

A META-ANALYSIS OF THE WHAT WORKS IN CHARACTER EDUCATION RESEARCH

Keith Johnson

Fairleigh Dickinson University

Robert E. McGrath

Fairleigh Dickinson University

Melinda Bier

University of Missouri-St. Louis

Mitch Brown

University of Arkansas

Marvin W Berkowitz

University of Missouri-St. Louis

The *What Works in Character Education* project (WWCE; Berkowitz & Bier, 2007) was a descriptive summary of significant findings in the empirical literature on character education for the period 1945 to 2004. It employed a broad definition of character education to include related positive youth development programs such as socioemotional learning. Drawing on studies carried out in K–12 education settings, it encompassed research on published character development curricula as well as research on particular instructional practices such as moral discussion and cooperative learning. The current meta-analysis revisits WWCE by applying meta-analytic techniques to the same set of studies collected as part of the WWCE review. While the current analysis did not include any studies not covered by WWCE, it did modify its approach to examining 5 systematic reviews for which Berkowitz and Bier (2007) looked to the overall conclusions. In contrast, the current analysis expanded these reviews to individually examine each of their respective studies. It complements the already substantial referencing and discussion of the WWCE that exists in the character education literature. Random effects meta-analysis of the original WWCE studies examined 836 comparisons across 64 studies involving 96,930 participants and yielded a small but statistically significant average effect size, $g = 0.33$, 95% CI = [0.21, 0.45]. This finding is consistent with prior evidence on the effect size associated with various educational interventions (Lipsey & Wilson, 1993) and reaffirms that character education, as conceptualized and analyzed in the original WWCE, is effective in positively influencing character outcomes. The meta-analytic techniques also yielded new evidence of substantial heterogeneity of effects across studies, suggesting the potential value of focusing more narrowly, in future meta-analyses, on the character education programs that appear most promising. In the meta-analysis reported here, there were also indications of possible bias in some studies' reporting of results, but overall conclusions regarding effectiveness remained valid even after correcting for possible bias.

Keywords: character education, socioemotional learning, positive youth development, What Works in

Character Education, meta-analysis

• **Correspondence concerning this article should be addressed to:** Robert McGrath, mcgrath@fdu.edu

Journal of Character Education, Volume 18(1), 2022, pp. 87–112
Copyright © 2022 Information Age Publishing, Inc.

ISSN 1543-1223
All rights of reproduction in any form reserved.

Character education has been a topic of renewed interest for many years (Berkowitz & Bier, 2007; Leming, 1993; Lickona, 1991; McClellan, 1999; Mulkey, 1997; Sojourner, 2012). Character education programs have garnered significant private and public support, with substantial funding from both corporations and government bodies (Berkowitz & Bier, 2007; Howard et al., 2004; Murphy, 2002; Sojourner, 2012).

The largest extant review of the character education literature was published by a project called What Works in Character Education (WWCE; Berkowitz & Bier, 2005, 2007). Though WWCE is now more than 15 years old, it has had considerable influence on perceptions of character education. WWCE is described in more detail below, though in short, it was a significant step toward establishing an empirical foundation for a rapidly growing educational movement that hoped to command the respect of scholars and the confidence of practitioners. WWCE brought a new level of credibility and guidance to character education by identifying effective programs and practices, including interventions that “worked” in the sense of often being linked to positive, character-related outcomes. However, WWCE was a purely descriptive analysis that generated no statistical summation of results from the individual studies reviewed and therefore could not offer any overall conclusions that were based on a comprehensive statistical analysis. Accordingly, we considered it worthwhile to reanalyze the studies reported in WWCE using meta-analytic techniques. As elucidated below, examining the data in this way would allow us to derive an overall conclusion about the effectiveness of interventions contained within the body of literature analyzed, determine the strength of evidence for coming to this conclusion, and investigate what specific variables contribute to this effectiveness. An updated literature search covering the years since WWCE and up to 2020 was also completed (Brown et al., 2021). The current study served as the first phase in that more comprehensive review.

Given the important role the WWCE played in the history of efforts to evaluate the effectiveness of character education programs, we believed the present study deserved to be published independently of the larger effort, as a complement to the earlier WWCE publications.

As background to this study, we will first briefly define what is meant by character, and by character education in the context of the WWCE project. Second, we will provide a summary and critical review of the character education literature as well as research covering related approaches to positive youth development. Given the sheer size of the literature, this review will focus on systematic reviews covering positive youth interventions. Finally, we will summarize the methodology of Berkowitz and Bier (2007) as it provided the starting point for which the current research was conducted.

Defining Character and Character Education

Character has to do with the extent to which a person consistently thinks and acts in ways that are considered prosocial. Behaviors and attitudes thought to be indicative of character are typically ones that contribute to the functioning of a social community, or at least are considered socially desirable. This allows for a very broad array of behaviors to be treated as indicators of character, including intellectual virtues (like curiosity), performance virtues (like perseverance), and self-regulatory functioning. WWCE did not exclude studies that examined virtues other than moral/prosocial ones as long as a given study included some aspect of behavior that was considered socially valued. Character was conceptualized as a complex, dynamic, multi-dimensional psychological construct reflected in moral, intellectual, and self-regulatory functioning (Berkowitz et al., 2017; Lickona & Davidson, 2005; Linkins et al., 2014). Understood in this way, character is the whole set of psychological characteristics that enable

one to operate as an effective member of society, flourish intellectually, and function as a moral agent.

Because of the complex multidimensionality of character, it is hardly surprising that definitions of character have varied among scholars and practitioners. Nor should it surprise us that there have been diverse conceptions of what constitutes “character education” (Lapsley, & Narváez, 2006; McGrath et al., 2021; Williams, 2000). Some character education approaches are pedagogical, focused on the direct teaching of character lessons or the use of pedagogical methods such as cooperative learning and project-based learning, while other character education initiatives focus on peer relationships and community involvement. Some approaches use fictional and/or nonfictional literature to teach character, while others emphasize mentorship. Still others focus on modifying school culture as a whole (Berkowitz & Bier, 2007). The original WWCE project also included programs that focused on the teaching of core values, prosocial behaviors, socioemotional reasoning, and conflict resolution skills (Berkowitz & Bier, 2007). As WWCE and the present study aimed to be inclusive of programs intended to promote development of character, character education was also defined as any program intended to encourage positive youth development (see Berkowitz & Bier, 2005, 2007; Berkowitz et al., 2017).

This ecumenical approach allows for a great deal of diversity in programs considered for inclusion in this review. In this context, “character education” encompasses any type of positive youth development program, rather than only including programs that explicitly labeled themselves as character education (Berkowitz & Bier, 2007). Programs focused on youth development can be considered character education if they target some aspect of social investment, including moral values, sociomoral reasoning, knowledge of ethical issues, moral emotional competencies, or moral identity; behavioral competencies such as conflict resolution skills; and/or characteris-

tics that support prosocial behavior such as traditional character strengths (Berkowitz & Bier, 2007). Since most character education programs focus on the period from kindergarten through high school, the examination of the character education literature can be bounded to some degree by focusing on this period.

The What Works in Character Education Project

As one of the first systematic reviews of the research literature on character education programming, the WWCE project yielded broad, yet important, conclusions. The single most important finding was that character education can positively affect the character development of school students. Other major findings included evidence that character education can affect many aspects of character development; character education programs take a wide variety of forms; and character education programs tend to include multiple components (Berkowitz & Bier, 2005, 2007).

Despite its historic importance as the first large-scale review of the character education literature, limitations of the project must be recognized. One was the focus on the results of significance testing. As has been widely discussed in recent years, this is a problematic inferential strategy. Assuming some homogeneity in effects associated with character education, one implication of power analysis is that the likelihood of significant findings is partly a function of sample size. Given that some studies involved multiple schools with sometimes thousands of students, interventions of even trivial size can achieve significance though they have limited clinical value. This is particularly problematic in studies where the data analysis did not correct for dependencies among children experiencing character education in a shared classroom or school. In recent years, there has been growing recognition that violation of the *independence of observations* assumption associated with using standard *t* tests or analyses of variance to evaluate interventions administered to groups

of individuals can substantially increase the probability of a Type I error, thereby suggesting a difference from a comparison treatment when that difference really reflects varying group effects for the students receiving the two treatments (e.g., Baldwin et al., 2005; Murray, 1998). Finally, the exclusion of articles with a balance of evidence suggesting no or negative effects for character education interventions meant the statistics reviewed were not a comprehensive survey of the character education literature. Accordingly, while the WWCE project was the most thorough compilation of research literature to that time, it addressed the question of whether the literature indicated character education *could be* efficacious, and if so under what conditions. It did not address the question of *how* efficacious that literature suggested character education programs to be.

Other Reviews of the Character Education Literature

Several other attempts have been made to provide insights into the effectiveness of character education programs and practices, each with their own strengths and weaknesses. Early such efforts focused on programs that emphasized moral development and values clarification interventions (Enright et al., 1983; Lockwood, 1978; Schlaefli et al., 1985), consistent with Kohlberg's influence on the rebirth of interest in character education. As a group, they reported evidence of growth in student's moral reasoning, but did not find much support for the hypothesis that those programs had an impact on secondary outcomes such as self-esteem, personal adjustment, drug use, or academic success. However, these earlier works noted various methodological concerns, including failure to use random assignment, lack of control groups, use of inappropriate statistics, and failure to control for potential confounds. The focus on morals and values also means they are of limited value as indica-

tors of character education effectiveness in general.

At approximately the same time as WWCE, the U.S. Department of Education sponsored a large-scale effort to evaluate cross-program effectiveness, referred to as the Social and Character Development Research Consortium (2010). This involved original research simultaneously evaluating multiple character education programs with a common set of instruments, using a prospective design over the period 2004–2007, involving a large variety of schools and participants, evaluated using appropriate statistical methods such as correction for nonindependence among students. The project revealed little evidence of positive effects, and the investigators concluded that programs may even have been associated with a detrimental impact on some student outcomes. However, this research had several limitations. Only seven programs were examined, and it is unclear to what extent they represented effective programs in general. For example, only three of them were on Berkowitz and Bier's (2007) list of programs for which there was existing evidence of effectiveness. Second, the substantial majority of teachers in the control groups reported integrating social and character development strategies into their classrooms.

More recent reviews of programs explicitly identifying as character education have employed meta-analytic methods and have reported modest positive effects on academic, behavior, social, and personality outcomes (Diggs & Akos, 2016; Jeynes, 2019). Jeynes (2019) found that longer programs were also associated with larger positive effects. Several meta-analyses have also been conducted focusing specifically on socioemotional learning programs (Durlak et al., 2011; Taylor et al., 2017). These have consistently generated positive effects for a number of outcomes, including socioemotional skills and academic performance. Overall, then, results across the literature that can be broadly conceived as character education have been inconsistent.

The Present Study

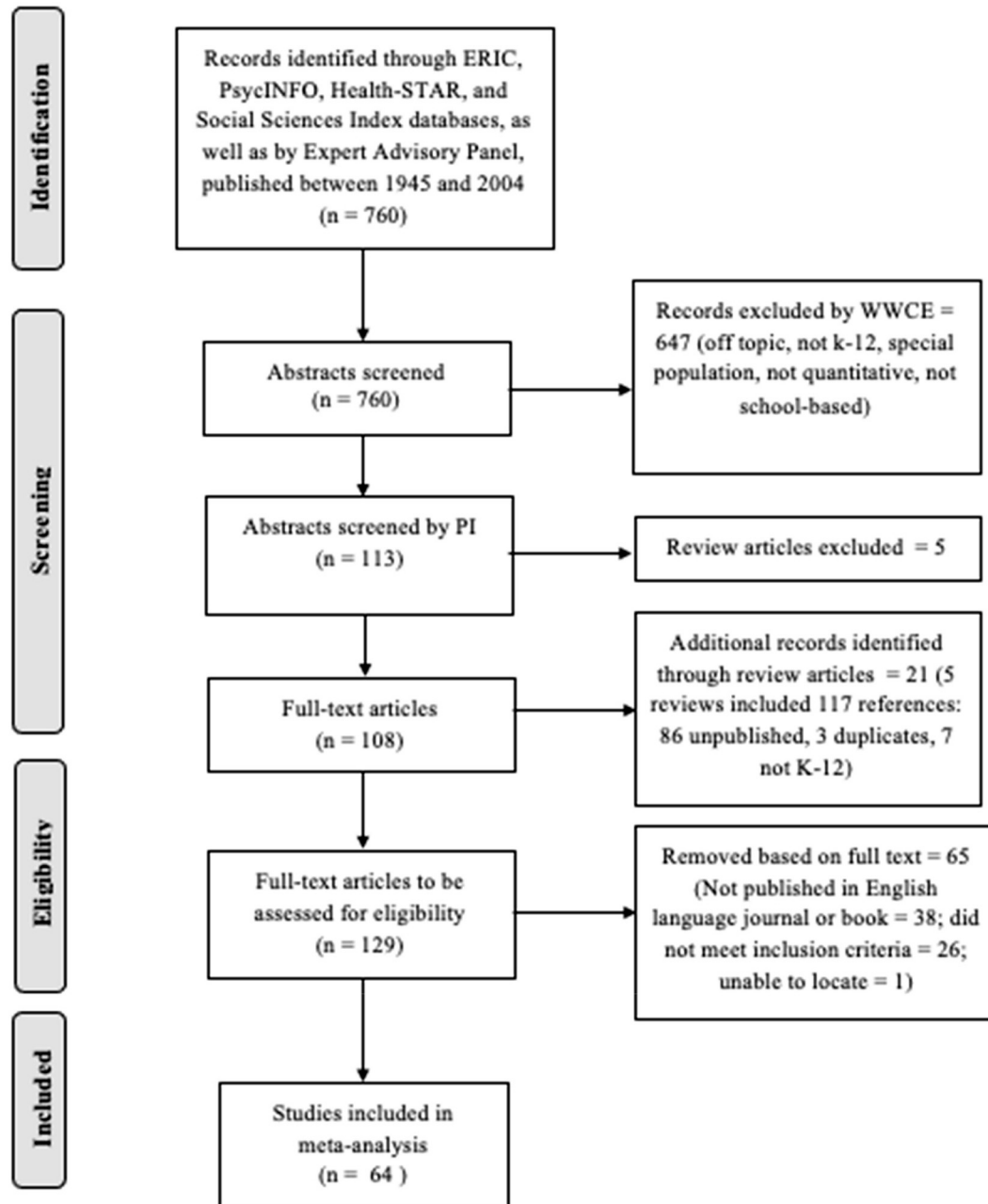
To summarize, prior attempts at reviewing the effectiveness of interventions that can be considered instances of character education have tended to use a restricted definition of character education or range of programs, or did not employ meta-analytic techniques to synthesize effects across the literature. The current study is intended to rectify these shortcomings and serve as a complement to Berkowitz and Bier (2007), augmenting their work by applying meta-analytic techniques to the WWCE literature base. Given the importance of WWCE in the story of the contemporary renewal of character education, this article focuses specifically on the effectiveness of character education in the studies that were considered for inclusion in WWCE, expanding the database beyond those that showed significant effects and changing the framing question from “What works?” to “Does it work?” The meta-analysis reported in this article examined the relationship between character education interventions and character outcomes across a range of programs, intervention targets, and outcome measures. As background to this effort, a more detailed description of the WWCE project’s methodology follows.

WWCE represented an extensive systematic review of the literature covering character education initiatives between the years 1945 and 2004 (Berkowitz & Bier, 2005, 2007). The WWCE project’s goal was to identify and review all scientifically sound research studies in character education that met their selection criteria. Given Berkowitz and Bier (2007)’s broad definition of character education, as described above, their review included a wide variety of programs and outcome measures. Various methods were used to find studies. The electronic search strategy for the WWCE project began with review of the ERIC, PsycINFO, Health-STAR, and Social Sciences Index databases for articles published between 1945 and 2004. Search terms included *school*, *evaluation*, *intervention*, *program*, and *project* as well as terms such as *prosocial*, *virtues*, *values edu-*

cation, *moral education*, *civics education*, *ethics*, *school culture*, *school climate*, and *social and emotional*. Berkowitz and Bier (2007) also assembled an advisory panel of five individuals considered experts in various domains of youth interventions, including service learning, socioemotional learning, violence prevention, drug and alcohol prevention, and teacher/classroom effects on students. Members of this panel were each asked to supply lists of studies from their fields that fit the project’s inclusion criteria. Berkowitz and Bier (2007) additionally contacted the developers and/or evaluators of then-existing programs to request further outcome evaluation reports. This process generated 760 studies (see Figure 1).

These studies were then reviewed for Berkowitz and Bier (2007)’s inclusionary criteria. To be included, studies had to have evaluated a program that could be considered character education, that is, a program that focused on youth personal development, including the teaching of moral values, prosocial behaviors, socioemotional reasoning, and conflict resolution skills. Other inclusionary criteria required that programs of interest were implemented in a kindergarten to Grade 12 setting, did not focus on students with mental and/or developmental disabilities (though samples described as “at-risk” were included), and were offered to the general student population; and the research was quantitative in nature. These criteria reduced the pool to 114 articles, one of which was not retrievable for the present study. Five of these were meta-analyses or other systematic reviews of prior literature. Berkowitz and Bier (2007) reviewed each of these again for additional criteria suggesting acceptable scientific quality and sound research design. This was defined as including a comparison group, a pre-post design or other approach to establishing equivalence between the experimental and control groups, and inclusion of significance test results. This step reduced the pool to 73 studies and five systematic reviews. These were further reduced to 64 studies and five systematic reviews, to include only those studies where the bulk of the tests reflected significant positive outcomes, as

FIGURE 1
Flow Diagram



the primary goal of WWCE was to identify attributes of effective programs rather than provide an overall judgment about the effectiveness of character education. This process resulted in 33 programs being deemed effective (Berkowitz & Bier, 2007).

Berkowitz and Bier (2007) focused on statistically significant evidence in support of character education programs from the studies featuring these 33 programs. For example, the Berkowitz and Bier (2007) reported that the most commonly reported implementation strategies among programs deemed to be effective were professional development, peer interactive teaching strategies, direct teaching strategies, family and community participation, modeling and mentoring, behavior management, and community service and service learning.

WWCE has been widely used in subsequent years to support the conduct of character education programs. The original article (Berkowitz & Bier, 2007) and spin-off publications (e.g., Berkowitz, 2011) have been cited hundreds of times, and provided the foundation for hundreds of trainings for faculty and administrators in character education. As noted, though, the project did not address the question of how well character education worked in the studies reviewed. The present study provides an addendum to this classic work, focusing on those studies drawn from the key 113 WWCE sources that provided sufficient information to generate estimates of the mean effect size between continuing education and comparison conditions. It therefore represents a meta-analysis specific to the WWCE endeavor.

METHOD

Search Strategy

Where WWCE focused on those studies of the character education literature for which the researchers reported significant results and found program effectiveness, the current research investigated whether that set of studies suggested character education generally is

effective. The meta-analysis began with the previously described pool of 113 articles. As noted above, these were all the studies or reviews Berkowitz and Bier (2007) were able to identify up through 2004 that summarized one or more quantitative evaluations of a character education program for a general or “at-risk” population of students in a K–12 setting (see Figure 1). The first step in the present project involved examining the five reviews included in this set for studies not included in the 113. That process yielded an additional 117 references. Of these, 86 were unpublished manuscripts, three duplicated prior results, and seven did not meet the selection criteria. The remaining 21 new references were added to the 108 primary sources from WWCE, raising the total to 129 studies considered for the present study.

Two graduate students conducted a full text review of each of these articles to evaluate compliance with the following additional selection criteria appropriate to the present project. The study had to be in a peer-reviewed publication; non-peer-reviewed reports, theses, and dissertations were excluded due to potential concerns about research quality. Second, studies had to include sufficient statistical information for the purposes of computing or estimating at least one effect size based on the standardized mean difference between character education and control conditions. In cases of disagreement on whether a study met these criteria, a third graduate student evaluated the study. This process resulted in the exclusion of an additional 65 studies, leaving the final pool of studies to be included in the analysis at 64. Of these, 42 (66%) were studies used in the WWCE review in Berkowitz and Bier (2007). The additional 22 studies (bolded in Table 1) either did not meet their criteria for inclusion, in that significant positive outcomes did not sufficiently outweigh negative and nonsignificant outcomes, or were articles we identified from one of the five systematic reviews covered in Berkowitz and Bier (2007), but not themselves reviewed on the individual study level.

A summary of the studies may be found in Table 1. These studies involved a total of

TABLE 1
Demographic and Other Variables From Studies

Study	Program Name	Program Length	N Intervention Group	% Female	M Age	Grade Range	Method of Assignment	Level of Assignment	Number of Schools
Aber (1998)	Resolving Conflict Creatively	> 365 days	1,799	48	—	2–6	Not random	School	15
Allen (1997)	Teen Outreach	≥ 60 days < 365	342	86	15.9	9–12	Random	Child	25
Allen et al. (1990)	Teen Outreach Program	> 365 days	632	70	15.6	9–12	Not random	Child	35
Allen et al. (1994)	Teen Outreach Program	> 365 days	1,020	71	15.7	9–12	Not random	Child	66
Battistich et al. (2000)	Child Development Project	> 365 days	826	51	—	3–6	Not random	School	24
Battistich et al. (2004)	Child Development Project	> 365 days	234	51	—	3–6	Not random	School	12
Beck et al. (1972)	Kolberg's Moral Development	≥ 60 days < 365	17	—	17	11	Not random	Classroom	2
Benninga et al. (1991)	Child Development Project	> 365 days	664	—	—	2–5	Not random	School	7
Blatt & Kohlberg (1975)	Kolberg's Moral Development	≥ 8 days < 60 days	21	44	—	—	Partially random	Child	4
Botvin et al. (1992)	Drug Abuse Prevention Intervention	≥ 60 days < 365	3,153	51	12	7	Random	School	47
Botvin et al. (1995a)	Drug Abuse Prevention Intervention	≤ 7 days	1,156	48	18	12	Partially random	School	56
Botvin et al. (1995b)	Generic Skills Training OR Culturally Focused Intervention	≥ 60 days < 365	228	53	15	9	Not random	School	6
Botvin et al. (1997)	Drug Abuse Prevention Intervention	≥ 60 days < 365	361	53	12.6	7	Not random	School	7
Caplan et al. (1992)	Positive Youth Development Program	≥ 60 days < 365	90	45	13	6–7	Matched random	School	2
Catalano et al. (2003)	Raising Healthy Children	> 365 days	497	45	7.43	9	Random	School	10
Cognetta & Sprinthall (1978)	Deliberate Psychological Education Curriculum	> 365 days	255	—	—	10–12	Not random	Classroom	1

Cognetta (1977)	Deliberate Psychological Education Curriculum	≥ 60 days < 365	15	—	—	10–12	Not random	Classroom	1
Colby et al. (1977)	Fenton's Social Studies Curricula	≥ 60 days < 365	174	—	15	8–12	Not random	Classroom	14
Conduct Problems Prevention Research Group (1999)	Promoting Alternative Thinking: Fast Track version	≥ 60 days < 365	3,357	—	6	1	Matched random	School	48
Cook et al. (2000)	Comer's School Development Program	> 365 days	1,288	—	12	6–8	Matched random	School	19
Elias et al. (1986)	Interpersonal Cognitive problem Solving	≥ 60 days < 365	79	49	—	5	Not random	School	4
Farrell et al. (2001)	Responding in Peaceful and Positive Ways (RIPP)	≥ 60 days < 365	305	50	11.7	6	Random	Classroom	3
Farrell et al. (2003)	Responding in Peaceful and Positive Ways (RIPP)	≥ 60 days < 365	239	54	12.8	7	Random	Classroom	2
Felner et al. (1982)	Transition Project	≥ 60 days < 365	59	—	—	9	Random	Child	1
Felner et al. (1993)	Transition Project	≥ 60 days < 365	1,204	—	—	9	Random	Child	1
Flay (2003)	Positive Action Program	> 365 days	24	—	16	7–12	Not random	School	36
Flay et al. (2001)	Positive Action Program	> 365 days	8,280	—	9	K–6	Not random	School	36
Gottfredson (1986)	Project PATHE (Positive Action Through Holistic Education)	> 365 days	1,928	—	—	7–12	Not random	School	9
Greenberg et al. (1995)	Promoting Alternative Thinking	> 365 days	83	42	8	2–3	Random	School	4
Grossman et al. (1997)	Second Step: Violence Prevention	> 365 days	418	46	8.2	2–3	Matched random	School	12
Harrington et al. (2001)	All Stars	≥ 60 days < 365	629	55	12	6–7	Random	School	14
Hawkins et al. (1992)	Seattle Social Development Project	> 365 days	919	48	10.5	1–4	Partially random	Child	8
Hawkins et al. (1999)	Seattle Social Development Project	> 365 days	149	50	11	5	Not random	Child	18

(Table continues on next page)

TABLE 1
(Continued)

<i>Study</i>	<i>Program Name</i>	<i>Program Length</i>	<i>N Intervention Group</i>	<i>% Female</i>	<i>M Age</i>	<i>Grade Range</i>	<i>Method of Assignment</i>	<i>Level of Assignment</i>	<i>Number of Schools</i>
Hawkins et al. (2001)	Seattle Social Development Project	> 365 days	200	50	12	1–6	Not random	School	18
Hayden (1981)	Kolberg's Moral Development	≥ 60 days < 365	12	100	12	7	Random	Child	1
Johnson et al. (1995a)	Generic peer mediation program	> 365 days	144	51	—	2–5	Not random	School	1
Johnson et al. (1995b)	Peer Mediation	> 365 days	311	50	9	2–5	Random	Classroom	1
Keasey (1973)	Kolberg's Moral Development	≤ 7 days	18	50	12	5–6	Not random	Child	1
Knoff & Batsche (1995)	Project Achieve	> 365 days	672	—	10	Pre-K–5	Not random	School	2
Krug et al. (1997)	Peace Builders	> 365 days	2,393	—	9	K–6	Random	School	7
Leming et al. (2000)	An Ethics Curriculum for Children	≥ 60 days < 365	426	—	9	1–6	Not random	School	4
McNeal et al. (2004)	All Stars	≥ 8 days < 60 days	1,822	54	12	—	Matched random	School	14
Metropolitan Area Child Study Research Group (2002)	Yes I Can	> 365 days	1,518	40	9	1–6	Random	School	16
O'Donnell et al. (1995)	Seattle Social Development Project	> 365 days	60	54	12.4	6	Not random	Child	8
Orpinas et al. (1995)	Second Step Violence Prevention Program	≥ 60 days < 365	223	53	12	6	Not random	Classroom	4
Orpinas et al. (2000)	Students for Peace	> 365 days	2,246	50	13	6–8	Matched random	School	8
Reid (1999)	Linking the Interests of Families and Teachers (LIFT)	≥ 60 days < 365	382	49	—	—	Partially random	School	12
Reyes & Jason (1991)	Transition Project	≥ 60 days < 365	77	—	—	9	Random	Child	1

Schultz et al. (2001)	Facing History and Ourselves	≥ 60 days < 365	346	52	14	8	Not random	Classroom	—
Shapiro et al. (2002)	Peacemakers	≥ 60 days < 365	1,293	50	—	4–8	—	School	9
Shope et al. (1998)	Substance Abuse Prevention Lesson	≤ 7 days	187	52	13	6–7	Partially random	Classroom	—
Shure & Spivack (1979)	Interpersonal Cognitive problem Solving	≥ 8 days < 60 days	39	—	4	Pre-K–K	Not random	School	—
Shure & Spivack (1980)	Interpersonal Cognitive problem Solving	≥ 8 days < 60 days	113	56	4.3	Pre-K–K	Not random	School	20
Shure & Spivack (1982)	Interpersonal Cognitive problem Solving	≥ 8 days < 60 days	113	56	4.3	Pre-K–K	Not random	School	8
Solomon et al. (2000)	Child Development Project	> 365 days	1,018	51	—	3–6	Not random	School	24
Spoth et al. (2002)	Life Skills Training (LST)	≥ 60 days < 365 days	503	47	12.5	7	Matched random	School	36
Taub (2002)	Second Step Violence Prevention Program	> 365 days	44	—	10	3–5	Not random	School	2
Taylor et al. (1999)	Across Ages	> 365 days	193	53	—	—	Random	School	3
Taylor et al. (2002)	Social Competency Program	> 365 days	96	50	12	6	Not random	Child	—
Turiel (1966)	Kolberg's Moral Development	≤ 7 days	11	—	—	7	Random	Child	—
Twenlow et al. (2001)	Social Systems/ Psychodynamic Antiviolence Intervention (Generic)	> 365 days	48	—	—	3	Not random	School	2
Van Schoiaek-Edstrom et al. (2002)	Second Step Violence Prevention Program	≥ 61 days < 365	398	51	13	6–8	Partially random	Classroom	5
Walker (1979)	Tracy and Cross (1973) Moral Development Paradigm.	≤ 7 days	44	100	16.2	10–11	Random	Child	1
Walker (1980)	Kolberg's Moral Development	≤ 7 days	12	50	12	4–7	Random	Child	2

96,930 participants. About 28% of studies studied a majority Black sample, participants in 38% of studies were predominantly White, and 9% of studies included predominantly Hispanic participants. The majority of studies (57%) took place in an urban environment, while 20% occurred in a suburban environment and 5% were in rural settings. Only 26 of the 64 studies discussed issues of risk among their participants. Of those, 40% identified their participants as being at-risk in some way. The average sample size was 709 (median = 280; range = 11–8,280 participants). The average number of schools per study was 12.7 (median = 7.5; range = 1–66 schools). In a majority of studies, experimental and control groups were housed in different schools. Forty-two percent of the 64 studies used random assignment at the individual, class, or school level; 50% of the studies did not, typically due to investigators using convenience samples from schools that had already been implementing the youth intervention program of interest. The remaining studies involved a mix of random and nonrandom assignment of schools or students, or did not describe the method of assignment sufficiently. For additional information about assignment and program characteristics, please refer to Table 2.

There were 36 different youth intervention programs featured across the 64 studies, including 29 established/branded programs, and seven site-specific programs. The most common target of intervention involved “productive” elements of character (43%) such as school adjustment, social skills, or problematic behaviors; followed by outcomes reflecting moral functioning such as moral reasoning, ethical sensibility, or racist attitudes. An additional 18% reflected more cognitive or intellectual contributors to character such as knowledge of various topics. A final 9% of outcomes did not fit into any of these three categories. Examples of such outcomes included expectations for peer/adult substance use, visits to the school nurse, and awareness of social norms.

Data Extraction

All data were initially extracted by graduate students, again with two graduate students reviewing and extracting data for each study. Discrepancies were resolved by one of the current authors, who double checked the data in question for accuracy. Since WWCE included all findings from an article, results for a study were aggregated across subgroups, outcome measures, interventions, and time points. In addition to outcome statistics, potential moderators were also extracted consistent with WWCE, including methods of program implementation and outcome targets.

Methods of program implementation included direct teaching methods, interactive teaching methods, behavioral management strategies, modeling or mentoring, family and community participation, community service or service learning, teacher professional development, and schoolwide organizational change. Elements of study design considered for the current analysis were the method of participant assignment (random assignment or other methods) and level of participant assignment (by individual, classroom, or school).

Finally, research quality was evaluated using criteria based on the *Cochrane Handbook for Systematic Reviews of Interventions* standards for identifying systematic bias (Higgins & Green, 2011). Nine potential contributors to poor quality were independently evaluated by two graduate students according to these guidelines. A detailed explanation of each of these can be found in Table 3. Raters evaluated each study on each of the nine contributors as demonstrating a high, low, or unclear risk to research quality, resulting in a total of 576 ratings. Of these, 220 (38.2%) were instances where risk of bias was unclear due to inadequate reporting of methodology in studies. There were 203 (35.2%) instances of a high likelihood of bias and only 153 (26.6%) of a low likelihood of bias.

TABLE 2
Program and Study Characteristics

<i>Variable</i>	<i>%</i>	<i>k^a</i>
Program Length		64
≤ 7 days	11	
8–60 days	8	
61–365 days	37	
> 365 days	45	
Program Implementor		62
Untrained staff	14	
Trained staff	62	
Specialist	34	
Community member	26	
Method of Implementation		64
Direct teaching methods	68	
Interactive methods	88	
Behavioral management	17	
Modeling/mentoring	15	
Family/community participation	28	
Community service/service learning	29	
Professional development	9	
Schoolwide organization	9	
Level of Participant Assignment		64
Active		
Child	26	
Class	17	
School	57	
Control		
Child	29	
Class	15	
School	55	
Type of Outcome		
Attitudes	26	
Growth	23	
School performance	8	
Drugs	16	
Other problem behavior	20	
Knowledge	2	
Positive behavior	9	
Distress	1	

Note: ^aNumber of studies on which percents are based.

TABLE 3
Threats to Study Quality

Study	Random Sequence Generation	Allocation Concealment	Blinding of Personnel	Blinding of Outcome Assessment	Incomplete Outcomes Data Addressed	Selective Reporting	Failure to Correct for Student Clustering	Contamination Across Treatments	Failure to Check for Treatment Fidelity
Aber (1998)	High	Low	High	Unclear	Unclear	Unclear	Low	Unclear	Low
Allen (1997)	Low	Low	Unclear	Unclear	Low	Unclear	High	High	Unclear
Allen et al. (1990)	High	High	High	High	Low	High	High	High	High
Allen et al. (1994)	High	High	High	Unclear	Low	Unclear	High	High	High
Battistich et al. (2000)	High	High	Unclear	Low	High	Unclear	High	Low	High
Battistich et al. (2004)	High	High	Unclear	Low	High	Unclear	High	Low	High
Beck et al. (1972)	High	High	High	High	Unclear	Low	High	High	Unclear
Benninga et al. (1991)	High	Unclear	Unclear	Unclear	Unclear	High	High	Low	High
Blatt & Kohlberg (1975)	High	Unclear	High	High	Unclear	Low	High	High	Unclear
Botvin et al. (1992)	Low	Unclear	Unclear	Unclear	Unclear	Unclear	High	Low	Low
Botvin et al. (1995)	Low	Unclear	Unclear	Unclear	Unclear	Unclear	High	Low	Low
Botvin et al. (1997)	High	Unclear	Unclear	Unclear	High	Unclear	High	Unclear	Unclear
Botvin et al. (1995)	High	High	High	Unclear	High	Unclear	High	Low	High
Caplan et al. (1992)	Low	Low	Unclear	Low	Low	Unclear	High	High	Unclear
Catalano et al. (2003)	Low	Low	Low	Low	Unclear	Unclear	Low	Low	Low
Cognetta (1977)	High	Unclear	High	High	Unclear	Unclear	High	High	Unclear
Cognetta & Sprinthall (1978)	High	Unclear	High	High	Unclear	Unclear	High	High	Unclear
Colby et al. (1977)	High	High	Unclear	Low	High	Low	High	High	High
CPPRG (1999) ^a	Low	Low	Unclear	High	Unclear	Unclear	Low	Low	Low
Cook et al. (2000)	Low	Low	Unclear	Unclear	High	Unclear	Low	Low	High
Elias et al. (1986)	High	Unclear	High	Unclear	Unclear	Unclear	High	Low	Unclear
Farrell et al. (2001)	Low	Unclear	Unclear	Low	Low	Unclear	Low	High	Unclear

Farrell et al. (2003)	Low	Unclear	Unclear	Low	Low	Low	High	Unclear
Felner et al. (1982)	Low	High	High	Unclear	Unclear	Unclear	High	Unclear
Felner et al. (1993)	Low	High	High	Unclear	Unclear	Unclear	High	Unclear
Flay (2003)	High	Unclear	High	Unclear	Unclear	Unclear	Unclear	Unclear
Flay et al. (2001)	High	Unclear	High	Unclear	Unclear	Unclear	Low	Unclear
Gottfredson (1986)	High	Unclear	High	Unclear	Unclear	Unclear	Unclear	Unclear
Greenberg et al. (1995)	Low	Low	Unclear	Unclear	Unclear	Unclear	Low	Low
Grossman et al. (1997)	Low	Low	Unclear	Low	High	Unclear	Low	High
Harrington et al. (2001)	Low	Low	High	Low	Low	Unclear	Low	Low
Hawkins et al. (1992)	High	Unclear	High	Unclear	Low	Low	High	Unclear
Hawkins et al. (1999)	High	Unclear	Unclear	Unclear	Low	High	Low	Unclear
Hawkins et al. (2001)	High	Unclear	High	Unclear	Low	Unclear	Unclear	Unclear
Hayden (1981)	Low	Unclear	High	Unclear	Low	Low	Unclear	Low
Johnson et al. (1995a)	High	Unclear	Unclear	High	Unclear	Unclear	High	Unclear
Johnson et al. (1995b)	Low	Unclear	Unclear	Low	Unclear	Unclear	High	Unclear
Keasey (1973)	High	Unclear	High	Unclear	Unclear	Low	Low	Unclear
Knoff & Batsche (1995)	High	High	High	Unclear	Unclear	High	Low	Unclear
Krug et al. (1997)	Low	Unclear	Unclear	Unclear	High	Low	Low	Unclear
Leming et al. (2000)	High	High	High	Unclear	Unclear	High	Low	Low
McNeal et al. (2004)	Low	Unclear	Unclear	Unclear	Low	Unclear	Low	Low
MACSRG (2002) ^b	Low	Unclear	Unclear	Unclear	Low	Unclear	Low	Low
O'Donnell et al. (1995)	High	Unclear	High	Unclear	Low	Low	High	Unclear
Orpinas et al. (1995)	High	Unclear	High	Unclear	Unclear	High	High	High
Orpinas et al. (2000)	Low	Low	High	Unclear	High	Unclear	Low	High
Reid (1999)	Low	Low	High	Low	High	Unclear	High	Low
Reyes & Jason (1991)	Low	Unclear	Unclear	Unclear	Unclear	Unclear	High	Unclear
Schultz et al. (2001)	High	High	High	Unclear	Low	Low	Low	Unclear

(Table continues on next page)

TABLE 3
(Continued)

<i>Study</i>	<i>Random Sequence Generation</i>	<i>Allocation Concealment</i>	<i>Blinding of Personnel</i>	<i>Blinding of Outcome Assessment</i>	<i>Incomplete Outcomes Data Addressed</i>	<i>Selective Reporting</i>	<i>Failure to Correct for Student Clustering</i>	<i>Contamination Across Treatments</i>	<i>Failure to Check for Treatment Fidelity</i>
Shapiro et al. (2002)	Unclear	Unclear	Unclear	Unclear	High	Unclear	High	Low	High
Shope et al. (1998)	High	Unclear	Unclear	Unclear	High	Unclear	High	High	High
Shure & Spivack (1979)	High	Unclear	Unclear	Low	Unclear	Unclear	High	Unclear	Unclear
Shure & Spivack (1980)	High	Unclear	Unclear	Low	High	Unclear	High	Low	Unclear
Shure & Spivack (1982)	High	Unclear	Unclear	Low	Unclear	Unclear	High	Low	Unclear
Solomon et al. (2000)	High	High	High	Low	Unclear	Unclear	High	Low	Low
Spoth et al. (2002)	Low	Unclear	Unclear	Unclear	Low	Unclear	Low	Low	Low
Taub (2002)	High	Unclear	Unclear	Unclear	Unclear	Low	High	Low	High
Taylor et al. (1999)	Low	Unclear	High	Unclear	Low	High	High	Low	High
Taylor et al. (2002)	High	High	High	High	Unclear	Unclear	High	High	Unclear
Turiel (1966)	Low	Low	Unclear	Unclear	Unclear	Low	Low	Unclear	Unclear
Twenlow et al. (2001)	High	High	High	High	Unclear	Unclear	High	Low	Unclear
Van Schoiack-Edstrom et al. (2002)	High	High	High	High	Unclear	Low	High	High	High
Walker (1979)	Low	Low	High	Unclear	Unclear	Unclear	Low	Unclear	Low
Walker (1980)	Low	Low	High	Unclear	Unclear	Low	Low	Unclear	Low

General Analytic Strategy

All analyses for which computation or estimation of a standardized mean difference was possible were included. Meta-analytic results were generated using Comprehensive Meta-Analysis, Version 3.0 (Borenstein et al., 2013). The first step involved computing an estimate of Cohen's d for each analysis. A few studies reported rates of dichotomous outcomes rather than dimensional outcomes. These were converted to an estimated standardized mean difference using the log odds ratio (Hasselblad & Hedges, 1995). There were 15 such analyses that were excluded because of a base rate of 0 in at least one group, that is, the targeted behavior never occurred in that group. These typically had to do with substance use in younger children. In instances where t or F values were used in a cluster randomized trial and were not analyzed using mixed models or a similar strategy that corrected for clustering, the significance test value was corrected per Hedges (2007) Equation 5. Except when computed from a t or F value, which had already been adjusted, standardized mean differences and standard errors were also corrected for clustering bias (Equations 18.19 and 18.20; Hedges, 2009). Researchers rarely provided the intraclass correlation coefficient reflecting the impact of case clustering on the results, so in all cases this statistic was assumed to be .15. This value is consistent with prior studies examining group effects in educational settings (Hedges & Hedberg, 2007; Schochet, 2008; What Works Clearinghouse, 2014). A number of unusually large effects were noted, so the top and bottom 2.5% of d values were then winsorized (Martin & Roberts, 2010), and standard errors were recomputed based on winsorized values. Final d values and standard errors were then used to generate a mean Hedges's g value for each study (Hedges, 2009).

RESULTS

As space does not permit adequate explanation each of the below statistical terms or why a given analytic procedure was judged appropriate, and as doing so is out of the scope of the current research, the current authors refer readers to consult the primary sources as cited below. Given the diversity of study populations, interventions, and outcomes, the random effects model was considered appropriate (Borenstein et al., 2010). Statistical procedures for detecting and adjusting for publication bias such as Egger's test (Egger, 1997), the funnel plot (Light & Pillemer, 1984), and the trim and fill method (Taylor & Tweedie, 1998) were also computed. Potential moderating variables were analyzed via meta-regression, as outlined in Borenstein et al. (2015). There were 836 comparisons for which standardized mean differences could be reported across the 64 studies. Of these, 83% were directly estimated from means, standard deviations, and group sizes.

On average, comparisons yielded a small, statistically significant effect size, $g = 0.33$, $p < .001$, 95% CI = [0.21, 0.45] with Hartung-Knapp-Sidik-Jonkman correction (IntHout et al., 2014) for heterogeneity in effects. These findings suggest that character education, broadly defined, is effective in positively influencing character outcomes relative to control conditions, in the kindergarten to grade 12 general education setting. However, this interpretation must be tempered by significant evidence of heterogeneity, $Q(63) = 540.9$, $p < .001$. This heterogeneity is amplified by both sampling error and by heterogeneity across the populations aggregated in this study. The I^2 statistic suggested that 88% of variation in study effects was due to population variability rather than sampling error. The standard deviation of population effects was estimated at $\tau = 0.40$, which suggests more than the typical level of variability across populations found in a random effects meta-analysis (Linden & Hönemann, 2021). Not surprisingly, the prediction interval of popula-

tion effects was quite large, [0.48, 1.14], indicating substantial variability in intervention effects across treatment populations, including the potential for medium-sized deleterious effects in some settings.

Possible Selection Bias

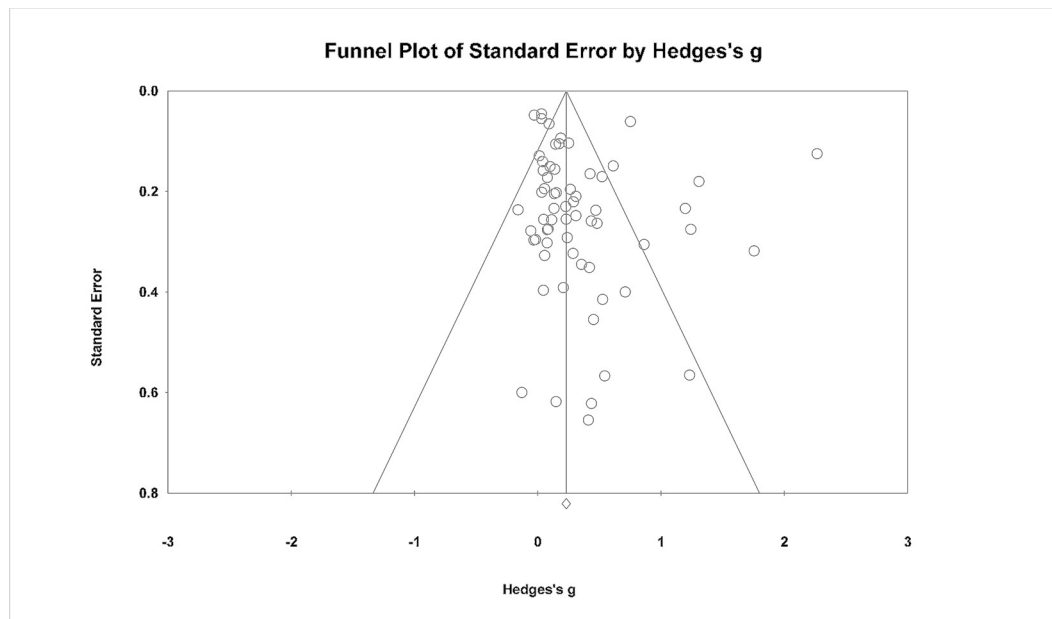
Qualitative examination of the funnel plot (see Figure 2) revealed asymmetry of effects suggestive of probable selection bias even with winsorizing of extreme values. Estimating the true effect size under the assumption that asymmetry in study effects was a result of selection bias generated a substantially reduced estimate of $g = 0.07$. However, this trim and fill method can be quite inaccurate under circumstances of heterogeneity among population effects (Peters et al., 2007; Terrin et al., 2003), which clearly applies in the present circumstances. It is also noteworthy that Egger's test of the regression intercept did not reach significance, $t(62) = 1.71$, $p = .09$, though the test of the rank correlation was, $z =$

2.8, $p < .01$. Vevea and Hedges (1995) also developed a likelihood ratio test for the null hypothesis of no bias. Using the *weightr* package in R (Coburn & Vevea, 2019) across four levels of p values for the mean effect for each study, this test was significant, $\chi^2(df = 3) = 44.99$, $p < .001$, suggesting evidence of bias. Assuming moderate one-tailed selection bias (Veeva & Woods, 2005), the corrected effect was estimated to be $g = 0.23$. Taken together, the bulk of the evidence suggests some degree of selection bias is present, resulting in some overestimation of the mean effect.

Moderating Variables

Meta-regression analyses were conducted evaluating eight program elements (direct teaching, student activities, behavioral management, school organizational change, mentoring, community/family participation in the program, service learning, and teacher professional development) as potential moderators of effects. We similarly examined 11 types of

FIGURE 2
Funnel Plot



outcome variables (personal competency, school attitudes, knowledge of character, risky behavior, antisocial behavior, school behavior, academic achievement, attendance, problem-solving, interpersonal competency, and sociomoral cognition) for variations in program impact. None of these analyses were significant. Because mentoring is of particular interest in this literature, we note that mentoring programs resulted in larger effects on average, mean g for mentoring programs = 0.38, 95% CI = [0.12, 0.64] versus 0.30 [0.20, 0.40] for those without, suggesting the possibility a larger sample of studies might have sufficient power to achieve significance.

DISCUSSION

The primary intention of this study was to provide an addendum to the WWCE project, generating an estimate of the effectiveness of the character education studies that were part of that project. Across 64 studies, character education demonstrated an overall positive impact compared to control groups, with a mean effect size of $g = .33$, indicating that on average the mean for students receiving character education was about 1/3 a standard deviation better than the mean for students in a control condition. Traditionally, this is considered a small to moderate effect across various disciplines (Cohen, 1988). However, such effects are not considered trivial, and can still be quite important in large populations such as that represented by K–12 students. The most famous example of this phenomenon is Rosenthal's (1990) demonstration that the effect of aspirin on prevention of heart attacks, which is now widespread practice, was far smaller than the effect reported here, but across the entire population has a profound effect on heart disease. It is noteworthy that the mean value reported here is the same as the mean value Lipsey and Wilson (1993) reported for all meta-analyses they found in educational research that compared interventions, .34.

However, it should be noted that estimates of the mean effect based on the assumption of selection bias generated smaller mean effects, with the procedure introduced by Vevea and Woods (2005) suggesting a mean effect of .23 even after winsorizing extreme effects. The existing information does not allow us to identify the cause(s) of potential selection bias, or even the extent to which these findings are due to heterogeneity in effects across treatment populations. There is a tendency to attribute such evidence to the "file drawer problem" (Rosenthal, 1979), that is, the tendency not to publish studies with null findings. This explanation would not seem particularly relevant in the present circumstances. Given the size of the samples and the amount of effort typically involved in conducting a school-based study, it is unlikely that researchers would forsake publication if possible. Large sample sizes are also likely to result in some significant findings assuming even a small effect, particularly given the tendency not to correct for clustered administration of treatment in the statistical analysis.

More likely, then, is the presence of some use of questionable research practices (Simmons et al., 2011), such as omitting non-significant analyses, adding covariates until an analysis becomes significant, and the aforementioned tendency to conduct tests without correction for clustered treatment. Even when corrected for selection bias, however, the mean effect is of a size that suggests the potential for practical value.

The heterogeneity of findings in this study also merits consideration. There was substantial variability in effects across studies. It is important to keep in mind that this finding does not imply ineffectiveness of treatment, only that we could not account for the variability in results based on the different educational approaches and targets we were able to compare to each other. None of the moderators we evaluated based on program or outcome variables examined in WWCE accounted for significant variation in effect sizes. A more likely interpretation has to do with variability in the

effectiveness of character education programs across settings. This interpretation is consistent with finding a prediction interval from the present results of [0.48, 1.14]. To understand the meaning of this interval, it is helpful to compare it to the confidence interval reported earlier in conjunction with the mean effect, [0.22, 0.45]. The latter helps us gauge how precise 0.33 is as a mean effect estimate across all the student populations, character education programs, and treatment outcomes considered in this meta-analysis. The fact that both endpoints of the interval point to a small-to-moderate mean effect suggests reasonable precision in the overall estimate. In contrast, the prediction interval attempts to estimate the range of effects across settings. The fact that the interval includes negative values as large as 0.48 suggests the potential for some character education programs in some settings with some populations to have deleterious effects on some outcomes.

Dramatic variation in the efficacy of different character education programs could also have contributed to the null findings of the contemporaneous Social and Character Development Research Consortium (2010), if the programs identified were of relatively low effectiveness. Unfortunately, we are unable to draw firm conclusions about the causes of relatively negative findings with the available data. For example, one of the two programs with the poorest outcomes (the largest negative mean effect sizes) was only evaluated in one study, while the other was associated with a positive mean effect in the only other study examining that program. Clearly, it would be important to identify programs that are particularly promising and conduct multiple studies to generate reliable evidence of effectiveness across settings.

Ultimately, the findings of this study support the overall effectiveness of the character education programs incorporated in WWCE. The effect size reported here is in line with similar findings from other studies in education. However, though this article represents an important addendum to the WWCE project,

that project is now more than 15 years old. Therefore what we present here cannot be considered a current profile of the character education literature. Even so, it points to goals for future character education research. Perhaps most important is the need to focus on programs of established effectiveness in order to more precisely determine the relative contribution of the different program components to overall program effectiveness.

AUTHORS' NOTES

- Keith Johnson <https://orcid.org/0000-0001-5707-6223>
- Robert E. McGrath <https://orcid.org/0000-0002-2589-5088>
- Mitch Brown <https://orcid.org/0000-0001-6615-6081>
- Marvin W Berkowitz <https://orcid.org/0000-0003-4335-6405>

Keith Johnson is now at the Central Western Massachusetts VA Healthcare System. We are grateful to the following individuals who contributed to extracting the data for this study: Francesca Bates, Tamar Blanchard, Peter Hart, Melissa Jermann, Megan McGrath, Allison Parente, Shirley-Anne Paul, Alec Twibell, Norah Wallace, and Bina Westrich.

This work is part of a larger project done in collaboration with the Center for Character and Citizenship at the University of Missouri-St. Louis, and was made possible through the support of a grant from the John Templeton Foundation (grant #61178 to Melinda C. Bier and Marvin W Berkowitz). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

REFERENCES

References marked with an asterisk indicate studies included in the meta-analysis.

- *Aber, J. L., Jones, S. M., Brown, J. L., Chaudry, N., & Samples, F. (1998). Resolving conflict creatively: Evaluating the developmental effects of a school-based violence prevention program in neighborhood and classroom context. *Development and Psychopathology*, 10, 187–213. <https://doi.org/10.1017/s0954579498001576>
- *Allen, J. P., Philliber, S., & Hoggson, N. (1990). School-based prevention of teen-age pregnancy and school dropout: process evaluation of the national replication of the Teen Outreach Program. *American Journal of Community Psychology*, 18, 505–524.
- *Allen, J. P., Kuperminc, G., Philliber, S., & Herre, K. (1994). Programmatic prevention of adolescent problem behaviors: The role of autonomy, relatedness, and volunteer service in the Teen Outreach Program. *American Journal of Community Psychology*, 22, 595–615.
- *Allen, J. P., Philliber, S., Herrling, S., & Kuperminc, G. P. (1997). Preventing teen pregnancy and academic failure: Experimental evaluation of a developmentally based approach. *Child Development*, 68, 729–742.
- Baldwin, S. A., Murray, D. M., & Shadish, W. R. (2005). Empirically supported treatments or Type I errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology*, 73, 924.
- *Battistich, V., Schaps, E., Watson, M., Solomon, D., & Lewis, C. (2000). Effects of the Child Development Project on students' drug use and other problem behaviors. *Journal of Primary Prevention*, 21, 75–99.
- *Battistich, V., Schaps, E., & Wilson, N. (2004). Effects of an elementary school intervention on students' "connectedness" to school and social adjustment during middle school. *Journal of Primary Prevention*, 24, 243–262. <https://doi.org/10.1023/B:JOPP.0000018048.38517.cd>
- *Beck, C., Sullivan, E., & Taylor, N. (1972). Stimulating transition to post conventional morality: the Pickering High School Study. *Interchange*, 3, 28–37.
- *Benninga, J. S., Tracz, S. M., Sparks Jr, R. K., Solomon, D., Battistich, V., Delucchi, K. L., Sandoval, R., & Stanley, B. (1991). Effects of two contrasting school task and incentive structures on children's social development. *The Elementary School Journal*, 92, 149–167.
- Berkowitz, M. W. (2002). The science of character education. In W. Damon (Ed.). *Bringing in a new era in character education* (pp. 43–63). Hoover Institution Press.
- Berkowitz, M. W. (2011). What works in values education. *International Journal of Educational Research*, 50(3), 153–158. <https://doi.org/10.1016/j.ijer.2011.07.003>
- Berkowitz, M. W., & Bier, M. C. (2004). Research-based character education. *The Annals of the American Academy of Political and Social Science*, 591, 72–85.
- Berkowitz, M. W., & Bier, M. C. (2005). *What works in character education: A research-driven guide for educators*. Character Education Partnership.
- Berkowitz, M. W., & Bier, M. C. (2007). What works in character education. *Journal of Research in Character Education*, 5, 29–48.
- Berkowitz, M. W., Battistich, V. A., & Bier, M. C. (2008). What works in character education: What is known and what needs to be known. In L. P. Nucci, T. Krettenauer & D. Narváez (Eds.), *Handbook of moral and character education* (p. 414–431). Routledge.
- Berkowitz, M. W., & Bier, M. C. (2014). Research-based fundamentals of the effective promotion of character development in schools. In L. P. Nucci, T. Krettenauer & D. Narváez (Eds.), *Handbook of moral and character education* (pp. 1–7). Routledge.
- Berkowitz, M. W., Bier, M. C., McCauley, B. (2017). Toward a science of character education. *Journal of Character Education* 13, 33–51.
- *Blatt, M. M., & Kohlberg, L. (1975) The effects of classroom moral discussion upon children's level of moral judgment. *Journal of Moral Education*, 4, 129–161.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, 1, 97–111.
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2013) *Comprehensive Meta Analysis Version 3.0*. Biostat.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. (2015). Regression in meta-analysis. *Comprehensive meta analysis manual*. <https://www.meta-analysis.com/downloads/MRManual.pdf>
- *Botvin, G. J., Dusenbury, L., Baker, E., James-Ortiz, S., Botvin, E. M., & Kerner, J. (1992). Smoking prevention among urban minority youth: Assessing effects on outcome

- and mediating variables. *Health Psychology*, 11, 290.
- *Botvin, G. J., Schinke, S. P., Epstein, J. A., Diaz, T., & Botvin, E. M. (1995). Effectiveness of culturally focused and generic skills training approaches to alcohol and drug abuse prevention among minority adolescents: Two-year follow-up results. *Psychology of Addictive Behaviors*, 9, 183–194.
- *Botvin, G. J., Baker, E., Dusenbury, L., Botvin, E. M., & Diaz, T. (1995). Long-term follow-up results of a randomized drug abuse prevention trial in a White middle-class population. *Journal of the American Medical Association*, 273, 1106–1112.
- *Botvin, G. J., Epstein, J. A., Baker, E., Diaz, T., & Ifill-Williams, M. (1997). School-based drug abuse prevention with inner-city minority youth. *Journal of Child & Adolescent Substance Abuse*, 6, 5–19.
- Brown, M., McGrath, R. E., Bier, M.C., Johnson, K., & Berkowitz, M. W. (in press). A comprehensive meta-analysis of character education programs. *The Journal of Moral Education*.
- *Caplan, M., Weissberg, R. P., Grober, J. S., Sivo, P. J., Grady, K., & Jacoby, C. (1992). Social competence promotion with inner-city and suburban young adolescents: Effects on social adjustment and alcohol use. *Journal of Consulting and Clinical Psychology*, 60, 56–63.
- *Catalano, R. F., Mazza, J. J., Harachi, T. W., Abbott, R. D., Haggerty, K. P., & Fleming, C. B. (2003). Raising healthy children through enhancing social development in elementary school: Results after 1.5 years. *Journal of School Psychology*, 41, 143–164.
- Coburn, K. M., & Vevea, J. L. (2019). *weightr: Estimating weight-function models for publication bias*. <https://CRAN.R-project.org/package=weightr>
- *Colby, A., Kohlberg, L., Fenton, E., Speicher-Dubin, B., & Lieberman, M. (1977). Secondary school moral discussion programmes led by social studies teachers. *Journal of Moral Education*, 6, 90–111. <https://doi.org/10.1080/0305724770060203>
- *Conduct Problems Prevention Research Group. (1999). Initial impact of the Fast Track Prevention Trial for Conduct Problems: II. Classroom effects. *Journal of Consulting and Clinical Psychology*, 67, 648–57.
- *Cognetta, P. (1977). Deliberate psychological education: A high school cross-age teaching model. *Counseling Psychologists*, 4, 22–24.
- *Cognetta, P. V., & Sprinthall, N. A. (1978). Students as teachers: Role taking as a means of promoting psychological and ethical development during adolescence. In N. Sprinthall & R. Mosher (Eds.), *Value development as the aim of education* (pp. 53–68). Character Research Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- *Cook, T. D., Murphy, R. F., & Hunt, H. D. (2000). Comer's school development program in Chicago: A theory-based evaluation. *American Educational Research Journal*, 37, 535–597. <https://doi.org/10.2307/1163533>
- Diggs, C. R., & Akos, P. (2016). The promise of character education in middle school: A meta-analysis. *Middle Grades Review*, 2, 1–19.
- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, 82, 405–432.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455–463.
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315, 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- *Elias, M. J., Gara, M. A., Ubriaco, M., Rothbaum, P. A., Clabby, J. F., & Schuyler, T. (1986). The impact of a preventative social problem solving intervention on children's coping with middle school stressors. *American Journal of Community Psychology*, 14, 259–275. <https://doi.org/10.1007/BF00911174>
- *Enright, R. D., Lapsley, D. K., Harris, D. J., & Shawver, D. J. (1983). Moral development interventions in early adolescence. *Theory Into Practice*, 22, 134–144.
- *Farrell, A. D., Meyer, A. L., & White, K. S. (2001). Evaluation of responding in peaceful and positive ways (RIPP): A school-based prevention program for reducing violence among urban adolescents. *Journal of Clinical Child Psychology*, 30, 451–463. https://doi.org/10.1207/S15374424JCCP3004_02

- *Farrell, A. D., Meyer, A. L., Sullivan, T. N., Kung, E. M. (2003) Evaluation of the responding in peaceful and positive ways (RIPP) seventh grade violence prevention curriculum. *Journal of Child and Family Studies*, 12, 101–120. <https://doi.org/10.1023/A:1021314327308>
- *Felner, R. D., Ginter, M., & Primavera, J. (1982). Primary prevention during school transitions: Social support and environmental structure. *American Journal of Community Psychology*, 10, 277–290.
- *Felner, R.D., Brand, S., Adan, A.M., Mulhall, P.F., Flowers, N., & Sartain, B. (1993) Restructuring the ecology of the school as an approach to prevention during school transitions: Longitudinal follow-ups and extensions of the School Transitional Environment Project (STEP). *Prevention in Human Services*, 10, 103–136.
- *Flannery, D. J., Vazsonyi, A. T., Liao, A. K., Guo, S., Powell, K. E., Atha, H., Vesterdal, W., & Embry, D. (2003). Initial behavior outcomes for the peacebuilders universal school-based violence prevention program. *Developmental Psychology*, 39, 292–309. <https://doi.org/10.1037/0012-1649.39.2.292>
- *Flay, B. R., Allred, C. G., & Ordway, N. (2001). Effects of the Positive Action program on achievement and discipline: Two matched-control comparisons. *Prevention Science*, 2, 71–89. <https://doi.org/10.1023/a:1011591613728>
- *Flay, B. R., & Allred, C. G. (2003) Long-term effects of the Positive Action program. *American Journal of Health Behavior*, 27, S6–S21.
- *Gottfredson, D. C. (1986). An empirical test of school-based environmental and individual interventions to reduce the risk of delinquent behavior. *Criminology*, 24, 705–731. <https://doi.org/10.1111/j.1745-9125.1986.tb01508.x>
- Glass, G. V., Smith, M. L., & McGaw, B. (1981). *Meta-analysis in social research*. SAGE.
- *Greenberg, M. T., Kusche, C. A., Cook, E. T., & Quamma, J. P. (1995) Promoting emotional competence in school-aged children: the effects of the PATHS curriculum. *Development and Psychopathology*, 7, 117–136.
- *Grossman, D. C., Neckerman, H. J., Koepsell, T. D., Liu, P. Y., Asher, K. N., Beland, K., Frey, K., & Rivara, F. P. (1997). Effectiveness of a violence prevention curriculum among children in elementary school: A randomized controlled trial. *Journal of the American Medical Association*, 277, 1605–1611.
- *Harrington, N. G., Giles, S. M., Hoyle, R. H., Fee-ney, G. J., & Yungbluth, S. C. (2001). Evaluation of the All Stars character education and problem behavior prevention program: Effects on mediator and outcome variables for middle school students. *Health Education & Behavior*, 28, 533–546.
- Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, 117, 167–178. <https://doi.org/10.1037/0033-2909.117.1.167>
- *Hawkins, J. D., Catalano, R. F., Morrison, D. M., O'Donnell, J., Abbott, R. D., & Day, L. E. (1992). The Seattle Social Development Project: Effects of the first four years on protective factors and problem behaviors. *Archives of Pediatrics and Adolescent Medicine*, 162, 1133–1141. <https://doi.org/10.1001/archpedi.162.12.1133>
- *Hawkins, J. D., Catalano, R. F., Kosterman, R., Abbott, R., & Hill, K. G. (1999). Preventing adolescent health-risk behaviors by strengthening protection during childhood. *Archives of Pediatrics and Adolescent Medicine*, 153, 226–234.
- *Hawkins, J. D., Guo, J. Hill, K. G., Battin-Pearson, S., & Abbott, R. D. (2001) Long-term effects of the Seattle Social Development Intervention on school bonding trajectories. *Applied Developmental Science*, 5, 225–236 http://10.1207/S1532480XADS0504_04
- *Hayden, B., & Pickar, D. (1981). The impact of moral discussions on children's level of moral reasoning. *Journal of Moral Education*, 10, 131–134.
- Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, 32, 151–179. <https://doi.org/10.3102/1076998606298040>
- Hedges, L. V. (2009). Effect sizes in nested design. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 337–355). Russell Sage Foundation.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60–87. <https://doi.org/10.3102/0162373707299706>
- Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane Handbook for Systematic Reviews Interventions Version 5.1.0*. The Cochrane Collaboration.
- Howard, R. W., Berkowitz, M. W., & Schaeffer, E. F. (2004). Politics of character education. *Edu-*

- cational policy, 18, 188–215. <http://10.1177/0895904803260031>
- Int'Hout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC medical research methodology*, 14, 14–25.
- Jeynes, W. H. (2019). A meta-analysis on the relationship between character education and student achievement and behavioral outcomes. *Education and Urban Society*, 51, 33–71.
- *Johnson, D. W., Johnson, R. T., Dudley, B., & Magnuson, D. (1995a). Training elementary school students to manage conflict. *The Journal of Social Psychology*, 135, 673–686. <http://10.1080/00224545.1995.9713971>
- *Johnson, D. W., Johnson, R., Dudley, B., Ward, M., & Magnuson, D. (1995b). The impact of peer mediation training on the management of school and home conflicts. *American Educational Research Journal*, 32, 829–844. <https://doi.org/10.3102/00028312032004829>
- *Keasey, C. B. (1973). Experimentally induced changes in moral opinions and reasoning. *Journal of Personality and Social Psychology*, 26, 30–38. <https://doi.org/10.1037/h0034210>
- *Knoff, H. M., & Batsche, G. M. (1995). Project ACHIEVE: Analyzing a school reform process for at-risk and underachieving students. *School Psychology Review*, 24, 579–603.
- *Krug, E. G., Brener, N. D., Dahlberg, L. L., Ryan, G. W., & Powell, K. E. (1997). The impact of an elementary school-based violence prevention program on visits to the school nurse. *American Journal of Preventive Medicine*, 13, 459–463.
- Lapsley, D. K., & Narváez, D. (2006). Character education. In A. Renninger & I. Siegel (Eds.) and W. Damon & R. Lerner (Editors-in-Chief.), *Handbook of child psychology: Vol. 4. Child psychology in practice* (6th ed., pp. 248–296). Wiley.
- Leming, J. S. (1993). In search of effective character education. *Educational Leadership*, 51, 63–71.
- *Leming, J. S. (2000). Tell me a story: An evaluation of a literature-based character education programme. *Journal of Moral Education*, 29(4), 413–427. <https://doi.org/10.1080/713679388>
- Lickona, T. (1991). *Educating for character*. Bantam Books.
- Lickona, T., & Davidson, M. (2005). *Smart & good high schools: Integrating excellence and ethics for success in school, work, and beyond*. Center for the 4th and 5th Rs/ Character Education Partnership. <https://www2.cortland.edu/centers/character/high-schools/SnGReport.pdf>
- Light, R. J., & Pillemer, D. B. (1984) *Summing up: The science of reviewing research*. Harvard University Press.
- Linden, A. H., & Hönkopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, 16(2), 358–376. <https://doi.org/10.1177/1745691620964193>
- Linkins, M., Niemiec, R. M., Gillham, J., & Mayer-son, D. (2014). Through the lens of strength: A framework for educating the heart. *The Journal of Positive Psychology*, 10, 64–68.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48(12), 1181–1209. <https://doi.org/10.1037/0003-066X.48.12.1181>
- *Lockwood, A. L. (1978). The effects of values clarification and moral development curricula on school-age subjects: A critical review of recent research. *Review of Educational Research*, 48, 325–364.
- McClellan, B. E. (1999). *Moral education in America: Schools and the shaping of character from colonial times to the present*. Teachers College Press.
- Martin, M. A., & Roberts, S. (2010). Jackknife-after-bootstrap regression influence diagnostics. *Journal of Nonparametric Statistics*, 22, 257–269. <https://doi.org/10.1080/10485250903287906>
- McGrath, R. E. (2017). The essentials and complexities of character: Reflections on Nucci's multifaceted model. *Journal of Character Education*, 13, 27–32.
- McGrath, R. E. (2018). What is character education? Development of a prototype. *Journal of Character Education*, 14, 23–35.
- McGrath, R. E., Han, H., Brown, M., & Meindl, P. (2021). What does character education mean to character education experts? A prototype analysis of expert opinions. *Journal of Moral Education*, 50, 1–19.
- *McNeal Jr, R. B., Hansen, W. B., Harrington, N. G., & Giles, S. M. (2004). How All Stars works:

- An examination of program effects on mediating variables. *Health Education and Behavior*, 31, 165–178. <https://doi.org/10.1177/1090198103259852>
- *Metropolitan Area Child Study Research Group. (2002). A cognitive-ecological approach to preventing aggression in urban settings: Initial outcomes for high-risk children. *Journal of Consulting and Clinical Psychology*, 70, 179–194. <https://doi.org/10.1037/0022-006X.70.1.179>
- Mulkey, Y. J. (1997). The history of character education. *Journal of Physical Education, Recreation & Dance*, 68, 35–37.
- Murphy, M. (2002). *Character education in America's blue ribbon schools: Best practices for meeting the challenge*. R&L Education.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford University Press.
- *O'Donnell, J., Hawkins, J. D., Catalano, R. F., Abbott, R. D., & Day, L. E. (1995). Preventing school failure, drug use, and delinquency among low-income children: Long-term intervention in elementary schools. *American Journal of Orthopsychiatry*, 65, 87–100.
- *Orpinas, P., Parcel, G. S., McAlister, A., & Frankowski, R. (1995). Violence prevention in middle schools: A pilot evaluation. *Journal of Adolescent Health*, 17, 360–371. [https://doi.org/10.1016/1054-139X\(95\)00194-W](https://doi.org/10.1016/1054-139X(95)00194-W)
- *Orpinas, P., Kelder, S., Frankowski, R., Murray, N., Zhang, Q., & McAlister, A. (2000). Outcome evaluation of a multi-component violence-prevention program for middle schools: The Students for Peace project. *Health Education Research*, 15, 45–58. <https://doi.org/10.1093/her/15.1.45>
- Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, 26, 4544–4562.
- *Reid, J. B., Eddy, J. M., Fetrow, R. A., & Stoolmiller, M. (1999). Description and immediate impacts of a preventive intervention for conduct problems. *American Journal of Community Psychology*, 27, 483–518. <https://doi.org/10.1023/A:1022181111368>
- *Reyes, O., & Jason, L.A. (1991) An evaluation of a high school dropout prevention program. *Journal of Community Psychology*, 19, 221–230
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, 45(6), 775–777. <https://doi.org/10.1037/0003-066X.45.6.775>
- Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, 33, 62–87. <https://doi.org/10.3102/1076998607302714>
- *Schlaefli, A., Rest, J. R., & Thoma, S. J. (1985). Does moral education improve moral judgment? A meta-analysis of intervention studies using the Defining Issues Test. *Review of Educational Research*, 55, 319–352.
- *Schultz, L. H., Barr, D. J., & Selman, R. L. (2001). The value of a developmental approach to evaluating character development programmes: An outcome study of facing history and ourselves. *Journal of Moral Education*, 30, 3–27. <https://doi.org/10.1080/0305724012003378>
- *Shapiro, J. P., Burgoon, J. D., Welker, C. J., & Clough, J. B. (2002). Evaluation of the peacemakers program: School-based violence prevention for students in grades four through eight. *Psychology in the Schools*, 39, 87–100. <https://doi.org/10.1002/pits.10040>
- *Shure, M. B., & Spivack, G. (1979). Interpersonal cognitive problem solving and primary prevention: Programming for preschool and kindergarten children. *Journal of Clinical Child & Adolescent Psychology*, 8, 89–94.
- *Shure, M. B., & Spivack, G. (1980). Interpersonal problem solving as a mediator of behavioral adjustment in preschool and kindergarten children. *Journal of Applied Developmental Psychology*, 42, 29–44.
- *Shure, M. B., & Spivack, G. (1982). Interpersonal problem-solving in young children: A cognitive approach to prevention. *American Journal of Community Psychology*, 10, 341–356.
- *Shope, J. T., Copeland, L. A., Kamp, M. E., & Lang, S. W. (1998). Twelfth grade follow-up of the effectiveness of a middle school-based substance abuse prevention program. *Journal of Drug Education*, 28, 185–197.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychologi-*

- cal Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Social and Character Development Research Consortium. (2010). *Efficacy of schoolwide programs to promote social and character development and reduce problem behavior in elementary school children* (NCER 2011-2001). National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.
- Sojourner, R. J. (2012). *The rebirth and retooling of character education in America*. <https://www.character.org/wp-content/uploads/Character-Education.pdf>
- *Solomon, D., Battistich, V., Watson, M., Schaps, E., & Lewis, C. (2000). A six-district study of educational change: Direct and mediated effects of the Child Development Project. *Social Psychology of Education*, 4, 3–51. <https://doi.org/10.1023/A:1009609606692>
- *Spoth, R., Redmond, C., Trudeau, L., & Shin, C. (2002). Longitudinal substance initiation outcomes for a universal preventive intervention combining family and school programs. *Psychology of Addictive Behaviors*, 16, 129–134.
- *Taub, J. (2002). Evaluation of the Second Step violence prevention program at a rural elementary school. *School Psychology Review*, 31, 186–200.
- *Taylor, A. S., LoSciuto, L., Fox, M., Hilbert, S. M., & Sonkowsky, M. (1999). The mentoring factor: Evaluation of the across ages' intergenerational approach to drug abuse prevention. *Child & Youth Services*, 20, 77–99.
- *Taylor, C. A., Liang, B., Tracy, A. J., Williams, L. M., & Seigle, P. (2002). Gender differences in middle school adjustment, physical fighting, and social skills: Evaluation of a social competency program. *Journal of Primary Prevention*, 23, 259–272. <https://doi.org/10.1002/jcop.20246>
- Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development*, 88, 1156–1171.
- Taylor, S. J., & Tweedie, R. L. (1998). Practical estimates of the effect of publication bias in meta-analysis. *Australian Epidemiologist*, 5, 14–17.
- Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, 22, 2113–2126.
- *Turiel, E. (1966). An experimental test of the sequentiality of developmental stages in the child's moral judgments. *Journal of Personality and Social Psychology*, 3, 611–618.
- *Twemlow, S. W., Fonagy, P., Sacco, F. C., Gies, M. L., Evans, R., & Ewbank, R. (2001). Creating a peaceful school learning environment: A controlled study of an elementary school intervention to reduce violence. *American Journal of Psychiatry*, 158, 808–810.
- *Van Schoiack-Edstrom, L., Frey, K. S., & Beland, K. (2002). Changing adolescents' attitudes about relational and physical aggression: An early evaluation of a school-based intervention. *School Psychology Review*, 31, 201–216.
- Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, 60, 419–435.
- Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, 10, 428–443.
- *Walker, L. J., & Richards, B. S. (1979). Stimulating transitions in moral reasoning as a function of stage of cognitive development. *Developmental Psychology*, 15, 95–103. <https://doi.org/10.1037/0012-1649.15.2.95>
- *Walker, L. J. (1980). Cognitive and perspective-taking prerequisites for moral development. *Child Development*, 51, 131–139. <https://doi.org/10.2307/1129599>
- What Works Clearinghouse. (2014). Procedures and standards handbook (Version 2.0). https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v2_1_standards_handbook.pdf
- Williams, M. M. (2000). Models of character education: Perspectives and developmental issues. *The Journal of Humanistic Counseling*, 39, 32–40. <https://doi.org/10.1002/j.2164-490X.2000.tb00091.x>