling data show that the moderate-to-high popular students may be driving these effects. Faris and Felmlee (2011; 2012), conducted social network analyses to discern levels of popularity (specifically, “network centrality”). The authors found that in high school the students who most frequently bullied others were not socially maladjusted or unpopular but were average or above average in popularity. Much of bullying occurred between the 75th and 93rd percentile in terms of popularity (Faris & Felmlee, 2012), with very little occurring at the very low and very high ends of the spectrum. That is, it was not a small sub-set of students with low empathy, poor perspective-taking or few friends who bullied others. Instead, bullying was carried out by a large set of mostly average teens who had reasonable numbers of friends and some level of popularity.

Finally, among younger children, externalizing behaviors were more likely to be co-morbid with bullying than among older adolescents (Cook et al., 2010). Said another way, among younger children, bullying was likely to be a correlate of a host of self-regulatory and social difficulties, while in later adolescence this was less true. For instance, Caravita et al. (2009) showed that among older adolescents *higher* levels of “cognitive empathy” predicted
more bullying, but not among younger children. This again suggests that perhaps later in development bullying others is less of an indicator of deficits in social skills, and more of a signal of the presence of strong interpersonal skills that are applied to harm others.

What accounts for this developmental shift in the correlations of social competence with bullying? In the prior section, we reviewed evidence that the form of bullying changed with development—moving from more overt forms such as hitting / insulting to more covert forms such as exclusion or spreading rumors, or to bullying involving sexual relationships. And these differential behaviors require different sets of skills to carry out successfully—the former relying more on physical strength, and the latter two on sufficient savvy to maximize the benefits for one’s popularity of harming others, while minimizing the potential for retribution from peers, adults, or potential sexual partners. Indeed, Andreau (2006) documented that overt aggression was predicted by a lack of strong social skills, while more covert forms of aggression were predicted by the presence of strong social skills (the latter was also shown by Kaukiainen et al., 1999). In general, deploying social, indirect bullying strategically requires advanced social skills and self-regulation. These are typically not fully developed in early childhood or early adolescence, but become more mature later in adolescence, perhaps leading to developmental differences in the correlates of social skills and of the efficacy of skill-based anti-bullying interventions.

All of this has at least two implications for intervention and policy. First, it should be apparent that attempting to remediate social skills should have positive effects for younger children and, to the extent that those skills facilitate more effective bullying, null or potentially even iatrogenic effects for older adolescents may be expected (Sutton et al., 1999). Second, there is cause for concern that anti-bullying programs in high school may be applying the wrong
stereotype of a “bully.” Hypothetically, it is possible that programs ask students to imagine the prototypical “victim” and “bully” might mislead school staff and even peer bystanders attempting to reduce bullying (Sutton et al., 1999). They may be looking for a low-intelligence, socially unskilled “bully” and may as a result believe that the aggression carried out by popular students does not count in their minds (Sutton et al., 1999).

Last, it is important to emphasize that it is possible for interventions to remediate factors relevant to social competence in early childhood and then reduce aggression later in development. For instance, interventions sometimes involve changing attributions about the ambiguous intent of others who might be harming a person—that is, the hostile attributional bias (Dodge, 2006; Hudley & Graham, 1993)—which is a key predictor of poor social competence. Dodge, Godwin, and The Conduct Problems Prevention Research Group (2013) tested this as a part of a comprehensive intervention among a group of students that was pre-selected to be likely to demonstrate anti-social behavior including physical aggression. These authors found that addressing social-cognitive deficits in kindergarten led to reductions in anti-social behavior until at least 9th grade. Thus, among students that comprise the group of students who were likely to bully physically in high school, social skills may have been a cause of physical bullying. Yet as noted physical bullying comprises a relatively smaller proportion of bullying later in high school as compared to earlier in childhood, and so this type of approach may be less effective at reducing overall prevalence of adolescent bullying and victimization.

**Motives.** Even if high school anti-bullying interventions target the age-appropriate type of bullying, they may have no effect if they do not successfully address the underlying motives for bullying. And motives for bullying may vary across age groups.
First, as adolescents transition to high school, peer groups become highly unstable (Cairns & Cairns, 1994). Illustrating this, in one study that surveyed 9th graders about their friendships every month for five months, roughly 50% of teenagers’ friends were different from one month to the next (Chan & Poulin, 2007). Such findings signify the social reality that, as teens learn about themselves, their identities, and their peers, they strike up new friendships and discard old ones. This can have many implications for teens’ behavior. In part this period of transition can signal a new opportunity to improve one’s standing and climb to higher levels of peer regard. During this time of change, teens may seek to keep lower-status peers from joining their group, or they may intentionally start rumors about others in order to neutralize peer threats.

Both qualitative and survey research illustrate that school transitions can be a “land grab” for status and peer influence (Crosnoe, 2011; Guerra et al., 2012; see Pelligrini & Long, 2002). To the extent that teens may believe that where the “dust settles” can have long-term implications for their social happiness, teens may be willing to engage even in extreme forms of social aggression to preserve or increase status.

A number of scholars have advanced this account (Cillessen & Mayeux, 2004; Ellis et al., 2012; Hawley et al., 2010; Heilbron & Prinstein, 2008; Volk et al., 2012; Sutton et al., 1999). One experiment directly investigated causality (Cohen & Prinstein, 2006). Cohen and Prinstein (2006) selected high school students with middle levels of popularity—not too low and not too high—on the basis of sociometric nominations. Next, students entered a chatroom that they believed was populated with actual peers from their class in school (it was in fact controlled by the computer). These “classmates” were characterized as either high status or low status (randomly assigned). After some initial activities, the “classmates” then voted to exclude another peer from the group, and the participants had to choose whether to cast the final,
deciding vote in favor of excluding the peer or not. Cohen and Prinstein (2006) found that these mid-level popularity teens were over thirty percentage points more likely to vote in favor of excluding the “peer” when they believed the high status classmates would see their vote, as compared to when they believed the classmates seeing their vote were low status. This finding provides causal evidence for the notion that a desire to improve social status leads to forms of social aggression (such as exclusion) that Figure 2 shows are most common in high school.

This has implications for developmental accounts of the causes of bullying during the transition to high school. It suggests that insecurity about social status, perhaps stemming from the ambiguity created by a difficult personal or school transition, can lead teens to join groups that engage in extreme behavior to protect their social standing (see Aikins & Litwack, 2011; Caravita & Cillessen, 2012; Cillessen & Mayeux, 2004; Faris & Felmlee, 2011; Ryan & Shim, 2008). This is a very different proposed cause of bullying than is commonly documented about younger children, which involves disliked and socially marginalized students with poor self-regulatory or social skills (Sutton et al., 1999).

This analysis also suggests that individual differences in social cognitions related to the drive for status—which can motivate a person’s need to demonstrate that status—may be increasingly important as causes of bullying during the transition to high school. That is, a potential threat to social status is not, of itself, a cause of bullying. Instead, that threat must be interpreted by the student in a way that conflicts with their goals (see, e.g. Dodge et al, 2006; Olson & Dweck, 2008). If one’s goal is to demonstrate high social status, then having a low-status peer joining one’s group could be interpreted as a threat. In this instance, social aggression may be interpreted as an effective method for meeting one’s social goal (Aikins & Litwack, 2011; Caravita & Cillessen, 2011; Cillessen & Mayeux, 2004; Ryan & Shim, 2008;
Declines in Efficacy of Anti-Bullying Programs

Volk et al., 2012).

Indeed, a growing amount of research is showing that individual differences in social demonstration goals (Rodkin, Ryan, Jamison, & Wilson, 2013; Rudolph, Abaied, Flynn, Sugimura, & Agoston, 2011; also see Erdley, Cain, Loomis, Dumas-Hines, & Dweck, 1997) can be powerful predictors of an adolescent’s likelihood of engaging in peer aggression. Specifically, teens whose goal is to demonstrate high status to others—that is, those who agree that “It is important to me to have ‘cool’ friends” or “My goal is to show other kids how much everyone likes me”—are more likely to respond to peer conflicts that threaten their status by resorting to aggression (Rodkin et al., 2013; Rudolph et al., 2011; also see Yeager, Trzesniewski, Tirri, Nokelainen, & Dweck, 2011; Yeager, Trzesniewski, & Dweck, 2013).

We have been suggesting that during the transition to high school, social goals may become more important, while simultaneously the social environment—with its constantly-changing friendship groups (Cairns & Cairns, 1994; Chan & Poulin, 2007)—seems to involve more opportunities for threats to those goals. Due to both influences, bullying may increasingly be shaped by social cognitions relative to status and popularity. To date, social-cognitive interventions to reduce aggression have not been updated to address social demonstration goals, instead focusing on important factors such as the hostile attribution bias, which have demonstrated efficacy for younger children but less efficacy for older children (Metropolitan Area Child Study Group, 2002). Indeed, Garandeau, Lee, and Salmivalli (2014) found that an intervention with strong overall efficacy had no effect on the most popular children, suggesting that, at least in the generally highly-successful KiVa program, the desire to demonstrate and attain social status may not have been fully addressed.
Causes related to sex, sexuality, and race/ethnicity. The possibilities noted above are not the only developments that could account for changing efficacy of anti-bullying interventions. There are also changes in sexual maturity that could alter both same-sex and cross-sex relationships. Indeed, qualitative work highlights the importance of competition over romantic partners as one cause of girls’ bullying in high school. For instance, girls who make themselves too alluring can be targeted by other girls for intense bullying (see Guerra et al., 2012; also see Vaillancourt, Miller, & Sharma, 2010). Also, bullying and victimization increases during later adolescence among gender nonconforming youth (Hong et al., 2011; Horn, Kosciw, & Russell, 2009) and lesbian, gay, or bisexual (LGB) youth (Espelage, Aragon, & Birkett, 2008; Robinson & Espelage, 2012). No evaluated anti-bullying programs address victimization based on sexual orientation and/or gender identity in great depth (Birkett, Espelage, & Koenig, 2009).

Potential declines in efficacy could also be related to race, ethnicity and culture. Although bullying others does not seem to vary significantly by race or ethnicity among adolescents (Hepburn, Azrael, Molnar, & Miller, 2012), ethnic minority and recent immigrant students report higher levels of being victimized by others (Peguero, 2012). One explanation for this may come from research by Killen and others (for a review, see Killen, Mulvey, & Hitti, 2012; also see Levy & Killen, 2008), which has found that children’s reasoning about exclusion based on race continues to develop into adolescence. Specifically, developmental intergroup theories note that majority group adolescents may use exclusion of out-group members as a method to maintain their own group’s status. And recall that bullying later in adolescence in general is increasingly experienced in the form of exclusion and seems to be motivated by efforts to maintain or gain status. Race- or ethnicity-based exclusion may not be reduced if an anti-bullying program uses a race-neutral approach that does not address the underlying social-
cognitive causes of this exclusion (see Killen et al., 2012). To the extent that many programs used in the U.S. were imported from European countries that are relatively racially homogenous, or from suburban, predominately White areas of the U.S., then these programs may not have included material that sufficiently addresses the underlying causes of race-based exclusion.

3. Changes in Responsiveness to Domain-General Behavior-Change Strategies

Heckman and Kautz (in press) recently reviewed a wide variety of interventions designed to affect consequential youth outcomes, such as intelligence, personality, health, or crime. These authors uncovered a striking trend. Early childhood interventions routinely seemed to show persistent and widespread effects on multiple domains of development, but interventions delivered later, in adolescence, often had null effects or benefits that quickly diminished. The authors use these data to support the claim that by adolescence certain patterns of behavior are simply fixed, like plaster, and that interventions are best directed toward younger children. Without denying the importance of early intervention, another explanation is possible: Perhaps the optimal delivery mechanism for effective interventions changes in adolescence, making it difficult for studies using traditional delivery mechanisms to have lasting effects. If true, and applied to the present case of anti-bullying programs, then even interventions grounded in accurate theory about the causes of bullying could have no effect or a dampened effect if delivered in developmentally-inappropriate ways. We explore this possibility here (for an analogous review, see Lapsley & Yeager, 2012).

Why might older adolescents need to be intervened-upon differently? Decisions about friendships that are the root of bullying—decisions about who to hang out with and who to exclude, who to like and who to dislike—are important personal choices. And so one challenge in attempting to influence anything having to do with social relationships is that these attempts
can trigger teens’ drive to assert their autonomy and rebel against what they perceive to be adults’ attempts to control a personal domain (Erikson, 1968; Hasebe, Nucci & Nucci, 2004; Nucci, Killen, & Smetana, 1996; also see Larson et al., 2002). Teens commonly express reactance in response to adults’ attempts to influence their personal goals (Brehm, 1966; Erikson, 1968; also see Grandpre, Alvaro, Burgoon, Miller, & Hall 2003; Henriksen, Dauphinee, Wang, & Fortmann, 2006), rejecting adult’s suggestions—or even endorsing their opposite—to reassert their autonomy (Lapsley & Yeager, 2012). Indeed, developmental research finds that older adolescents (i.e., age 16), compared to younger children (i.e., age 8-10), increasingly invoke their right to make personal choices and not have them controlled by adults in school (Ruck, Abramovitch, & Keating, 1998).

Experimental work shows that even small changes in how a request is made can trigger reactance or, alternatively, allow a teen to feel autonomous while accepting the suggestion. Much of this has grown out of self-determination theory (see Vansteenkiste, Lens, & Deci, 2006). For instance, in some laboratory research, positive feedback about how a student performed (“You did well”) improved motivation to continue in a task, but using controlling language (“You did well, as you should”) undermined motivation (Deci, Koestner, & Ryan, 1999). A field experiment with high school students similarly suggested the importance of autonomy for adolescents. When a new activity was presented to teens in less controlling terms (e.g. “You might decide to learn more”) as opposed to in more controlling terms (“You should decide to learn more”), students processed the information more deeply and were more likely to apply it in their lives a few days later (Vansteenkiste, Simons, Lens, Sheldon, & Deci, 2004; also see Vansteenkiste et al., 2006; Reeve, Jang, Carrell, Jeon, & Barch, 2004; Ryan & Deci, 2000).
Although past experiments have not been conducted in bullying prevention, one prominent example illustrates the power of controlling language to undermine well-intentioned change efforts in public health, as well as the power for autonomy-supportive processes to promote behavior change. This example is the truth™ anti-smoking campaign. The truth™ campaign is thought to have produced dramatic reductions in teenage smoking rates by depicting non-smoking teens as rebellious and independent for standing up to “Big Tobacco” (e.g., with television commercials in which teens pile body bags outside a tobacco company building while yelling at executives through a megaphone). This method has proven much more effective than direct injunctions from adults to “just say no” to smoking—a strategy that, in some rigorous evaluations, was found to ironically promote smoking, presumably by triggering the reactance processes described above (Bauer, Johnson, Hopkins, & Brooks, 2000; Farrelly et al., 2002).

Some have argued that the key to the truth™ campaign’s success relative to the more-controlling “just say no” strategies was its portrayal of tobacco companies as the “authority” seeking to manipulate teens and of non-smoking teens as rebellious in standing up to them (Bauer et al., 2000; Evans, Price, Blahut, Hersey, Niederdeppe & Ray, 2004).

Building on these possibilities, some developmental experiments explicitly tested the special importance of autonomy in high school as compared to earlier in development. Grandpre et al. (2003) found that telling 10th graders not to smoke seemed to increase their willingness to try a cigarette, while making suggestions in an autonomy-supportive way (as in “you might try” not to smoke or “you could try” to say no) seemed to decrease intentions to smoke. Among younger teens in 4th or 7th grade, the controlling versus autonomy-supportive conditions did not differ, suggesting that concerns about autonomy are especially important later in adolescence and perhaps less important earlier in childhood.
Declines in Efficacy of Anti-Bullying Programs

These theories and findings raise the possibility that even if past anti-bullying interventions were built on appropriate theories of adolescent bullying they may have been less effective than desired due to use of controlling language (such as “don’t bully” or “be respectful”) or due to controlling messaging that forbids certain behaviors in a way that insufficiently emphasizes personal choice (see Figure 3). If this were true, then one might expect smaller effect sizes specifically in high school, the age when teens’ concerns about autonomy might be greatest (for discussions of “stealthy” intervention strategies that avoid controlling injunctions, see Yeager & Walton, 2011; also see Robinson, 2010). Supporting these possibilities, it is interesting that a correlational study by Roth et al., (2011) found that students who reported that their teachers were autonomy-supportive were also less likely to bully peers.

Summary

This review provided reasons to predict developmental differences in the effects of anti-bullying programs (see Table 1). Programs that seek to reduce direct physical or verbal bullying, remediate basic social skills such as empathy or perspective-taking, and use direct instruction or skill repetition, may well be effective for younger children in elementary school. As adolescents age, however, the forms and causes of bullying may shift, as may the efficacy of domain-general behavior-change strategies (Table 1). Adolescents transitioning to high school, for instance, are more likely to engage in covert, indirect aggression such as exclusion, or to bully in ways related to sexuality, such as targeting gender non-conforming or promiscuous peers. These adolescents also may do so in order to defend or attain social status or strengthen romantic relationships, and may sometimes employ sophisticated social skills in order to be effective at this. Further, they may resent the potentially insulting implication that they lack basic skills, and may show reactance against controlling commands from adults to change what seems like a personal
choice. As such, strategies that may show modest efficacy for young children could have null or even iatrogenic effects for older adolescents.

**A New Meta-Analysis is Needed**

Given the review above, it may be surprising that past meta-analyses have shown that anti-bullying programs are more effective for adolescents as compared to children (e.g., Ttofi & Farrington, 2011). And recently, Smith et al. (2012) criticized the between-study approach employed in that meta-analysis (Ttofi & Farrington, 2011). Smith et al. (2012) noted that when looking at prominent examples of interventions that delivered the same content to a wide age range, in every case younger children benefitted more than older adolescents (Kärnä et al. 2011a, b; 2012; Olweus & Limber, 2010; Smith, 2010). Ttofi and Farrington (2012), in a reply, located eight total studies in which within-study comparisons were available, and conceded that many showed the same trend of a declining effectiveness with age. However, these analyses by Ttofi and Farrington (2012) are limited in a number of important ways. First, they included a small group of studies. The authors did not search for additional examples or contact authors to ask for new calculations of raw data. Second, these analyses did not involve formal statistical tests of the within-study slopes. All told, Ttofi and Farrington (2012) concluded that their age-related data were inconclusive. Both sets of authors called for a new meta-analysis that aggregates within-study moderation by age and conducts appropriate statistical tests of this relation (Smith et al., 2012; Ttofi & Farrington, 2012).

**PART 2: A META-ANALYSIS OF WITHIN-STUDY AGE MODERATION**

Building on the theory outlined above, the second half of the present paper involves a novel meta-analysis of within-study moderation by age of anti-bullying program effectiveness. This involved calculating an effect size for each age group in each relevant study that appeared...
in past meta-analyses, in addition to inclusion of new studies that were published since the previous meta-analyses stopped collecting data. For statistical analyses, we build on recent advances in three-level meta-analysis (Cheung, in press; Konstantopoulos, 2011; Marsh et al., 2009; Van den Noortgate et al., 2012). Indeed, Cheung (in press) notes that traditional meta-analyses can be considered two-level models with participants at level 1 and the studies at level 2, while three-level meta-analyses add another level that allows the effect sizes to be correlated within a cluster. In the present case, level 2 involves the multiple effect sizes for the different age groups in each study, while the level 3 “clusters” correspond to each study. Analyses were conducted using the metaSEM package in R (Cheung, 2012), which generalizes the structural equation modeling framework to model the nested dependencies of effect sizes.

In doing our analyses, we initially tested for linear trends. However, the theoretical predictions outlined above focused mostly on a discontinuous “step” function—a switch from effectiveness in early childhood and adolescence and to ineffectiveness later in adolescence, near the transition to high school. We therefore tested for this discontinuity. Finally, we compared this within-study analysis to the more traditional between-study moderation test, in order to demonstrate the potentially misleading nature of past procedures.

Methods

Data

**Inclusion and exclusion criteria.** To be included in the present meta-analysis, a study had to meet several criteria. First, all studies included in the meta-analysis needed to have evaluated the effectiveness of an anti-bullying intervention with measured outcomes of bullying others. Second, because differential program effects across ages were of primary interest to this meta-analysis, a study had to involve data on multiple age groups to be included. Third, the study
had to provide the statistical information required to derive the effect size either from the report itself or from personal communication with the authors. Lastly, studies had to have appeared in electronic search databases or have been available to us by September 1, 2012.

Our study focused on the effects of anti-bullying interventions across different age groups, thus we did not include studies with multi-age aggregated groups because they did not provide the possibility of estimating age-specific effect sizes.

**Report identification and selection.** Multiple strategies were used to locate all relevant studies that met the inclusion criteria. A flowchart depicting these is presented in Figure 4. First, we reviewed previous meta-analyses and narrative reviews in order to re-analyze previous interventions included in older reviews (i.e., Brown, 2009; Farrington & Ttofi, 2011; Merrell et al., 2008; Polanin, Espelage, & Pigott, 2012; Smith et al., 2004; Vreeman & Carroll, 2007). In addition, we searched for any new meta-analyses before September 2012; however, we identified no relevant reviews. We extracted all the included studies of the previous meta-analyses, which, based on the title and abstract, seemed to meet our inclusion criteria. This yielded 89 studies.

In order to ensure that we obtained all studies conducted after the previous meta-analyses, we searched for documents catalogued after 2008 (after the most recent of the previous meta-analyses) and before September 2012. The databases searched were PsycINFO, ERIC (Educational Resources Information Clearinghouse), Proquest Dissertations and Theses, Google Scholar, Social Science Citation Index, EBSCO, ASSIA, PubMed, Sociological Abstracts, GALE, Academic Search Complete, MedLine, Campbell Collaboration, and Cochrane Collaboration. For each database, a series of search terms were employed to identify any evaluations of anti-bullying interventions, applying the appropriate truncation and Boolean techniques to achieve an inclusive yet focused search. The complete search strategy is provided
in Appendix 2. Our search strategy incorporated all search terms from all previous meta-analyses. This search yielded 14,889 potential studies. Then, a backward search was employed. Furthermore, we emailed experts in the field and asked them for relevant published studies. Upon evaluating titles and abstracts, we identified 84 more potential evaluations that required full-texts to further evaluate for inclusion.

Once this search strategy was employed, and all citations had been retrieved, full-texts for these studies were judged for relevance, resulting in three sets of studies. The first of which were excluded studies that did not meet our inclusion criteria upon full-text review ($k = 116$). The second set ($k = 5$) included studies that met all the inclusion criteria including effect size data by age or grade. The third set ($k = 50$) had studies that met all but one inclusion criteria, which was the effect size data by age or grade. The studies in this final set reported a multi-age sample but no effect size data separated by age.

Our next step was to contact all the authors of the identified studies that seemed to meet inclusion criteria but did not report the necessary effect size data. We requested that they calculate the necessary means, standard deviations and sample sizes so we could compute effect size data by age. If first authors did not respond, co-authors were contacted next in sequential order. Out of 50 studies, we received 18 responses from authors. Twelve of the responding authors provided the necessary data to calculate effect sizes per age group. The other 6 responses either could no longer access the data ($k = 5$) or provided another contact person that later did not respond ($k = 1$).

This left us with a final sample of 17 reports with the necessary statistical information for computing within-study age trends. See Figure 4 and Table 2.

**Measures**
Outcomes. The outcome was reductions in reports of bullying other students. Whenever a composite index was used by the primary study, we coded the results based on the composite score instead of the individual items. For example, a study may report physical bullying and verbal bullying scores as well as a total bullying score; the latter was used when possible. If a composite score was not used and multiple indices of bullying were reported, then when computing our effect sizes we used two-stage meta-analyses that first aggregated effects across all dependent variables to produce a single effect size estimate. This was done using standard formulas for averaging dependent effect sizes in the Comprehensive Meta-Analysis software (Version 2.2; Borenstein, Hedges, Higgins, & Rothstein, 2005; for more background, see Cheung & Chan, 2004; Marín-Martínez & Sánchez-Meca, 1999; Rosenthal & Rubin, 1986).

Age. Because studies often did not report exact ages of students, we used grade level as a proxy. Some countries (such as Finland) number their grade levels differently from the United States, and so grade levels for these countries were re-coded to correspond to the United States system.

Computing effect size estimates

We used the standardized mean difference to estimate the effect of the anti-bullying intervention. This calculation results in a measure of the difference between the two group means expressed in terms of their common standard deviation. When this information was not reported in a study, corresponding inference test statistics (e.g., \(t\)-statistic, \(F\)-statistic, \(p\)-values) were used to derive an effect size. Effect sizes reported as odds ratio were converted to Cohen’s \(d\) effect sizes using commonly-employed formulas (Chinn, 2000).

A weighting procedure was used to calculate average effect sizes (Cohen’s \(d\)) and standard errors across independent samples. Each effect size was first multiplied by the inverse
of its variance; then, the sum of these products was divided by the sum of their inverses. This procedure allows for more weight to samples of larger size, which is generally preferred (Hedges & Olkin, 1985). We used the Comprehensive Meta-Analysis software to calculate effect sizes (Borenstein et al., 2005). Table 2 reports all effect sizes and sparklines showing trends across grade levels.

Results

Do Anti-Bullying Programs Decline in Effectiveness with Age?

Preliminary analyses. In the full sample of 68 effect sizes, the simple correlation between age and effect size was $r = -0.22$, such that older participants showed weaker effects, as expected by theory. Next, we estimated the within-study correlation between age and effectiveness for each study, calculating an $r$ value for each. The average within-study correlation was $r = -0.62$, and the median was $r = -0.86$. Furthermore, 80% of studies showed negative correlations, and in a sign test this was significantly different from 50%, $p = .02$. This preliminary analysis is consistent with theoretical predictions of a negative relationship between age and anti-bullying program effectiveness. It is also informative because unlike standard meta-analytic techniques the sign test prevents a single study with a large sample size to exert an influence that misrepresents broader trends. However, this analysis does not take into account clustering of the data or potential non-linearity, nor does it involve a formal meta-analytic test of significance that weights for sample size, something we consider next.

Three-level meta-analysis: linear and non-linear effects. We conducted a three-level random-effects meta-analysis, with level 1 corresponding to participants (whose raw data are not available), level 2 corresponding to effect sizes at different age groups within studies, and level 3 corresponding to studies (Cheung, in press; 2012). Maximum likelihood estimation was used.
When treating age as a continuous linear variable, there was a marginally significant negative linear trend, \( b = -0.02, Z = -1.91, p = .056 \).

Raw effect sizes, unweighted for sample size, are plotted in the left panel of Figure 5. The left panel in Figure 5 suggests a drop in effectiveness toward the end of middle school, consistent with our theory-based predictions of a discontinuity in effect sizes. This suggests that a linear analysis might not be the appropriate comparison. Confirming this, in a three-level meta-analysis, the dichotomous comparison of effect sizes in grades 1-7 vs. grades 8-13 was significant, \( b = -0.11, Z = 1.99, p = .047 \). Follow-up three-level meta-analytic tests found that in grades 1-7 there was a significant average effect of anti-bullying interventions, \( d = .13, Z = 4.42, p < .001 \), while in grades 8+ there was no significant average effect, \( d = .02, Z = .48, p = .63 \).

Further inspecting the left panel in Figure 5, however, seems to reveal that effect sizes continue to decline after 8th grade. We therefore tested a step function by creating two predictors: a variable that increased linearly from 1st grade to 7th grade, and a second variable that increased linearly from 8th grade to 13th grade. When both were entered into the model, there was no significant trend from 1st to 7th grade, \( b = .005, Z = .33, p = .74 \), but there was a significant decline from 8th to 12th grade, \( b = -.06, Z = 2.29, p = .02 \). Using this model, values for each grade level were estimated; these are plotted in the right panel in Figure 5. Remarkably, these estimated values suggest that by roughly 10th grade anti-bullying efforts are in the direction

\[2\] We also examined whether the elementary school years (up to grade 5) differed from middle school years that showed treatment effects (grades 6 and 7). They did not.

\[3\] Because only one study (conducted in Europe) involved 13th grade, analyses were re-conducted dropping that age group, to ensure it was not driving effects. When this was done, the linear decline in effect sizes from 8th to 12th grade remained significant, \( b = -.05, Z = 1.97, p = .049 \).
of causing harm, on average. Although this iatrogenic effect for high schoolers was not significant on average, it was significant in some individual studies (e.g., Hanewinkel, 2004).  

**Comparison to Standard Between-Study Techniques**

Finally, we sought to replicate past studies’ between-study moderation analyses. This analysis averaged the grade levels for each study and the effect size for each study, to produce only one effect size estimate and one age group per study. Conducting this analysis is helpful for the present purposes in confirming that the differences between our results and those from past meta-analyses is due to our use of proper analytic techniques, and not due to the more limited sample of studies that met our inclusion criteria of having multiple effect sizes across ages.

Recall that many past studies found a non-significant age moderation, such that program efficacy was slightly but not significantly lower for older students (Ferguson et al., 2007; Fossum, Handegård, Martinussen, & Mørch, 2008; Merrell et al., 2008; Smith et al., 2004; Vreeman & Carroll, 2007; Wilson & Lipsey, 2007), while one study found a significant positive relation (Ttofi & Farrington, 2011). Using data from our set of studies, we replicated the former finding of no significant relation between age of participants and effect size, $r = .01$, meta-regression $p = .78$. Raw data and a less smoothing curve are depicted in Figure 6. This demonstrates that while traditional meta-analytic techniques might lead to a conclusion that anti-bullying programs are just as effective among older adolescents compared to younger ones, such conclusions would be premature unless they were confirmed using within-study moderation tests.

---

4 While the present meta-analysis focused on age-related trends in reports of bullying others, for a sub-set of our studies we were also able to obtain within-study estimates for reports of being the victim of bullying ($k = 15$, 62 effect sizes). Replicating the step-function findings noted for bullying others, there was no age trend for victimization from 1st to 7th grade, $b = -.01$, $Z = .59$, $p = .55$, and a significant decline from 8th to 13th grade, $b = -.05$, $Z = 1.98$, $p = .048$. 

---
Discussion

Are anti-bullying programs effective for older adolescents, the age at which many highly visible tragedies involving bullying have occurred? Past theoretical and empirical reviews have suggested they may not be (Berger, 2007; Smith et al., 2012), but past meta-analyses conducting between-study tests of moderation found either no significant differences across age or a significant positive relationship between age and program effectiveness. The present study obtained new calculations of past studies’ effect sizes and conducted a novel three-level meta-analysis of within-study age trends. This analysis showed the theoretically predicted negative relationship, signifying reduced efficacy for older teens (or, rather, no efficacy for older teens). Using the same dataset, we were able to replicate the null or slightly positive relationship shown in past between-study moderation analyses, thus demonstrating that past findings were an artifact of a misleading methodology.

The age trends reported here are corroborated by non-experimental evidence (Finkelhor, Vanderminden, Turner, Shattuck, & Hamby, 2014). Recent analyses of a national probability sample of thousands of young people in the United States showed that elementary-aged children exposed to high-quality bullying programs reported significantly reduced bullying, as in the present study. However, adolescents in high school showed a non-significant trend in the direction of increased bullying when exposed to a high-quality program. These findings closely mirror the present study’s meta-analytic results.

The present research has a number of implications. The first calls into question the wisdom of state mandates regarding anti-bullying programs for high schools. The null effects alone suggest that it might be a misuse of resources until improved interventions are developed. The possibility of a negative effect suggests ethical concerns. Of course, ultimately it may be
effective to require schools to implement anti-bullying programs, but perhaps not until new interventions have been adapted to successfully implement the latest developmental theories and evaluated to ensure efficacy. At a minimum, the present research suggests that, at a program implementation level, it is no longer sufficient to simply “age up” existing materials that are tested with younger children, for instance by switching out the examples or the graphic art used in the activities. Instead, it may be important to fundamentally re-think the theoretical target for a given program and, in some cases, re-build the program from the ground up to address the qualitatively different causes of bullying and the changing efficacy of delivery mechanisms for older students.

Next, the present findings provide some support for existing theories of the causes and solutions to childhood bullying. Replicating past investigators, we find that anti-bullying interventions were modestly effective for younger children in grades 1-7, roughly $d = .15$, at least in the selected sub-sample of studies that met our inclusion criteria.

Yet our results also suggest that bullying in high school and the final year of middle school may be more poorly understood than in lower grades on average, at least among designers of the kinds of interventions that are a part of the published record. In the first part of this paper, we attempted to review findings and theories that may explain this. These involved two broad theoretical areas: developmental theory specific to bullying and developmental theory about domain-general behavior-change strategies (Table 1). We hope the present review will lead to improved and developmentally-appropriate anti-bullying intervention, and be informative for other social problems, including substance abuse, obesity, or academic achievement.

We do not have direct evidence that the theoretical reasons to expect age-related moderation does, in fact, apply to the meta-analytic findings presented here. However some
initial evidence for the possibility that older adolescents react more negatively to anti-bullying lessons within a comprehensive social-emotional-focused middle school program is beginning to emerge. For instance, early analyses of data from a school-level randomized experiment conducted by Espelage, Low, Polanin, and Brown (2013) shows no declines in teacher-rated engagement with the anti-bullying content for youth in grades 6 to 7th, but a sharp and statistically significant decline in 8th grade, such that 8th graders were much more disengaged from the lessons on bullying compared to younger peers.

Next, this paper reinforces the methodological point that has been made by some leading meta-analytic experts (e.g., Cooper & Patall, 2009), namely that between-study tests of moderation should be treated with caution (see Figure 1). Between-study analyses, at least in the field of bullying, introduce undesirable confounds in that both the content of the bullying intervention and the age of the participants vary across studies. At best, this introduces measurement error, leading to a need for a very large set of studies in order to detect reliable between-study relations. At worst, this can lead to systematic biases. Building on the logic of ecological fallacies and Simpson’s Paradox, between-study meta-analyses aggregate up a level—from age groups to studies—producing one effect size estimate when in fact several were present. Just as comparing average levels of a personality trait across nations often leads to different conclusions about the correlates of that trait than comparing average levels within nations, so too can the relationship at the study level misrepresent significant variance within studies (for a discussion of this, see Cooper & Patall, 2009). Concretely, in the present case, the between-study correlation of age with effectiveness was $r = .02$, while the average within-study correlation with age was $r = -.67$. Clearly, the former masked an important relation.
More generally, the present findings may serve as a cautionary tale about the interface between public policy and developmental science. Existing meta-analytic evidence provided some justification for the push to mandate anti-bullying programs for high school students. However sometimes it may be important to go the extra step—in the present case, obtaining raw data calculations for within-study effect sizes—to truly address the sub-groups of students affected by the policies (and it is also important for primary study authors, ourselves included, to report effect sizes and confidence intervals within important sub-groups whenever possible). When such additional data collection is not possible or severely limited, then it will be important to consult theory and be modest about conclusions and recommendations.

**Potential Limitations**

In this spirit, we note that the present investigation has a number of limitations. First, our dataset is limited by non-response from authors. It is possible that some of the excluded studies might have showed different age-related trends than those included. However, it is important to note that the included studies involved a great deal of data—over 350,000 participants in grades 1 – 13 from around the world. In addition, the studies that were not able to provide data were systematically those that (a) were older, and (b) used non-experimental, non-causal methodology. Our response rate for experimental or quasi-experimental studies with large samples conducted in the past 3-8 years was high. Thus, the present meta-analysis can to some extent be reasonably considered a reflection of the latest and best efforts at bullying prevention.

Next, it is possible for some of the negative effects reported in evaluations of past programs to be due to well-known problems with reference bias in self-reports (Biernat, 2003). That is, it is possible that an intervention that makes adolescents aware that certain behaviors—like exclusion—count as “bullying” may lead them to report that they do it more often (for other
research showing that a group’s frame of reference may change post-intervention, leading to counter-intuitive directional effects of self-reports, see Tuttle et al., 2013, or Dobbie & Fryer, 2013). Investigating this possibility, however, is beyond the scope of the present meta-analysis and is perhaps better addressed through future research studies that employ improved observational or behavioral measures. Furthermore, if reference bias is a possibility, then the positive findings too must be suspect. Thus, the reference bias problem does not plausibly seem to be a cause of the age-related trends documented here—that is, unless it could be shown that there are developmental differences in the direction of reference bias.

Finally, while we found that on average anti-bullying interventions had no (or a negative) effect among older adolescents in high school, this does not mean that there are no cases of programs that detected significant differences in this age group. For instance, our meta-analysis only included studies that could provide estimates from multiple age groups, but some studies that were excluded from our search had been conducted only with older adolescents. By inspecting lists of studies included in past meta-analyses (e.g., Ttofi & Farrington, 2011) we found two studies that reported at least one significant beneficial effect in at least one sub-group of students (Evers, Prochaska, Van Marter, Johnson, & Prochaska, 2007; Huey & Rank, 1984). Yet many other excluded studies found null effects or harmful effects (Cowie & Olafsson, 2000; Lee, Hallberg, & Hassard, 1979; Merrell, 2004; Schumacher, 2007; Peterson & Rigby, 1999; Tierney & Dowd, 2000). While it may be tempting to look at the former group to learn about effective intervention for older teens, the results of the present meta-analysis and the failures to find beneficial effects in many other cases suggests we cannot rule out the possibility that past significant beneficial effects were due to chance or due to other analytic procedures that might have resulted in false positive findings (Simmons, Nelson, & Simonsohn, 2011).
Conclusion

Altogether, the present analysis supports the conclusion that we cannot yet confidently rely on anti-bullying programs for grades 8 and above. It seems safer to do so with younger children and adolescents in grades 7 or below, but with the caveat that these effects are modest (at least in our estimation) and likely to vary across context and implementation.

Does this mean that schools and researchers should not attempt to change bullying among older adolescents? No. There have been exciting advances in psychological theory that have provided a wide variety of possible targets for intervention in this age group (e.g., Rodkin et al., 2013; Rudolph et al., 2011; Yeager et al., 2013). At the same time, there have been important improvements to behavior-change techniques—that is, techniques that “nudge” people’s behavior or that “redirect” their thinking (Thaler & Sunstein, 2008; Wilson, 2011; see Yeager & Walton, 2011), but these are under-utilized in existing programs (for a rare exception, see Paluck & Shepard, 2012). One high priority will be to move this basic psychological research more quickly into programs that reduce bullying. A related priority will be to test whether simply changing the delivery mechanisms for existing interventions, making them more developmentally appropriate and psychologically “wise,” might unlock latent effectiveness for existing approaches. Despite the difficulties faced by past research, we believe that a worthy investment of research resources is to address the highly important—and, we continue to believe, solvable—problem of reducing bullying among older adolescents.
References


http://dx.doi.org/10.1177/0143034306067286

http://dx.doi.org/10.1002/ab.20000

http://dx.doi.org/10.1001/jama.284.6.723

http://dx.doi.org/10.1177/082957350401900105


Declines in Efficacy of Anti-Bullying Programs


Declines in Efficacy of Anti-Bullying Programs


Heckman, J., & Kautz, T. (in press) Fostering and measuring skills: Interventions that improve


Marín-Martínez, F., & Sánchez-Meca, J. (1999). Averaging dependent effect sizes in meta-
Declines in Efficacy of Anti-Bullying Programs


Declines in Efficacy of Anti-Bullying Programs


http://dx.doi.org/10.1111/j.1467-8624.2010.01437.x


http://dx.doi.org/10.1037/a0029389


http://dx.doi.org/10.1037//0033-2909.99.3.400

Journal of Educational Psychology, 81, 654-666. http://dx.doi.org/10.1348/2044-8279.002003


Schumacher, P. J. (2007). To what extent will a sixty second video on bullying produced by high school students increase students’ awareness of bullying and change their attitudes to reduce acceptance of bullying in a high school environment? (Unpublished master’s thesis), University of Pennsylvania, PA.


Smith, P. K. (2011). Why interventions to reduce bullying and violence in schools may (or may not) succeed: Comments on this Special Section. *International Journal of Behavioral Development, 35*, 419-423. http://dx.doi.org/10.1177/0165025411407459


http://dx.doi.org/10.1111/1467-9604.00151


Declines in Efficacy of Anti-Bullying Programs 56

Ann Arbor, MI: Inter-university Consortium for Political and Social Research (ICPSR).

Retrieved from http://dx.doi.org/10.3886/ICPSR33081.v1


Child Development, 84, 970-988. http://dx.doi.org/10.1111/cdev.12003
Table 1. Reasons to Expect Developmental Differences in Anti-bullying Program Efficacy

<table>
<thead>
<tr>
<th>Developmental differences</th>
<th>Explanation for differential intervention efficacy</th>
<th>Application to bullying prevention</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Changes in manifestation of problematic behavior</td>
<td>The manifestation of the behavior targeted by the intervention occurs at less problematic levels at later developmental stages; other manifestations reach more problematic levels.</td>
<td>Across adolescence, directly observable forms of aggression / bullying (hitting or insulting) become less prevalent than more indirect forms of aggression / bullying (rumors or exclusion).</td>
</tr>
<tr>
<td>2. Changes in underlying causes of the problematic behavior</td>
<td>With age, a causal antecedent of a problematic behavior may become less influential, and so interventions addressing that antecedent could be less effective.</td>
<td>Bullying interventions that are effective for younger children may be addressing causes of bullying that are less relevant for older adolescents.</td>
</tr>
<tr>
<td>Social skills</td>
<td>The self-regulatory or social skills required to refrain from a problematic behavior at one age may differ from those required at a later age.</td>
<td>Bullying may be caused by poorer social skills among younger children, but among older adolescents these skills may facilitate indirect forms of bullying (e.g. exclusion / spreading rumors).</td>
</tr>
<tr>
<td>Motives or beliefs</td>
<td>The motives and beliefs that contribute to a behavior at an older age may differ from those that matter at a younger age.</td>
<td>Older adolescents may become more motivated to demonstrate social status than at younger ages; failing to address this may lead to reduced program effectiveness.</td>
</tr>
<tr>
<td>Settings</td>
<td>The social structure of the environment or the available choices can differ at later ages.</td>
<td>Educational transitions, such as the transition to high school, can shuffle friendship networks; this may create a greater concern about social relationships and therefore conformity to peers. AND older adolescents may have greater access to technology, making it easier for them to engage in online forms of bullying.</td>
</tr>
<tr>
<td>3. Changes in the efficacy of domain-general behavior-change techniques</td>
<td>Techniques for changing behavior among younger children may not compel older adolescents.</td>
<td>Bullying program elements—direct instruction, role-playing, school assemblies—may be less effective in high school than in elementary school.</td>
</tr>
<tr>
<td>Building social skills</td>
<td>Older adolescents may resent the implication that they lack basic social or emotional skills.</td>
<td>Among older adolescents it may be more important to make skill-building non-stigmatizing and non-remedial—i.e., an honorific program to create materials that will benefit younger students.</td>
</tr>
<tr>
<td>Changing motives or beliefs</td>
<td>Older adolescents may resist adults’ transparent attempts to control what feels like a personal choice (beliefs, motives); they may require a more indirect approach.</td>
<td>Among older adolescents, it may be more important to utilize autonomy-supportive persuasive tactics (i.e., self-persuasion, descriptive norms) and language (“you might” instead of “you should”).</td>
</tr>
<tr>
<td>Adjusting settings</td>
<td>Detailed systems of reward and punishment may be less effective later in adolescence, while “nudge” approaches—which change choice architecture and avoid explicit persuasion—may be increasingly effective.</td>
<td>Among older adolescents, bullying behavior may be difficult to spot and may rapidly change with technology; so policies that directly forbid certain bullying behaviors may become quickly outdated.</td>
</tr>
<tr>
<td>First author (year)</td>
<td>Type of document</td>
<td>Research design</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Baldry (2004)</td>
<td>Journal article</td>
<td>Experimental; pre-/post-test</td>
</tr>
<tr>
<td>Beran (2005)</td>
<td>Journal article</td>
<td>Experimental; pre-/post-test</td>
</tr>
<tr>
<td>Cross (2012)</td>
<td>Journal article</td>
<td>Single-group; pre-/post-test</td>
</tr>
<tr>
<td>Ertesvag (2007)</td>
<td>Journal article</td>
<td>Single-group; pre-/post-test</td>
</tr>
<tr>
<td>Flannery (2003)</td>
<td>Journal article</td>
<td>Experimental; pre-/post-test</td>
</tr>
<tr>
<td>Frey (2005)</td>
<td>Journal article</td>
<td>Experimental; pre-/post-test</td>
</tr>
<tr>
<td>Gollwitzer (2006)</td>
<td>Journal article</td>
<td>Experimental; pre-/post-test</td>
</tr>
<tr>
<td>Hanewinkel (2004)</td>
<td>Book Chapter</td>
<td>Single-group; pre-/post-test</td>
</tr>
<tr>
<td>Hunt (2007)</td>
<td>Journal article</td>
<td>Experimental; pre-/post-test</td>
</tr>
<tr>
<td>Kärnä (2011a)</td>
<td>Journal article</td>
<td>Experimental; pre-/post-test</td>
</tr>
<tr>
<td>First author (year)</td>
<td>Type of document</td>
<td>Research design</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Kärnä (2011b)</td>
<td>Journal article</td>
<td>Cohort-longitudinal</td>
</tr>
<tr>
<td>Kärnä (2012)</td>
<td>Journal article</td>
<td>Experimental; pre-/post-test</td>
</tr>
<tr>
<td>Koivisto (2004)</td>
<td>Book chapter</td>
<td>Single-group; pre-/post-test</td>
</tr>
<tr>
<td>Salmivalli (2005)</td>
<td>Journal article</td>
<td>Single-group; pre-/post-test</td>
</tr>
<tr>
<td>Wong (2011)</td>
<td>Journal article</td>
<td>Quasi-experimental</td>
</tr>
</tbody>
</table>

Note: * Red dots in sparklines correspond to negative (harmful) effects. Descriptions of interventions adapted from summaries provided by previous meta-analyses such as Ttofi and Farrington (2011).
Figure 1. Simplified illustration that (A) within-study and (B) between-study moderation tests in meta-analyses can produce age-related trends in the opposite direction. Note: Studies were selected to more clearly illustrate the potential for the two techniques to produce opposite developmental trends; see Figure 5 for all effects from all studies. Grade levels were converted to United States grade levels for comparability. The top panel involves separate effect sizes for each of the age groups in the study. The bottom panel involves the average effect across all grade levels, and the average grade level, for each of the three studies.
Figure 2. The form of bullying changes across development: Evidence from a nationally-representative sample. Note: Values reflect answers to survey questions asking students to report whether they were the victims of bullying. Loess smoothing curves. Gray areas represent 1 standard error of the mean at a given grade level. Each line’s slope is significantly different from zero and from each other at ps < .001. Excludes students who reported not being bullied. Source: Authors’ analyses of the aggregated 2009 and 2011 United States National Crime and Victimization Survey (NCVS).
Figure 3. An anti-bullying effort that may not be as effective with high school students.
Source: Robers, Kemp, Truman, and Snyder (2012).
Figure 4. Flow chart for creation of sample of effect sizes used in meta-analysis.

Note:

- \( n \) = number of meta-analyses
- \( k \) = number of studies

Declines in Efficacy of Anti-bullying Programs 65
Figure 5. Anti-bullying efficacy declines with age: Raw and predicted effects. Note: Loess smoothing curves. In the left panel, each dot corresponds to an effect at a given age within a study; within studies, effects within studies are centered on the grand mean to facilitate comparisons across studies but are unweighted for sample size. The left panel does not weight for sample size. In the right panel, values are estimates that were predicted from the three-level meta-analysis regression, and it weights observations based on sample size.
Figure 6. Between-study moderation by age produces no effect of age on efficacy: Raw data. Note: Loess smoothing curve.
Appendix 1. Figure 2: The National Crime and Victimization Survey, 2009-2011.

Here we outline the data sources and analysis methods used to produce Figure 2.

Data Source and Sampling

In the analysis of the developmental trends of bullying and victimization, we used data from the School Crime Supplement (SCS) to the National Crime Victimization Survey (NCVS). The SCS is a supplement to the NCVS that is designed to collect national representative information about students and school characteristics associated with school crime. The SCS has been conducted since 1989, 1995, 2001, 2003, 2005, 2007, 2009, and 2011, and our analysis includes data collected in 2009 and 2011.

The SCS survey sample was selected from the NCVS household members aged 12 through 18. In terms of household sampling, the NCVS adopted a rotating panel design. Households were initially selected into the NCVS sample by using a stratified, multistage cluster design. The individual sampling criteria for the SCS require students to be aged between 12 and 18; to be in grades 6 through 12; to be currently enrolled in primary or secondary education programs leading to a high school diploma; and not to have been exclusively homeschooled during the school year. Only one person from a household can become a panel of the SCS survey. These individuals had been enrolled sometime during the 6 months prior to the first interview. The SCS was administered to all eligible respondents between January and June of the year of data collection. In 2009, total 4,414 respondents completed the questionnaires related to students’ experiences with bullying whereas in 2011, total 5,755 respondents were included in the analysis.

Measures

To analyze the content of bullying, we combined questionnaire items into two categories: Direct, and Indirect Bullying. First, Direct Bullying was obtained by the number of student reporting “Yes” to at least one of below two questionnaires:

- “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already. During this school year, has any other student bullied you? That is, has another student.... Made fun of you, called you names, or insulted you? (VS0073)”

- “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already. During this school year, has any other student bullied you? That is, has another student.... Pushed you, shoved you, tripped you, or spit on you? (VS0076)”

Second, to examine the prevalence of Indirect Bullying, below questionnaire items were used. Same as Direct Bullying, we calculated the number of student reporting “Yes” to at least one of
these two questions. Missing cases—“Don’t know”, “Refused”, “Residue”, or “Blank (SCS noninterview or Out of universe)”—were excluded from the final analysis sample.

- “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already. During this school year, has any other student bullied you? That is, has another student.... *Spread rumors about you?* (VS0074)”

- “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already. During this school year, has any other student bullied you? That is, has another student.... *Excluded you from activities on purpose?* (VS0078)”

The overall prevalence of any type of bullying was defined by the total number of student in each grade reporting “Yes” to at least one of below seven questionnaires.

- “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already. During this school year, has any other student bullied you? That is, has another student.... *Made fun of you, called you names, or insulted you?* (VS0073)”

- “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already. During this school year, has any other student bullied you? That is, has another student.... *Spread rumors about you?* (VS0074)”

- “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already. During this school year, has any other student bullied you? That is, has another student.... *Threatened you with harm?* (VS0075)”

- “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already. During this school year, has any other student bullied you? That is, has another student.... *Pushed you, shoved you, tripped you, or spit on you?* (VS0076)”

- “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already. During this school year, has any other student bullied you? That is, has another student.... *Tried to make you do things you did not want to do, for example, give them money or other things?* (VS0077)”

- “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events
you told me about already. During this school year, has any other student bullied you? That is, has another student.... Excluded you from activities on purpose? (VS0078)”

• “Now I have some questions about what students do at school that make you feel bad or are hurtful to you. We often refer to this as being bullied. You may include events you told me about already. During this school year, has any other student bullied you? That is, has another student.... Destroyed your property on purpose? (VS0079)”

To investigate the developmental trends of bullying, a variable for students’ grade level was also used. In this questionnaire, the students were instructed to choose one answer from: Fifth or under; Sixth; Seventh; Eighth; Ninth; Tenth; Eleventh; Twelfth; Other; College/GED/Post-graduate/Other noneligible. However, we only included the cases of sixth through twelfth grade.

• “What grade are you in? (VS0017)”.

**Data Analysis**

In the present data analysis, we aggregated the 2009, and the 2011 SCS data into single dataset in order to obtain more robust developmental trends. Prior to the analysis, we applied the SCS final person weight variable (VS0142) to the raw responses in order to estimate population-level trends of bullying behavior. Therefore, the number of responses used in the analysis reflects statistically estimated responses from the corresponding age group.

To compare the developmental trends of direct and indirect bullying behaviors, we first calculated $B_{grade\ i}$, the overall prevalence of bullying by grade 6 through 12. More specifically, $B_{grade\ i}$ is defined by the total number of students who reported “Yes” at least once to the questionnaires VS0073 through VS0079. And then, $D_{grade\ i}$, the prevalence of direct bullying was calculated in grade 6 through 12. Here, $D_{grade\ i}$ is expressed by the total number of students who reported “Yes” at least once in the questionnaires VS0073 (insulting) and VS0076 (hitting). Likewise, $I_{grade\ i}$, the prevalence of indirect bullying was obtained by the total number of students who reported “Yes” at least once in the questionnaires VS0074 (spreading rumors) and VS0078 (excluding from activities). Finally, the proportions (%) of direct bullying, and indirect bullying were obtained by grade level ($6^{th}$ through $12^{th}$) as the following equations:

$$%\ Direct\ Bullying_{grade\ i} = \frac{D_{grade\ i}}{B_{grade\ i}}$$

$$%\ Indirect\ Bullying_{grade\ i} = \frac{I_{grade\ i}}{B_{grade\ i}}$$

Using these equations, we were able to rule out the possible confounding effect caused by the differential level of overall prevalence of bullying in each school grade.
Appendix 2. Search Procedures for Meta-Analysis

Search Strategy for Meta-Analyses of Anti-Bullying Interventions

(meta-analysis OR review OR synthesis) AND (violence OR aggression OR bully* OR victim*)

Search Strategy for Individual Evaluations of Anti-Bullying Interventions

(intervention OR prevention OR training OR program) AND (violen* OR aggress* OR bully* OR bullies OR victim* OR “conflict resolution”) AND (experiment* OR quasi-experiment* OR evaluation OR effect OR outcome)

Note: * indicates truncation to allow various forms other word to appear.