Routledge
Taylor & Francis Group

Journal of Moral Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjme20

A comprehensive meta-analysis of character education programs

Mitch Brown, Robert E. McGrath, Melinda C. Bier, Keith Johnson & Marvin W. Berkowitz

To cite this article: Mitch Brown, Robert E. McGrath, Melinda C. Bier, Keith Johnson & Marvin W. Berkowitz (2022): A comprehensive meta-analysis of character education programs, Journal of Moral Education, DOI: 10.1080/03057240.2022.2060196

To link to this article: https://doi.org/10.1080/03057240.2022.2060196

View supplementary material

Published online: 13 May 2022.

Submit your article to this journal

Article views: 293

View related articles

View Crossmark data

JOURNAL OF MORAL EDUCATION
https://doi.org/10.1080/03057240.2022.2060196

A comprehensive meta-analysis of character education programs

Mitch Brown[a], Robert E. McGrath[b], Melinda C. Bier[c], Keith Johnson[b] and Marvin W. Berkowitz[c]

[a]Department of Psychological Science, University of Arkansas, Fayetteville, Arkansas, United States; [b]School of Psychology and Counseling, Fairleigh Dickinson University, Teaneck, New Jersey, United States; [c]Center for Character and Citizenship, University of Missouri-St. Louis, St. Louis, Missouri, United States

ABSTRACT

This study represents a large meta-analytic investigation of character education studies. We identified 214 studies up until 2017 (N_{Total} = 307,512) with at least one computable effect size comparing a character education program to a control condition. Initial analyses indicated a small, significant average positive effect for character education (g = 0.24, 95% CI [0.19, 0.28]). Single-session interventions and mentoring programs were associated with larger effects. However, all treatment lengths were associated with significant treatment effects, and the number of single-session interventions was small. Three programs—Cognitive Problem-Solving, Kohlberg's Moral Dilemma Discussion and Strong Kids—were associated with above-average effects in multiple assessments. Evidence emerged suggesting selection bias in published literature. Correcting this bias indicated lower estimates of mean effects while remaining positive and significant. We consider heterogeneity in reporting standards and discuss how to address biases.

KEYWORDS

Meta-analysis; character education; social-emotional learning; selection bias

The implementation of interventions regarded as 'character education' has become increasingly widespread over the past few decades. A substantial corpus of research has emerged attempting to identify which interventions are most capable of optimizing youths' positive outcomes (e.g., Berkowitz & Bier, 2007; Leming, 1993; Lickona & Davidson, 2005). Character education programs have additionally attracted substantial monetary support from a variety of private and public sources (Howard et al., 2004; Murphy, 2002; Sojourner, 2012). Such investment necessitates evaluating whether these implementations have been effective, and the conditions under which positive effects emerge.

A small number of reviews of the character education literature have been completed. The largest attempt at a comprehensive review of this literature remains What Works in Character Education (WWCE) (Berkowitz & Bier, 2005, 2007). WWCE involved summarizing results from 33 programs across 69 publications. Though extremely influential in the ensuing years, WWCE also had its limitations. In particular, it focused on which aspects of character education were most consistently associated with significant positive

CONTACT Mitch Brown mb103@uark.edu Department of Psychological Science, University of Arkansas, 72701 Fayetteville, Arkansas, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/03057240.2022.2060196

outcomes (i.e., what works?). The subsequent emergence of meta-analysis as the dominant approach to research synthesis prioritizes a different issue, i.e., the overall effectiveness of an intervention or class of interventions (i.e., does it work?).

A reanalysis of publications used in the WWCE project using meta-analytic methods found modest effect sizes, with a mean Hedge's *g* of 0.33 (Johnson et al., in press). However, the body of literature underlying the WWCE project is now 15 years old and does not address results published after WWCE. Two meta-analyses including more recent studies are also available. One consisted of 11 studies published since WWCE that exclusively considered middle school/early adolescence; they found a mean *d* value of 0.15 for academic variables, −0.20 for behavior (suggesting greater deterioration in students receiving character education), and 0.26 for perceptual variables (Diggs & Akos, 2016). This analysis did not provide an overall mean effect but weighting these values by the number of outcome variables in each category generates a mean *d* of .04. Another review indicated mean *d* values between 0.24 and 0.28 across various subsets of 52 articles with students from pre-kindergarten to college freshman levels (Jeynes, 2017).

The present research builds on the work of Johnson et al. (in press), presenting a meta-analytic review of the character education literature through 2017. This year was the start of data collection of the publication pool, implicating 2017 an effective cut-off date. This study incorporates both those studies identified from the WWCE literature search as accessible to meta-analytic techniques in this meta-analysis, as well as literature that has emerged after WWCE was completed. We further considered possible moderators of effect sizes, including the content of the interventions and their outcomes, selection bias, and study quality. These analyses serve the original WWCE goals of identifying which components of character education may be most effective at eliciting positive outcomes in a more extensive and contemporary context.

Defining character and character education

Character refers to the extent a person is motivated to and habitually thinks and behaves in prosocial ways. Behaviors and attitudes thought indicative of character typically contribute to the functioning of a social community. This allows for a very broad array of outcomes (e.g., nonviolence and abstinence from drugs) to be considered for inclusion, and the WWCE project did not restrict the domain of outcomes so long as they were considered socially valued at one end of the continuum or the other and were potentially indicative of character.

This inclusiveness has led to defining an intervention as character education being described as a 'semantic minefield' for researchers (Berkowitz & Bier, 2014, p. 249). A definition of character education consensually satisfying researchers has proven difficult, as there are differing opinions on what should constitute character education and myriad features researchers find to be central to its definition (Berkowitz, 2021; Lapsley & Narváez, 2006; McGrath et al., in press; Williams, 2000). Character has been conceptualized as a complex, dynamic, multidimensional psychological construct encompassing moral, intellectual, self-regulatory, and civic functioning (Shields, 2011), with particular emphasis being placed on the moral domain (Berkowitz et al., 2017; Lickona & Davidson, 2005). There is a wide array of attempts to define character. Character.org defines it as the composite of cognitive, emotional/motivational, and behavioral characteristics in service of core ethical values. From a virtue ethics perspective, character may be defined as the composite of virtues (Wright et al., 2020). More limited definitions of character consider performance character elements that promote school success (e.g., grit and curiosity; Tough, 2012). Recognizing the variability of definitions, we choose to define it as the whole set of psychological characteristics motivating and enabling one to operate as an effective member of society, to flourish intellectually, to strive for excellence, and to serve as a moral agent.

Substantial variability also exists in the strategies used to nurture the development of character, further muddying the definition (McGrath, 2018). Although some character education programs are pedagogical, others focus on peer relationships and community involvement, while still others emphasize mentorship. The WWCE project included programs employing a variety of interventions, including teaching core values, prosocial behaviors, socio-emotional reasoning, and conflict resolution (Berkowitz & Bier, 2007). As WWCE and the present study aimed to be inclusive of programs intended to promote a broad conception of character, character education was defined as any program intended to foster positive student development (see Berkowitz & Bier, 2005, 2007; Berkowitz et al., 2017). According to this approach, youth development programs can be considered character education if they target some aspect of social investment, including moral values, socio-moral reasoning, knowledge of ethical issues, moral emotional competencies, or moral identity; behavioral competencies such as conflict resolution skills; and/or characteristics that support prosocial behavior. One of the few restrictions on range implemented by WWCE was limiting settings to kindergarten through high school, so this restriction was carried forward into the present study.

Current study

Prior meta-analyses specific to character education have had relatively limited focus. The current meta-analysis was intended as the most extensive to date on the topic, including examination of the relationship between interventions and across a range of options. In particular, the present review includes about four times as many studies as the largest existing meta-analyses, and slightly more than the widely cited and influential meta-analysis of social emotional learning studies (Durlak et al., 2011). Building on the WWCE-specific meta-analysis reported by Johnson et al. (in press), the current study offers an extension of that piece of research.

Method

Search strategy

Searching occurred in three separate stages referred to as the WWCE and the post-WWCE stages (see Figure 1 for the PRISMA diagram for both stages). This project started in 2017, which served as the basis for our last year for consideration in the analysis. The WWCE stage resulted in 64 studies through 2004 that were part of a preliminary analysis intended to provide a meta-analytic complement to WWCE (Johnson et al., in press). The initial search for the post-WWCE stage was conducted

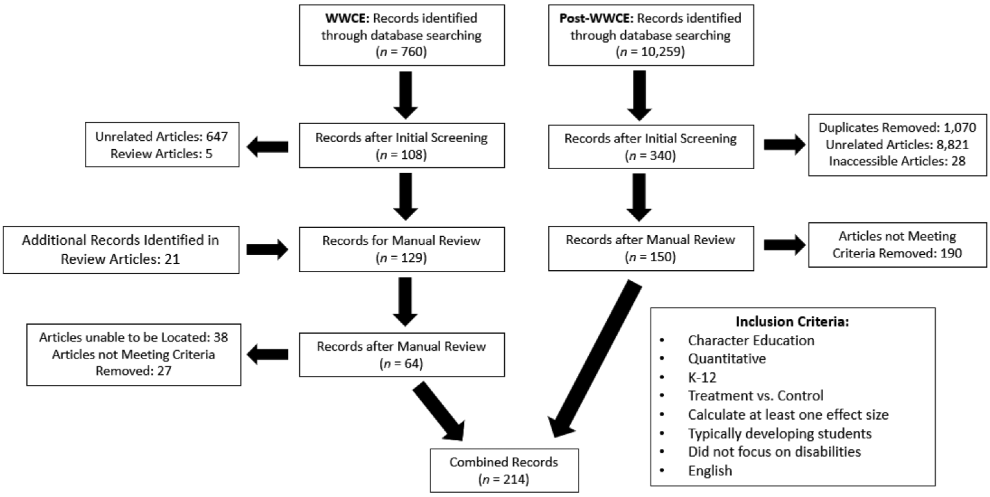
WWCE: Records identified through database searching (n = 760)
Post-WWCE: Records identified through database searching (n = 10,259)
Unrelated Articles: 647
Review Articles: 5
Records after Initial Screening (n = 108)
Records after Initial Screening (n = 340)
Duplicates Removed: 1,070
Unrelated Articles: 8,821
Inaccessible Articles: 28
Additional Records Identified in Review Articles: 21
Records for Manual Review (n = 129)
Records after Manual Review (n = 150)
Articles not Meeting Criteria Removed: 190
Articles unable to be Located: 38
Articles not Meeting Criteria Removed: 27
Records after Manual Review (n = 64)
Combined Records (n = 214)
Inclusion Criteria:
Character Education
Quantitative
K-12
Treatment vs. Control
Calculate at least one effect size
Typically developing students
Did not focus on disabilities
English

Figure 1. PRISMA chart describing selection of articles from WWCE and post-WWCE collections for inclusion.

in two steps. The authors of the WWCE subsequently reviewed the literature for the years 1999–2014; the period overlapped with the original WWCE search to identify any references missed in the WWCE stage.

Results were further augmented by a search covering the period 2014–2017, the year in which the project started. Terms used for the two post-WWCE searches are in Appendix A. Because of differences in access to databases, the first two searches focused on the Ebscohost databases (Education Full Text, Educational Administration Abstracts, ERIC, PsycINFO, and Social Work Abstracts) and PubMed. The Ebscohost set of databases was not available for the 2014–2017 search, so that search included ERIC, Education Source, PubMed, Social Work Abstracts, and PsycINFO. This search resulted in studies that were peer-reviewed or served as theses and dissertations to be indexed by these databases.

As described by Johnson et al. (in press), the WWCE stage began with identification of 64 studies out of 113 sources that met the following conditions:

(1) Studies must have evaluated a program that met the broad definition of character education provided above.
(2) It was implemented in a kindergarten to grade 12 setting.
(3) It focused on a general or ‘at-risk’ student population; studies focusing on populations with identified special needs (e.g., children with behavioral disorders) were excluded.
(4) The study involved a control condition.
(5) The research was quantitative and allowed computation of at least one standardized mean difference between groups.

Results from the two post-WWCE searches were combined, resulting in an initial pool of 10,259 new citations. As outlined in Figure 1, a final list of 340 references was reviewed by two graduate students for compliance with the inclusionary criteria

Table 1. Indicators for research quality.

Quality Indicator	Rating of Threat (with number of studies in each category)
Randomization	**High**: Non-random assignment (82) **Unclear**: Method of assignment not adequately described (1) **Low**: Random assignment by student, class, or school (131)
Providers Blind	**High**: Personnel know which treatment participants would receive (180) **Unclear**: Issue not addressed (32) **Low**: Personnel did not know treatment before study enrollment (2)
Assessors Blind	**High**: Personnel aware of treatment group (125) **Unclear**: Issue not addressed (56) **Low**: Personnel unaware of treatment group (33)
Attrition	**High**: Different rates of incomplete data across groups ignored in analysis (20) **Unclear**: Incomplete data not discussed (89) **Low**: Effect of incomplete data addressed in some way (105)
Selective Reporting	**High**: Clear evidence some results (groups, measures) not reported (8) **Unclear**: Issue not addressed (124) **Low**: Authors indicate all groups and measures are included (82)
Possible Contamination	**High**: Multiple treatments administered in same school or class (34) **Unclear**: Different treatments administered in different classes of same school (57) **Low**: Researchers apply different treatments in different schools (123)
Failure to Check for Treatment Fidelity	**High**: Researchers provide no evidence of evaluating fidelity in treatment (31) **Unclear**: Fidelity evaluation may not have been adequate (80) **Low**: Researchers described methods of addressing fidelity (103)
Failure to Check for Reliability	**High**: Researchers provide no evidence for considering reliability of measures (17) **Unclear**: Inconsistent evaluation of measures (38) **Low**: Some effort made to ensure reliability for reported measures (159)

Note. Values in parentheses indicate the number of studies receiving that rating.

described above. This resulted in a post-WWCE pool of 150 articles. Combined with the 64 studies from the initial meta-analyses, the result was 214 studies for the present meta-analysis.

Data extraction

Data were extracted for each study by two graduate students and discrepancies subsequently resolved by one of the authors. In addition to outcome statistics, potential moderators were also extracted, including the length of the program, program variables, program targets, program components, and types of outcomes. The moderators recorded are listed in Appendix B. We also rated each study on eight indicators of study quality based on the Cochrane Handbook for Systematic Reviews of Interventions standards (Higgins & Green, 2011). Raters classified each study as demonstrating a high, unclear, or low risk of poor quality (see Table 1).

General analytic strategy

We included all comparisons between groups from the 214 studies for which computation or estimation of a standardized mean difference was possible. For the entire set of 214 studies, 2,472 analyses were extracted, for an overall mean of 11.60 analyses per study.

We generated most meta-analytic results 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. For the majority of analyses (79%), the original authors either provided an estimate of *d* themselves or means and standard deviations from which we computed *d*. Another 10% were based on rates or odds ratios for dichotomous outcomes rather than dimensional outcomes. These were converted to an estimated standardized mean difference using the log odds ratio (Hasselblad & Hedges, 1995). The other 11% of effect sizes were estimated from an inferential statistic (*t* or *F*) or another effect size (e.g., a regression coefficient).

Statistical decisions were consistent with those described by Johnson et al. (in press). 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 and Hedberg (2007) Eq. 5. Except when computed from a *t* or *F* value, which had already been adjusted, standardized mean differences and standard errors were then corrected for clustering bias (Eq. 18.19 and 18.20; Hedges, 2009). When correcting for participant clustering (Hedges, 2007), the intraclass correlation coefficient was always 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). *d* values that exceeded the top and bottom 2.5% were winsorized and the standard error was recomputed. This resulted in adjustment of 124 *d* values. Random effects meta-analysis of final *d* and standard error estimates was then used to generate a mean Hedge's *g* for each study, and the set of studies as a whole. This was the appropriate meta-analytic model given the combination of student demographics, interventions, and outcome variables, suggesting the studies should be considered a mix of treatment populations.

Results

Sample characteristics and descriptive statistics

The 214 studies included a total of 307,512 participants. Among studies reporting ethnic composition of the sample, 40.2% of studies had predominantly White participants, 16.8% of studies reported a majority Black sample, and 7.9% of studies included predominantly Hispanic participants. Samples on average were 49.8% girls (range = [0, 100]), and 53% of participants were eligible for free/reduced-cost lunch. Most studies reporting a type of community took place in an urban environment (26% of studies), while 10% occurred in a suburban environment and 3% were rural; however, most were unclear or used multiple sites. The average age of students was 11.12 years across studies providing a mean (range = [5, 18]).

The average number of schools from which data were collected for each study was 16.78 (range = [1, 153]). The most common venue for distributing the intervention was the classroom (62%), while 20% of programs were schoolwide. Only seven were offered as after-school programs, and these were offered to students on an individual basis. Random assignment of clusters occurred in 61% of studies. Only 30% of studies in which administration was clustered used mixed models for statistical analysis of the results.

A positive point-biserial correlation emerged between year of publication and use of statistics correcting for clustering (r = .29). This suggests increased use over time; no study corrected for student clusters until 1997. In the subsequent 21 years, there were only seven years when the majority of articles included correction for non-independence in the computation of significance tests. In cases where significance tests were not corrected, correction of t values for clustering were reduced by an average of 44%. This represents an issue of continuing misrepresentation of results in this literature.

There were 152 distinct youth intervention programs across the 214 studies. Appendix B lists the program attributes, targets, methods, and outcome variables that we coded, and the percent of studies or analyses (for outcome variables) fitting each category. Note that variables listed were not exclusive of each other.

Primary analyses

For an initial estimate of effect size, we generated a single mean effect for each study by aggregating across subgroups, experimental interventions, outcome measures, and outcome measurement points. These were then averaged after weighting each study effect by its inverse variance estimate. The result was a small but significant effect, $g = 0.24$, $SE = 0.02$, $p < .001$, 95% CI = [0.19, 0.28]. Knapp-Hartung-Sidik-Jonkan correction for heterogeneity (IntHout et al., 2014) resulted in no change in these values. This finding replicates the main conclusion by Johnson et al. (in press) using a subset of the present studies: character education as a broad umbrella concept has a small, positive influence on average on character outcomes in the K-12 setting. However, the prediction interval was [−0.22, 0.69]. This interval reflects the best estimate of the range of effects across the populations included in the random effects meta-analysis. The inclusion of negative values in this interval suggests that in some settings examined there is risk of overall deleterious outcomes for some interventions.

Heterogeneity was apparent across studies, $Q(213) = 891.36$, $p < .001$. The I^2 suggested 24% of variability was due to sampling error, 76% due to variability in treatment effects across populations. The standard deviation of population effects was estimated at $\tau = 0.23$. The addition of more articles seems to have resulted in less variability than was reported by Johnson et al. (in press), and less variability than appears typical in meta-analyses (Linden & Hönekopp, 2021).

Moderation analyses

We conducted moderation analyses to determine the extent to which various factors influenced these reported effects (see Appendix B). A significant outcome emerged for program length, $F(3, 203) = 17.69$, $p < .001$, $\eta_p^2 = .194$. Pairwise contrasts across the four levels indicated single-session interventions were associated with a significantly larger mean effect than programs of any greater duration, and programs of no more than one week or one month duration were associated with a significantly larger mean effect than longer programs. There were also significant effects suggesting programs that incorporated elements outside school or focused on moral sentiments were associated with significantly lower mean effects. However, in both cases simultaneous regression with the length variable eliminated these effects, suggesting that they tended to be associated with longer programs. The omnibus test for the four mentoring options was not significant, but pairwise contrasts indicated programs incorporating mentors had better outcomes than programs offering no mentoring. This effect remained significant even when controlling for program length. No other pairwise contrasts for mentoring were significant.

It should be noted that only seven studies examined single-session programs, necessitating cautious interpretation of the findings for unusually brief programs. However, the consistent finding that programs of more than one month were associated with smaller effects than programs of briefer durations suggests concerns about the relative effectiveness of long-term programs. Table 2 provides effect sizes. It is also noteworthy that all four means were significantly greater than zero, indicating that programs of any length on average had positive impacts.

Particularly given the presence of negative values in the prediction interval, it was considered important to evaluate program as a moderator. However, 130 of 152 programs were the focus of only one study meeting our criteria for inclusion. Only 14 programs were the focus of three or more investigations, which could be considered a bare minimum for considering a finding to have been reliably demonstrated:

- Child Development Project,
- Drug Abuse Prevention Intervention,
- Good Behavior Game,
- Interpersonal Cognitive Problem Solving,
- KiVa,
- Kohlberg's Moral Development,
- Promoting Alternative Thinking Strategies,
- Positive Action,
- School-Wide Positive Behavioral Interventions and Supports,
- Seattle Social Development Project,
- Second Step,
- Strong Kids,
- Teen Outreach,
- and Transition Project.

This lack of replication limited our ability to draw conclusions about specific programs. Given the practical and financial difficulty in conducting each individual study multiple times, this limitation is unsurprising. In all 14 cases, the mean effect was positive. Significance tests were under-powered, so the 68 studies in which one of these

Table 2. Mean effect sizes for significant moderators.

Moderator	*k*	Mean *g*	*SE*
Program Length			
1 Session	7	0.86	0.41
< 1 Week	14	0.35	0.13
< 1 Month	118	0.25	0.03
> 1 month	68	0.14	0.02
Mentoring			
None	129	0.21	0.02
Formal Mentor	25	0.39	0.11

Note. *k* = # of studies. All means are significantly different than zero, *p* < .05. A full list of moderators is available in Appendix B.

14 programs was evaluated were combined. The mean effect size was $g = 0.29$, $SE = 0.04$, $p < .001$, 95% CI = [0.20, 0.37], slightly higher than the mean effect overall. In reviewing these programs, however, we noted an important cause for caution. For 5 of the 14 programs, all studies seem to have been generated by a single research group as indicated by overlap in the list of authors. This allows for the possibility that effect sizes are larger than would be seen in other settings due to allegiance effects (Hollon, 1999). Of the remaining nine where at least one study had no overlapping authors with other evaluations, six (Child Development Project, Good Behavior Game, Promoting Alternative Thinking Strategies, School-Wide Positive Behavioral Interventions and Supports, Second Step, and Transition Project) were associated with mean effects <0.20. In contrast, Interpersonal Cognitive Problem-Solving was associated with a mean g of 1.10, Kohlberg's moral development with a mean of 0.50, and Strong Kids with a value of 0.33, suggesting the strongest evidence available among the tested programs.

We also examined the mean effect at posttest (upon program completion) versus measurements at follow-up periods. The mean effect at posttest was 0.26. This dropped to 0.17 for follow-ups, suggesting immediate effects on average were larger than delayed effects.

Finally, we classified studies as high- and low-quality based on our assessment of study quality through the eight indicators of bias (see Table 1), arbitrarily considering a study to be of higher quality when the number of ratings of low risk of some bias in the design exceeded the number of ratings of high risk by at least four ($k = 33$). For those studies deemed as higher quality, the overall positive effect was smaller but still significant, $g = 0.17$, $SE = 0.04$, $p < .001$, 95% CI = [0.08, 0.25].

Selection bias

We also considered the degree to which selection bias influenced our reported findings. Both Kendall's tau (.23) and Egger's regression intercept (0.79) were significant, indicating evidence of selection bias. The asymmetrical nature of the funnel plot also offered evidence of selection bias (see Figure 2). There were a number of studies with mean effects >1.0 even with winsorizing and none with a mean effect < −0.25. Correcting for this effect using the trim-and-fill method (Duval & Tweedie, 2000) reduced the estimated mean effect to 0.11 with a 95% CI = [0.06, 0.15]. However, concerns have been raised about the accuracy of the trim-and-fill method in the presence of heterogeneity across populations (Schwarzer et al., 2010).

Vevea and Hedges (1995) described a general linear model that allows for the computation of a likelihood ratio test evaluating the null hypothesis of no bias based on the distribution of mean p values across studies, as well as estimation of the mean effect size in the absence of selection bias. The final 2,472 d value estimates and corresponding sampling variances were used for this analysis. Analyses were conducted using the *weightr* package in R (Coburn & Vevea, 2019). The likelihood ratio test was significant, $\chi^2(df = 13) = 1787.55$, $p < .001$, again suggesting significant evidence of bias.

This approach also allows for presetting the distribution of the p intervals, and Vevea and Woods (2005) suggested reasonable estimates for setting these distributions. Based on the assumption of moderate selection bias in favor of the low tail of the p distribution (significant outcomes), the estimated mean effect was reduced to 0.11, the same value

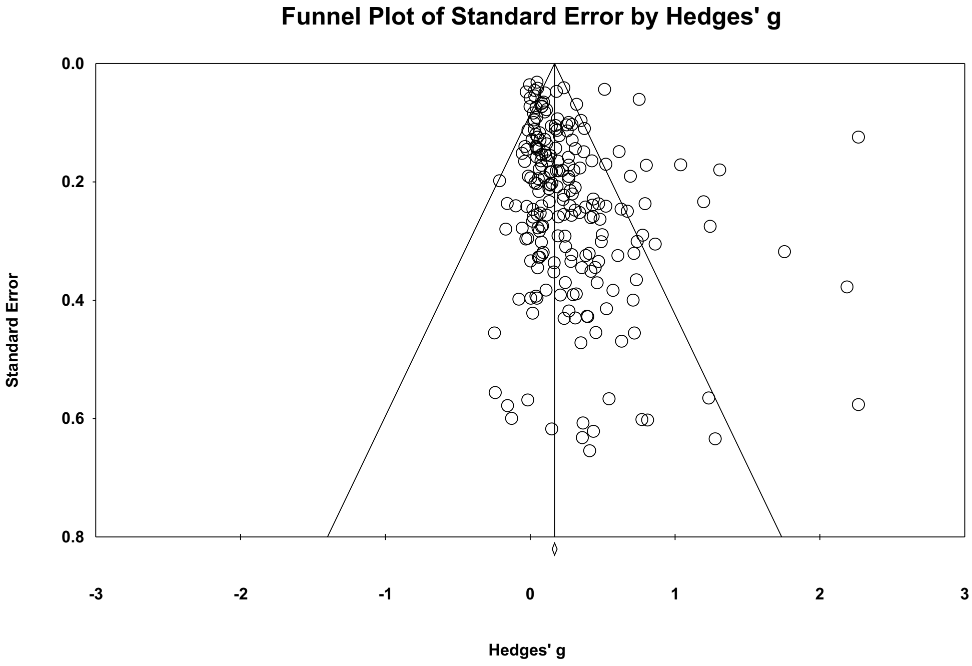

Figure 2. Funnel plot of studies included in the meta-analysis.

suggested by the trim and fill method. These analyses were repeated using the mean effect from studies, and with Hedges' *g* values, with similar results. Tests of selection bias were consistently positive and suggested the potential for substantial bias in the results.

Discussion

The goal of this study was to provide a substantial meta-analytic investigation of the effectiveness of character education across an inclusive, and broadly defined, array of interventions. These findings specifically build upon the original WWCE project (Berkowitz & Bier, 2007), which only considered interventions deemed effective, by considering the presence of null and negative findings to afford researchers a more accurate understanding of character education as a field. In analyzing these 214 studies, interventions demonstrated an overall positive impact, with a mean effect size of $g = 0.24$. Though such an effect is considered small by widely accepted statistical conventions (e.g., Cohen, 1988), in contexts of large populations such effects can be quite important. For example, previous research indicated that the effect of aspirin on prevention of heart attacks is markedly less than this, but across the entire at-risk population represents an important intervention (Rosenthal, 1990). It may also speak to the overall difficulty in implementing character education programs effectively (see Prentice & Miller, 1992). Most importantly, given the broad range of quality and types of character education, and the variability in evaluation methods and particular outcome assessments, it would be expected that there is also a broad range in actual effectiveness and ability to measure effects, thus generating a pool of outcome research studies both expected to and not expected to demonstrate significant implementation impacts. Finding an overall significant effect becomes even more meaningful, despite this variability. At the same time, the focus on published studies creates the possibility that even this small effect exaggerates the true mean effect.

This overall value is comparable in magnitude with previously reported effects for character education and social-emotional learning programs. It is better than the mean of .04 Diggs and Akos (2016) reported for 11 studies and is consistent with the range of point estimates ([0.24, 0.28]) reported by Jeynes (2017). The latter is more consistent with the present study in terms of scope (including elementary and high-school students) and number of studies, so probably represent more defensible representations of the character education literature in general. Meta-analyses investigating school-based social-emotional learning programs have suggested that effects for those programs are larger when the outcome measure is a measure of social-emotional skills. For all other outcomes (e.g., attitudes and behavior), the effects of social-emotional learning programs are largely equivalent to those reported in the present review at both post-test (Durlak et al., 2011) and follow-up (Taylor et al., 2017).

However, the finding that some studies were associated with overall negative effects, as well as the inclusion of negative values in the prediction interval, raises the caution that character education can sometimes be a deleterious undertaking. WWCE (Berkowitz & Bier, 2007) reported finding a small number of studies (programs) that had iatrogenic effects but given that that project focused on 'what works,' those studies were excluded from analysis. In particular, in the present study none of the 21 studies of programs lasting less than one week was associated with a negative mean *g* value. In contrast, there were 22 instances where a program longer than one week had deleterious effects on average (12%). Unfortunately, most character education programs were the focus of only one study, making it difficult to identify programs that are reliably harmful or reliably superior, given the relative infrequency in implementing and/or assessing various programs due to various budgetary and practical restraints.

We also found 43% of studies that examined clusters of students analyzed the results as if all participants were independent. This is an incorrect analytic strategy, one that increases the likelihood of Type I errors (detecting a difference from control in the absence of a difference). It is very likely that the number of significant hypothesis tests is inflated in a substantial portion of character education programs (Hedges & Hedberg, 2007). We urge the use of mixed models analyses correcting for cluster effects in all studies that randomize participants in clusters. The estimation of the clustering effect in every study will also allow future meta-analyses to use study-specific values for those effects rather than the generic value of .15 used in the present review.

Issues with selection bias

The results also highlight some important issues in this literature, particularly the troubling evidence of selection bias. Given the effort needed to mount a comparative study of the effectiveness of character education, the finding from this analysis that character education is generally associated with positive effects, and the likelihood of obtaining significance considering the large samples available to educational researchers and the neglect of correction for clustering, we suspect this bias is not largely a function of the so-called 'file drawer problem' (Rosenthal, 1979), the tendency not to pursue publication in response to failing to obtain significant results. More likely, though, the finding primarily reflects a tendency towards opportunistic biases in the conduct of the research (DeCoster et al., 2015). Such biases could include selectively omitting outcomes that contradict the desired conclusion, ad hoc addition of covariates or moderators to achieve significance, and modifying hypotheses to match post hoc analyses that proved significant (Simmons et al., 2011). Even if publication bias is not a significant issue in the character education literature, selection bias in the presentation of findings seems to be.

Our efforts to correct for these biases still suggest character education is effective, but the true mean effect for character education could be substantially lower than our best single estimate of 0.24. Recent work on replicability of research suggests this issue of overestimation of effect sizes is potentially widespread and not specific to character education research (e.g., Fraser et al., 2018; John et al., 2012; Simmons et al., 2011). We thus recommend future researchers interested in character education employ newly emerging best practices, including pre-registration of study designs, variables, and hypotheses (Nosek et al., 2018, 2019) and transparency statements that one has reported all experimental manipulations and variables (Simmons et al., 2011).

Moderation effects

Though we tested a large variety of potential moderators, length of mentorship was the only one that merited interpretation beyond the moderation analyses. That is, one-shot or very brief interventions were associated with substantially larger mean effects than longer interventions. One possible explanation for this finding was noted above, that (for whatever reason) longer interventions run a greater risk of deleterious outcomes. Another possibility is that brief programs may be better controlled, reducing the potential for other factors or statistical noise to dampen the effect. Alternatively, this may be an example of the economic principle of diminishing returns: shorter interventions may afford researchers the opportunity to detect transitory state-level changes that are not enduring and most germane to specific targeted outcomes that may be less generalizable outside of the experimental setting (Duncan et al., 2007). This interpretation is consistent with the trend across the four length groups towards smaller mean effects. It is also important to remember that the mean effects for all four lengths of treatment were significant, so on average there is reason to believe character education is helpful no matter what the length. Nonetheless, our discussion of this possibility remains speculative and warrants additional research in understanding the longevity of intervention effects.

Attempts to draw conclusions about the effectiveness of individual programs were hampered by the lack of replication efforts, and by the even rarer instances of replication across research teams. Ours seems to be the first study to raise issues of the need for greater replication efforts in the study of character education. The widespread practice of conducting a single investigation of a program undermines efforts at creating a cumulative science of character education intervention. Despite the limitations of the available literature, we were able to conclude that three programs—Cognitive Problem-Solving, Strong Kids, and the Kohlberg Moral Development program—generated mean effects that were particularly strong across multiple research groups. The paucity of replication studies also made it impossible to evaluate whether effectiveness of specific programs varied across levels of the moderator variables.

In terms of study quality, the average study was judged low on 3.4 threats to study quality and high on 2.3 threats. The most common high-risk finding was lack of blindness in the treatment providers (84% of studies), which is a ubiquitous limitation in all psychosocial intervention research. However, 58% were also judged as failing to ensure blindness in assessors, while limitations in random assignment to treatment was evident in 38% of studies. Of course, practical considerations (e.g., outcomes are only available from treatment providers or participants, or the conditions under which schools agree to participate negates the possibility of random assignment) can often render it impossible for a program to mitigate these risks. When considering only studies with a balance of experimental rigor, the overall effect shrank but remained significant.

Limitations and future directions

The current study has several limitations that future research may be able to address. First, a primary interest in this work was to minimize the number of low-quality studies in character education by considering only those with adequate rigor, so we largely excluded unpublished findings (see Cooper, 2003, for a critique of this view). Famously, Lipsey and Wilson (1993) found unpublished effects on average were only 2/3 the size of published effects. We also were limited to studies published in English, which could represent a cultural bias in our findings. It remains possible that our conceptualization of character education literature may be rather homogeneous and less sensitive to varying definitions (McGrath et al., in press). These results could represent what cross-cultural psychologists regard as a WEIRD (Western, Educated, Industrialized, Rich, and Democratic) definition of character education that might not represent the majority of the world's population (e.g., Henrich et al., 2010). Variability does exist cross-culturally in terms of what constitutes character and how character is conceptualized (McGrath & Brown, 2020). Although many interventions could target more communal traits, different definitions of communality and flourishing may exist within collectivistic societies compared to interventions that may have more individualistic bases behind them in Western cultures (Cai et al., 2011; Markus & Kitayama, 2010).

Even within the U.S., several cross-cultural differences exist that lead to region-specific values for engaging others. For example, the 'Culture of Honor' in the American South emphasizes ingroup loyalty (Brown, 2022; Brown et al., 2021). The aggressive response toward those who have affronted a man's honor is viewed as a moral imperative (see Cohen et al., 1996). Such viewpoints could be at odds to other regions within the country and may point to considerable variability on a regional basis. Different conceptualizations of which behaviors constitute character could influence the effectiveness of interventions of specific outcomes. A future investigation would benefit from addressing the U.S. studies in identifying potential within-country differences.

Endemic to science is an additional lack of access to unpublished findings, results in technical reports that have not been peer-reviewed, or work in other languages. A replication of this research would benefit from soliciting data from these sources specifically. Furthermore, the time necessary to conduct this analysis with its onset cut-off date of 2017 requires consideration of further character education studies published following 2017.

Another critical limitation emerges in the low number of high-quality studies. With this meta-analysis's goal of providing an exhaustive overview of previous findings, it becomes problematic for researchers to make broad generalizations of the reported effects when only 33 studies could be deemed high-quality. Meta-analytic investigators cannot control the quality of studies in the extant literature, but we hope that future syntheses of the literature can identify a larger group of high-quality studies. Additionally, this approach is entirely quantitative and necessarily requires complementary analyses to provide a wider scope of character education in various outcomes. Future research would benefit from conducting broad-level analyses of qualitative data to understand additional nuance in these outcomes.

Conclusion

Character education has become a widespread set of interventions designed to improve positive outcome for youth. This popularity necessitates careful evaluation of whether and to what extent these interventions are indeed effective. The current meta-analysis provides the largest evidence base to date for the potential effectiveness of character education, while highlighting various limitations in the available literature. We hope this will spark more rigorous investigations of this important component of non-cognitive education going forward.

Note

Data and materials for this meta-analysis are available through Supplemental Data and at: https://osf.io/t3z6c/

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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.

Notes on contributors

Mitch Brown is Faculty at the University of Arkansas. His primary research area is in evolutionary psychology, investigating the interplay between social perception and motivations in how they shape interpersonal preferences. Much of this work considers the trade-offs individuals make in mating and affiliative decisions. His involvement in this project is based on a postdoctoral fellowship investigating character education.

Robert E. McGrath is a Professor of Psychology at Fairleigh Dickinson University, and Senior Scientist for the VIA Institute on Character. He has published extensively on the structure and development of character. His latest book is *The Power of Character Strengths*, co-authored with Ryan Niemiec. He also maintains research programs on the assessment of public safety candidates and professional issues in psychology.

Melinda C. Bier is Co-Director for the Center for Character and Citizenship (CCC), at UMSL where she has led program evaluation, the design and implementation of K-12 educational innovations, and professional development for teachers and school leaders. Most recently, Bier, co-PI, Marvin W Berkowitz and, expert practitioner colleagues, have been developing and testing servant leadership-oriented training and mentoring for deep understanding and implementation of character education in K-12 schools and Colleges of Education. With support from the Kern Family Foundation, John Templeton Foundation, and Templeton World Charities Foundation these models of individual and organizational change are being pilot tested in St. Louis, Kenya, Colombia, and Mexico.

Keith Johnson received his Ph.D. in clinical psychology from Fairleigh Dickinson University. Research projects have included conducting a meta-analysis of the What Works in Character Education Literature and examining use of performance validity instruments in neuropsychological testing. He is currently a neuropsychological postdoctoral resident at the Central Western Massachusetts Veterans Healthcare System.

Marvin W. Berkowitz is the McDonnell Professor of Character Education and Co-Director of the Center for Character and Citizenship at the University of Missouri–St. Louis, and UM System President's Thomas Jefferson Fellow. He is a developmental psychologist specializing in character development and education. He is author of *You Can't Teach Through a Rat, Parenting for Good* and more than 100 book chapters and journal articles, and is co-editor of the *Journal of Character Education*. His newest book, *PRIMED for Character* was published in April 2021. He is recipient of the Character Education Partnership's Lifetime Achievement Award (2006) and the Association for Moral Education's Good Work Award (2010) and Kuhmerker Career Award (2013).

References

A full list of refereces, including those articles used in the meta-analysis is available in OSF link.

Berkowitz, M. W. (2021). *PRIMED for character education: Six design principles for school improvement*. Routledge.

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*(1), 29–48.

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*(1), 33–51.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2013). *Comprehensive meta-analysis version 3*.

Brown, M. (2022). Preliminary evidence for an aversion to atheists in long-term mating domains in the Southern United States. *Journal of Social and Personal Relationships*, *39*(3), 711–733. https://doi.org/10.1177/02654075211045051

Brown, M., Tracy, R. E., Young, S. G., & Sacco, D. F. (2021). Crowd salience heightens tolerance to healthy facial features. *Adaptive Human Behavior and Physiology*, *7*(4), 432–446. https://doi.org/10.1007/s40750-021-00176-2

Cai, H., Sedikides, C., Gaertner, L., Wang, C., Carvallo, M., Xu, Y., O'Mara, E. M., & Jackson, L. E. (2011). Tactical self-enhancement in China: Is modesty at the service of self-enhancement in East Asian culture? *Social Psychological and Personality Science*, *2*(1), 59–64. https://doi.org/10.1177/1948550610376599

Coburn, K. M., & Vevea, J. L. (2019). *weightr: Estimating weight-function models for publication bias*. https://CRAN.R-project.org/package=weightr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An "experimental ethnography. *Journal of Personality and Social Psychology*, *70*(5), 945–960. https://doi.org/10.1037/0022-3514.70.5.945

Cooper, H. (2003). Psychological bulletin: Editorial. *Psychological Bulletin*, *129*(1), 3–9. https://doi.org/10.1037/0033-2909.129.1.3

DeCoster, J., Sparks, E. A., Sparks, J. C., Sparks, G. G., & Sparks, C. W. (2015). Opportunistic biases: Their origins, effects, and an integrated solution. *American Psychologist*, *70*(6), 499–514. https://doi.org/10.1037/a0039191

Diggs, C. R., & Akos, P. (2016). The promise of character education in middle school: A meta-analysis. *Middle Grades Review*, *N2*(4), 4. http://scholarworks.uvm.edu/mgreview/vol2/iss2/4

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

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*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

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*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS One*, *13*(7), e0200303. https://.org/10.1371/journal.pone.0200303

Harlacher, J. E., & Merrell, K. W. (2010). Social and emotional learning as a universal level of student support: Evaluating the follow-up effect of strong kids on social and emotional outcomes. *Journal of Applied School Psychology*, *26*(3), 212–229. https://doi.org/10.1080/15377903.2010.495903

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*(1), 167–178. https://.org/10.1037/0033-2909.117.1.167

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, *32*(2), 151–179. https://doi.org/10.3102/1076998606298040

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *32*, 341–370.

Hedges, L. V. (2009). Effect sizes in nested designs. In L. V. Hedges, H. Cooper, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis, 2nd Ed* (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*(1), 60–87. https://.org/10.3102/0162373707299706

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29. https://.org/10.1038/466029a

Higgins, J. P., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 4). John Wiley & Sons.

Hollon, S. D. (1999). Allegiance effects in treatment research: A commentary. *Clinical Psychology: Science and Practice*, *6*(1), 107–112. https://.org/10.1093/clipsy.6.1.107

Howard, R. W., Berkowitz, M. W., & Schaeffer, E. F. (2004). Politics of character education. *Educational Policy*, *18*(1), 188–215. https://doi.org/10.1177/0895904803260031

IntHout, 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*(1), 25. https://doi.org/10.1186/1471–2288-14-25

Jeynes, W. H. (2017). A meta-analysis: The relationship between parental involvement and Latino student outcomes. *Education and Urban Society*, *49*(1), 4–28. https://doi.org/10.1177/0013124516630596

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Johnson, K., McGrath, R. E., Bier, M. C., Brown, M., & Berkowitz, M. W. (in press). A meta-analysis of the what works in character education research. *Journal of Character Education*.

Lapsley, D. K., & Narváez, D. (2006). Character education. In A. Renninger, I. Siegel, W. Damon, & R. Lerner (Eds.), *Handbook of child psychology: Vol. 4. child psychology in practice* (6th, pp. 248–296). Wiley.

Leming, J. S. (1993). In search of effective character education. *Educational Leadership*, *51*(3), 63–71.

Leming, J. S., Hendricks-Smith, A., & Antis, J. (2000). An evaluation of the heartwood institute's "An ethics curriculum for children". *Presented at the Annual Meeting of the American Educational Research Association in New Orleans, LA.*

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

Linden, A. H., & Hönekopp, 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

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

Markus, H. R., & Kitayama, S. (2010). Cultures and selves: A cycle of mutual constitution. *Perspectives on Psychological Science*, *5*(4), 420–430. https://doi.org/10.1177/1745691610375557

McClellan, B. E. (1999). *Moral education in America: Schools and the shaping of character from colonial times to the present*. Teachers College Press.

McGrath, R. E. (2018). What is character education? Development of a prototype. *Journal of Character Education*, *14*(2), 23–35.

McGrath, R. E., & Brown, M. (2020). Using the VIA classification to advance a psychological science of virtue. *Frontiers in Psychology*, *3442*(7). https://doi.org/10.3389/fpsyg.2020.565953

McGrath, R. E., Han, H., Brown, M., & Meindl, P. (in press). What does character education mean to character education experts? A prototype analysis of expert opinions. *Journal of Moral Education*. https://doi.org/10.1080/03057240.2020.1862073

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*, 2600–2606.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*(1), 160–164. https://doi.org/10.1037/0033-2909.112.1.160

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 ng 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*(1), 62–87. https://doi.org/10.3102/1076998607302714

Schwarzer, G., Carpenter, J., & Rücker, G. (2010). Empirical evaluation suggests Copas selection model preferable to trim-and-fill method for selection bias in meta-analysis. *Journal of Clinical Epidemiology*, *63*(3), 282–288.

Shields, D. L. (2011). Character as the aim of education. *Phi Delta Kappan*, *92*(8), 48–53.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. *SSRN 2160588*. https://doi.org/10.2139/ssrn.2160588

Sojourner, R. J. (2012). The rebirth and retooling of character education in America. https://www.character.org/wp-content/uploads/Character-Education.pdf

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*(4), 1156–1171. https://doi.org/10.1111/cdev.12864

Tough, P. (2012). *How children succeed: Grit, curiosity, and the hidden power of character*. Houghton Mifflin Harcourt.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. https://doi.org/10.1007/BF02294384

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428–443. https://doi.org/10.1037/1082-989X.10.4.428

What Works Clearinghouse. (2014). *Procedures and standards handbook*. Version 3.0.

Wike, T. L. (2012). The effectiveness of a social skills intervention for preventing aggression in children: an evaluation of the making choices intervention. [*Unpublished doctoral dissertation*].

Williams, M. M. (2000). Models of character education: Perspectives and developmental issues. *The Journal of Humanistic Counseling*, *39*(1), 32–40. https://doi.org/10.1002/j.2164-490X.2000.tb00091.x

Wright, J. C., Warren, M. T., & Snow, N. E. (2020). *Understanding virtue: Theory and measurement*. Oxford University Press.

Appendix A

Search Terms

WWCE Search

school OR evaluation OR intervention OR program OR and project as well as terms such as prosocial OR virtues OR values education OR moral education OR civics education OR ethics OR school culture OR school climate OR social and emotional

1999–2014 Search

(non-cognitive OR noncognitive OR prosocial OR values education OR social and emotional OR SEL OR socioemotional OR socio-emotional OR character education OR peace education OR character strengths OR school climate OR school culture OR virtues OR positive psychology OR moral education) NOT ethics AND (school OR education) AND (evaluation OR evaluat* OR intervention OR program OR project)*

Limiters were also set for databases as follows:

- ERIC: Education level: elementary education, elementary secondary education, grade 1 to 12, high schools, middle schools, primary education, and secondary education; Language: English

- PsychINFO: Age groups: school age, adolescence; Methodology: Empirical study; Language: English; Population: human
- PubMed: Article Types: Clinical Trial; Evaluation Studies; Meta-Analysis; Randomized Control Trial; Systematic Review; Ages: Child; Adolescent; Species: Human; Languages: English
- Ebscohost (general): academic journals and dissertations

2014–2017 Search

(non-cognitive OR noncognitive OR prosocial OR values education OR social and emotional OR SEL OR socioemotional OR socio-emotional OR character education OR peace education OR character strengths OR school climate OR school culture OR virtues OR positive psychology OR moral education) NOT ethics AND (school OR education) AND (evaluation OR evaluate* OR intervention OR program OR project)*

Appendix B

Proposed Moderators

Moderators without description were yes-no judgments. Percentages represent the number of each category or the number of yes judgments for moderators without descriptions.

Program Attributes

- Length
  - One session (3.2%)
  - Up to one week (6.5%)
  - Up to one month (55%)
  - Longer (32%)
- Control group assigned activities (13.1%)
- Out-of-school component (19.6%)
- Student population ‘at risk’ (25.7%)
- Student population described in ways that indicated disadvantaged (32.7%)

Program Targets

- Moral Principles (32.7%)
- Moral Sentiment (71.5%)
- Emotion Regulation (61.7%)
- Performance Outcomes (72.9%)
- Civic Outcomes (11.2%)
- Intellectual Outcomes (9.3%)

Program Methods

- Direct Teaching Strategies (79.4%)
- Interactive Teaching Strategies (82.2%)
- Behavioral Management (31.3%)
- Organizational Change (20.5%)
- Mentorship Model
  - None (60.3%)
  - Formal mentoring (11.7%)
  - Modeling (14.5%)
  - Curriculum example cases (9.3%)
- Community Participation (26.6%)
- Service Learning (6.5%)
- Professional Development (26.6%)

Program Outcomes

- Intrapersonal Competency (31.7%)
- School Attitudes (6.5%)
- Prosocial Behavior (11.2%)
- Character Knowledge (10.7%)
- Risky Behavior (14.9%)
- Antisocial Behavior (21.9%)
- School Behavior (9.3%)
- Academic Achievement (4.67%)
- Attendance/Absence (2.8%)
- Problem-Solving (13.5%)
- Interpersonal Competency (33.6%)
- Victimization (5.6%)
- Social-Moral Cognition (8.4%)