

Meta-Analysis of Social-Emotional Learning (SEL) Interventions in Primary Schools

Edwin^{1,*}, Josiharu Edmund Franky¹, Albert Djohanes², Togap Parlindungan Simanjuntak², & Sachiko Takahi³

¹Theology Study Program, Sekolah Tinggi Teologi Kingdom, Indonesia

²Theology Study Program, Sekolah Tinggi Teologi Bethel Indonesia, Indonesia

³Center for the Studies of Higher Education, Tokai National Higher Education and Research System Nagoya University, Japan

*Corresponding email: edwin.sttkingdom@gmail.com

Received: 25 July 2026

Accepted: 25 August 2026

Published: 14 September 2026

Abstract: Universal social-emotional learning (SEL) is widely implemented in primary schools, yet reported effects range from negligible to implausibly large, and syntheses that separate methodological quality from true intervention effects remain scarce. This meta-analysis aimed to estimate the effect of universal, school-based SEL interventions, relative to conventional schooling, on primary-school children's social-emotional competence, and to test how that estimate depends on study design, risk of bias, and small-study effects. Following PRISMA 2020, 4,043 records were identified from Scopus (2016–2025), 882 reports were screened at title and abstract, and 93 full texts were assessed for eligibility; 18 studies met the inclusion criteria (15 controlled between-group and three single-group pre-post designs), representing roughly 11,900 pupils across 13 countries. Study quality was appraised with ROBINS-I, and effect sizes (Hedges' g) were pooled with random-effects models using the DerSimonian-Laird, REML, and Paule-Mandel estimators with Hartung-Knapp confidence intervals, complemented by meta-regression, influence diagnostics, and trim-and-fill. The naïve pooled estimate was large ($g = 0.89$) but extremely heterogeneous and driven by three higher-risk outliers; after excluding them, the effect stabilized at a moderate g of 0.53 under DerSimonian-Laird and 0.55 under REML and Paule-Mandel, falling to $g = 0.18$ when restricted to lower-risk studies. At the same time, small-study bias persisted after outlier removal. We conclude that universal SEL interventions are associated with a genuine but modest-to-moderate improvement in social-emotional competence (credible $g = 0.53$; conservative low-risk $g = 0.18$) when judged by adequately controlled, lower-risk designs, whereas the broader literature overstates effectiveness because underpowered, higher-risk studies exert disproportionate influence.

Keywords: social-emotional learning, primary school, meta-analysis, effect size, risk of bias.

Article's DOI: <https://doi.org/10.23960/jpp.v16i3.pp1784-1808>

■ INTRODUCTION

The central challenge in evaluating social-emotional learning (SEL) is not whether such programs are popular but whether their measured effects can be trusted: across primary-school studies, reported outcomes for essentially the same construct range from negligible to implausibly large, and this inconsistency in how social-emotional outcomes are measured and reported makes the evidence base difficult to

interpret. Schooling is increasingly understood as an enterprise that must develop not only academic knowledge but also the intrapersonal and interpersonal capabilities on which learning, wellbeing, and later life outcomes depend (Dickinson & Peterson, 2015; Ferguson-Patrick, 2023; Hubbard, 2019; Mosley & Niwano, 2012; Sancassiani et al., 2015). Within this paradigm, social-emotional learning (SEL) has become one of the most frequently advocated

frameworks in primary education (Browne & Shapiro, 2016; Cornell et al., 2017; Durlak et al., 2015; Gómez-Carmona et al., 2025; Greenberg et al., 2017). Organized around the acquisition of self-awareness, self-management, social awareness, relationship skills, and responsible decision-making (Swisher & Doyle, 2024) SEL programming is claimed to build the competencies that underpin classroom engagement, peer relationships, and mental health, and to do so most effectively during the primary-school years, when social-emotional capacities are most plastic (Bird & Sultmann, 2010; Close et al., 2023; Lohr et al., 2021; Sklad et al., 2012).

Despite this enthusiasm, the empirical picture is difficult to interpret. Primary studies differ markedly in design, sample size, intervention duration, outcome measurement, and statistical reporting, and the reported benefits vary from essentially null in some of the largest cluster-randomized trials (Kiviruusu et al., 2016) to several standard deviations in small quasi-experiments—magnitudes that strain credulity for any classroom intervention (Wigelsworth et al., 2016). Landmark syntheses have reported average effects on social-emotional skills in the range of 0.3 to 0.6 (Cipriano et al., 2023; Durlak et al., 2011; Sklad et al., 2012; Taylor et al., 2017). However, these estimates typically pool across school levels, designs, and risk-of-bias strata, leaving open the question of how much of the apparent benefit for primary-school children specifically reflects genuine pedagogical effect and how much reflects artifacts of methodological quality.

The present study addresses this gap through a meta-analysis that treats heterogeneity not as a nuisance to be averaged away but as a substantive object of inquiry. It estimates the effect of universal, school-based SEL interventions, relative to conventional schooling, on primary-school children's social-emotional competence;

it distinguishes controlled from single-group evidence and compares outcome domains; and it evaluates small-study and publication-bias effects directly. Three research questions guide the analysis. First, what is the pooled effect of SEL interventions on primary-school children's social-emotional competence? Second, how does the estimate depend on study design, methodological quality, and outcome domain? Third, to what extent is the literature affected by small-study effects, and what is the best-supported effect once these are taken into account?

The study's contribution is therefore twofold. Substantively, it offers educators and policymakers a defensible, quantitative estimate of how much universal SEL programming improves primary-school children's social-emotional competence, together with an indication of how confident one can be in that estimate. Methodologically, it demonstrates a transparent workflow for distinguishing credible from inflated effects in a literature dominated by small quasi-experiments, by combining a pre-specified outlier rule, design- and quality-based subgrouping, leave-one-out diagnostics, and regression-based tests of small-study effects. In doing so, the study responds to repeated calls for greater methodological rigor and reporting transparency in educational intervention research (Grant et al., 2017; Wigelsworth et al., 2016), and it models a form of critical synthesis that other areas of educational research could adopt.

Social-emotional learning traces its intellectual lineage to research on emotional intelligence (Cherniss et al., 2006; Salovey & Mayer, 1990) and to developmental models of social competence, and it was consolidated as a field by the Collaborative for Academic, Social, and Emotional Learning, whose framework organizes the construct into five interrelated competence clusters: self-awareness, self-management, social awareness, relationship skills,

and responsible decision-making (Durlak et al., 2015). In its mature form, school-based SEL involves sequenced, active, focused, and explicit instruction the SAFE criteria delivered universally to all pupils, typically by classroom teachers, and reinforced through classroom climate and whole-school practices (Durlak et al., 2011; Oberle et al., 2016). These design features matter for the present review, because they define the boundary of the construct: interventions that merely add isolated activities to otherwise conventional teaching are weaker instances of the model than those that embody systematic, curriculum-embedded competence instruction (Swisher & Doyle, 2024).

The primary-school years are widely regarded as the optimal window for universal SEL provision. Social-emotional capacities develop rapidly between the ages of five and twelve, children's peer ecologies stabilize, and the competencies acquired in this period predict adolescent adjustment, academic attainment, and adult outcomes including employment and mental health (Bird & Sultmann, 2010; Close et al., 2023; Jones et al., 2015; Lohr et al., 2021). Universal delivery in primary classrooms also avoids the stigmatization associated with targeted provision and reaches children who would not otherwise be identified as at risk (Greenberg et al., 2017). A large family of manualized programs—Promoting Alternative Thinking Strategies (PATHS), Second Step, the RULER approach, Together at School, Papilio, and many national adaptations—has been developed for this age range, alongside curriculum-infused approaches in which SEL content is woven into regular subjects (Domitrovich et al., 2017). Primary studies of these programs report outcomes ranging from robust competence gains (Green et al., 2021; Kourmoussi et al., 2018) to null effects in large, well-controlled trials (Berry et al., 2016; Kiviruusu et al., 2016), a dispersion that motivates the design-sensitive synthesis undertaken here.

Beyond classic curricular programs, recent research has diversified SEL delivery. Mindfulness-based practices target the attentional and self-regulatory substrates of social-emotional competence (Lin et al., 2025); physical education, traditional games, and dance embed cooperation and emotion management in movement contexts (Filippou et al., 2019; Hartanto et al., 2021; Nur et al., 2025); computerized trainings scaffold socio-emotional cognition through observation and inference (Mayer et al., 2025); and cross-curricular vehicles such as school reading plans and computational-thinking projects carry emotional-intelligence content into academic instruction (Candeias et al., 2025; Marín-Marín et al., 2023). This diversity broadens the reach of SEL but also widens the methodological spectrum of the evidence base, because studies of movement- and technology-based variants are often smaller and less rigorously controlled than trials of established curricula. A synthesis that pools across these modalities must therefore attend closely to design quality, lest the enthusiasm of small pilot studies masquerade as the approach's effect.

Previous syntheses have been foundational in establishing that SEL programming works on average. Durlak et al. (2011) reported a mean effect of 0.57 on social-emotional skills across 213 universal school-based studies; Sklad et al. (2012) and Taylor et al. (2017) reported comparable short-term benefits and smaller but durable follow-up effects; and Cipriano et al. (2023), synthesizing 424 studies to 2020, reported significant but more modest improvements. However, these reviews generally combine primary and secondary school samples, mix randomized and uncontrolled designs, and stop short of interrogating whether the reported magnitudes are trustworthy. Where risk of bias has been examined, effects have been found to shrink substantially in better-controlled studies (Grant et al., 2017; Wigelsworth et al., 2016), and recent large-scale European trials have

repeatedly failed to replicate the headline effects of earlier small studies (Berry et al., 2016; Kiviruusu et al., 2016). This is a consequential omission. If the magnitude of the effect is partly a product of underpowered designs and selective reporting, then practice and policy recommendations built on headline effect sizes will be miscalibrated. The present study therefore adopts an explicitly critical stance: it estimates the effect of universal SEL interventions on primary-school children's social-emotional competence, but it also treats the dispersion of effect sizes, the role of study design and quality, and the presence of small-study effects as primary findings in their own right, and it separates controlled from uncontrolled evidence rather than averaging indiscriminately across them.

Accordingly, the present meta-analysis addresses three research questions:

1. What is the pooled effect of universal, school-based SEL interventions, relative to conventional schooling, on primary-school children's social-emotional competence?
2. To what extent do study design, risk of bias, sample size, and outcome domain moderate this effect?
3. Is there evidence of small-study effects or publication bias that would qualify the pooled estimate?

■ METHOD

Research Design

This review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement (M. Page et al., 2024; M. J. Page et al., 2021, 2022). The eligibility criteria, search strategy, extraction procedure, and analysis plan are described in the subsections below, and the flow of records is summarised in Figure 1. The review protocol was not registered in an online database (such as

PROSPERO or OSF) prior to the study; as an investigator-initiated secondary synthesis conducted within a limited timeframe, no prospective registration was undertaken. To preserve transparency, however, this article reports the full eligibility criteria, search strategy, extraction procedure, and analysis plan, and the complete data and analysis code are made openly available as supplementary materials.

Search Strategy

The primary corpus was retrieved from Scopus, restricted to documents published between 2016 and 2025 in English. The search combined social-emotional learning terms with intervention terms and primary-school population terms using the following Boolean string:

(*“social-emotional learning”* OR *“SEL”* OR *“emotional intelligence”* OR *“social skills”*) AND (*“intervention”* OR *“program”* OR *“strategy”* OR *“approach”*) AND (*“primary school”* OR *“elementary school”* OR *“children”* OR *“youth”*)

The search identified 4,043 records. Because full-text statistical extraction was required, the corpus was narrowed to the 887 documents for which full texts could be sought; 882 full-text reports were successfully retrieved and constituted the screening corpus. All reports were held to the same eligibility, extraction, and risk-of-bias standards.

The 882 retrieved reports were first screened on title and abstract against the population, intervention, outcome, and design criteria; 789 were excluded because the population was not a mainstream primary-school sample ($n = 267$), the sample was clinical or special-educational ($n = 149$), the report was not an intervention study ($n = 124$), no quantitative evaluation design was employed ($n = 72$), the sample was preschool-only ($n = 48$), the population was otherwise ineligible ($n = 48$), or

the report was a protocol ($n = 42$), a review ($n = 22$), or lacked a social-emotional outcome ($n = 17$). The remaining 93 reports were assessed in full text. Of these, 75 were excluded, most commonly because the study was correlational or otherwise not an eligible intervention evaluation ($n = 17$), the sample was clinical or special-educational ($n = 16$), the age group fell outside the primary-school range ($n = 14$), no social-emotional competence outcome was measured ($n = 12$), the reported statistics were not extractable or contained anomalies precluding valid effect-size computation ($n = 9$), the report was qualitative or perception-only ($n = 5$), or it

was a protocol without outcome data ($n = 2$). The final synthesis therefore comprised 18 studies: 15 controlled between-group studies (Calhoun et al., 2020; Coelho et al., 2023; Filippou et al., 2019; Green et al., 2021; Hartanto et al., 2021; Ishchenko et al., 2024; Kiviruusu et al., 2016; Ko³odziejczyk et al., 2025; Kourmoussi et al., 2018; Lechner et al., 2025; Lin et al., 2025; Marín-Marín et al., 2023; Mayer et al., 2025; Resurrección et al., 2021; Santamaría-Villar et al., 2021) and three single-group pre-post studies (Candeias et al., 2025; Moreira et al., 2025; Nur et al., 2025) with a combined sample of approximately 11,900 pupils across 13 countries.

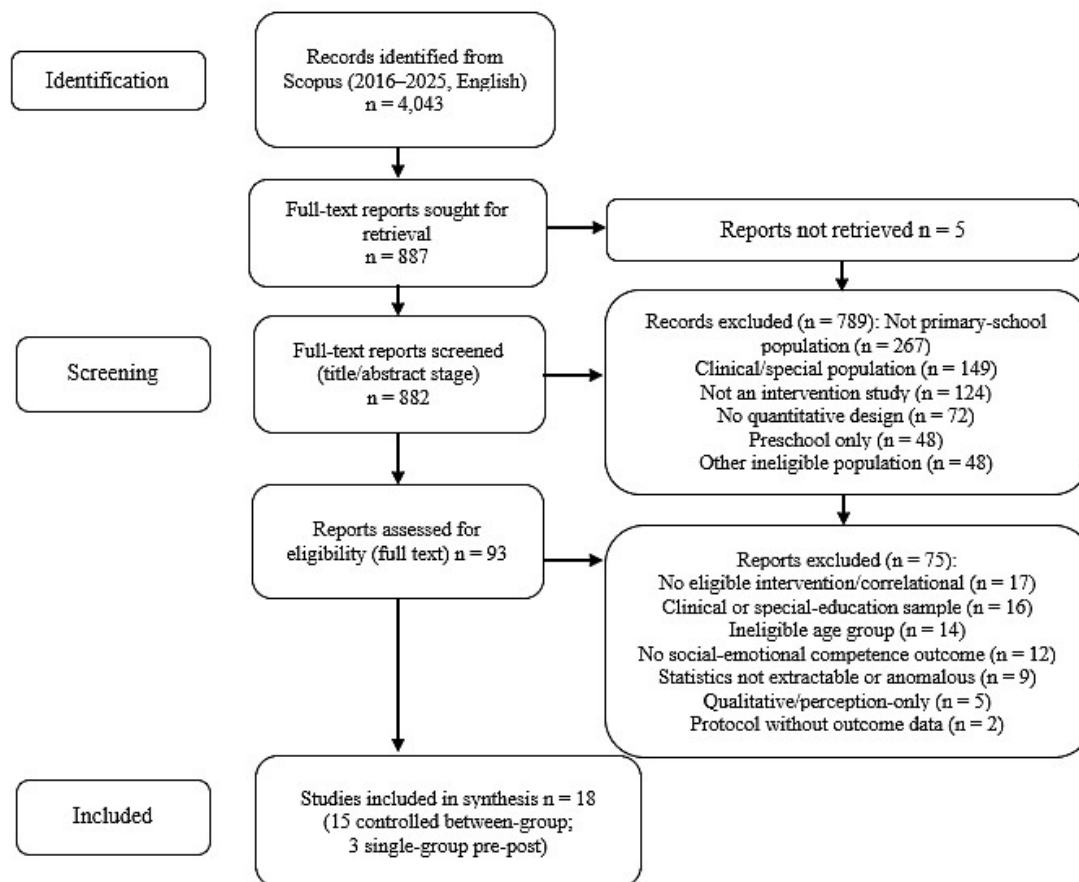


Figure 1. PRISMA 2020 flow diagram for study identification, screening, and inclusion

Inclusion Criteria

Studies were eligible if they implemented a universal social-emotional learning intervention including curriculum-based programs,

mindfulness-based practices, and SEL-oriented physical education, games, dance, or technology-supported variants as the primary instructional treatment in mainstream primary (elementary)

schools; sampled general primary-school populations of approximately five to twelve years of age; measured a social-emotional competence outcome, namely social-emotional skills, social skills or social competence, emotional intelligence, or emotion recognition and regulation, with a quantitative instrument; used either a between-group controlled design comparing the intervention with conventional schooling or a single-group pre-post design; reported statistics sufficient to compute a standardized effect size, such as group means and standard deviations, a *t* or *F* statistic with sample sizes, a partial eta-squared, or a reported Cohen's *d*; and were available in English. Studies of clinical or special-education samples (for example, autism spectrum disorder, ADHD, or specific learning disorders), preschool-only or secondary-only samples, correlational and validation studies, reviews, study protocols, qualitative or perception-only evaluations, and reports lacking extractable statistics were excluded to preserve construct validity and analytic rigor.

Data Analysis

Data extraction and effect-size computation. For each study, the authors extracted bibliographic details, country and grade level, program type, design, sample sizes, intervention and comparison conditions, outcome instrument and informant, and the statistics required for effect-size computation. For controlled pre-post designs reporting cell means and standard deviations, the pre-post-control standardized mean difference of Morris (2008) was computed, standardizing the difference in pre-to-post change between groups by the pooled pre-test standard deviation. Where studies reported a time-by-group interaction *F* statistic, the corresponding change-score contrast was converted to the raw-score metric assuming a pre-post correlation of 0.60, which is reported transparently (Borenstein, 2009); ANCOVA *F*

statistics and partial eta-squared values were converted through the standard relations $d = t / \sqrt{(1/n_1 + 1/n_2)}$ with $t = \sqrt{F}$, and $d = 2\sqrt{(\eta^2/(1-\eta^2))}$. One study reporting only a Mann-Whitney *U* statistic was converted through the normal approximation to *z* and thence to *d*. For single-group designs, a standardized mean change was computed from pre-test and post-test means and the pre-test standard deviation, or from the paired *t* statistic, again imputing a pre-post correlation of 0.60. Where a study reported several conceptually parallel social-emotional competence subscales, subscale effects were aggregated into a single study-level composite assuming a within-study correlation of 0.60 (Borenstein et al., 2009, 2011) so that each study contributed one effect and no study was represented more than once in any pooled model. All estimates were converted to Hedges' *g* by applying the small-sample correction (Hedges, 1981). The provenance and derivation of every effect size are documented in the supplementary analysis scripts (Table S2). For the one study reporting only a Mann-Whitney *U* statistic (Resurrección et al., 2021), the rank-based statistic was converted to a standardized effect by transforming *U* to a *z* score and then to $r = z / \sqrt{N}$ before conversion to *d*, following the distribution-free procedures described by Rosenthal (1991). A single pre-post correlation of $r = .60$ was imputed for all change-score and composite-aggregation variances, a conventional value in the absence of reported test-retest correlations (Borenstein et al., 2009); a sensitivity analysis varying *r* across .50, .60, and .70 left the credible pooled estimate essentially unchanged (Hedges' *g* = 0.53 at $r = .50$ and .60, and 0.62 at $r = .70$), confirming that the aggregation of dependent effects did not distort the between-study variance.

Risk-of-bias assessment. The methodological quality of the included studies was appraised with the Risk of Bias in Non-

randomized Studies of Interventions (ROBINS-I) tool, because a majority of included studies employed quasi-experimental or single-group designs rather than individually randomized controlled trials (Sterne et al., 2016). Each study was rated across domains covering confounding, participant selection, missing data, outcome measurement, and selection of the reported result. Domain-level and overall judgments were expressed as low, moderate, or serious risk of bias, with single-group pre-post designs assigned at least serious risk on the confounding domain because they cannot separate the intervention from maturation, practice, and regression to the mean. Each of the 18 studies was independently appraised by two authors using the ROBINS-I tool, with each domain scored separately and each rater blinded to the other's judgments. Disagreements on the overall risk-of-bias rating were reconciled by discussion to consensus, with a third author adjudicating any residual disagreement; the domain-level profiles shown in Table 2 are those of the primary appraiser, whereas each study's overall rating is the reconciled consensus. Inter-rater reliability before reconciliation was quantified with Cohen's κ on the overall risk-of-bias ratings ($\kappa = 0.63$, 95% CI [0.30, 0.95], indicating substantial agreement), computed in SPSS (Crosstabs) and cross-checked with the *irr* package in R, and interpreted following the benchmarks of Landis & Koch (1977). The consensus ratings were used descriptively and in sensitivity analyses rather than to weight studies; the pre-specified outlier rule (studies whose 95% confidence-interval lower bound exceeded 1.5) together with the leave-one-out and influence diagnostics served as complementary checks, with all three signals converging on the same three studies (S4, S9, S12).

Because between-group and single-group designs estimate different parameters and operate on different scales, they were analyzed separately, with controlled studies forming the primary

analysis and single-group studies a secondary analysis. Effects were pooled with random-effects models. To ensure that conclusions did not depend on the choice of heterogeneity estimator, the between-study variance was estimated with three methods, namely DerSimonian-Laird, restricted maximum likelihood, and Paule-Mandel, and confidence intervals were additionally computed with the Hartung-Knapp adjustment (DerSimonian & Laird, 2015; Inthout et al., 2014; Veroniki et al., 2016). Heterogeneity was quantified by Cochran's Q , the I^2 statistic, and the between-study variance, and the dispersion of true effects by 95% prediction intervals.

Robustness was assessed through leave-one-out analysis, influence diagnostics (studentized deleted residuals and Cook's distance), a Baujat plot, and univariate random-effects meta-regression on publication year and the logarithm of total sample size. A pre-specified, transparent outlier rule flagged an effect as statistically influential if the lower bound of its 95% confidence interval exceeded a Hedges' g of 1.5. This threshold was set a priori because standardized effects above 1.5 are rarely observed in universal classroom interventions and typically signal methodological artifacts or reporting anomalies rather than genuine instructional gains (Kraft, 2020); the influence diagnostics were used to confirm, not define, the flagged studies. A further sensitivity analysis restricted the pool to studies rated at low-to-moderate overall risk of bias. Small-study effects were evaluated with funnel plots, Egger's regression, and the Duval and Tweedie trim-and-fill procedure (Duval & Tweedie, 2000; Egger et al., 1997). All analyses were conducted in Python 3 using the NumPy, SciPy, and Matplotlib libraries, with random-effects and heterogeneity computations implemented from the formulae in Borenstein et al. (2009); all analysis scripts and the dataset are provided in the supplementary materials. The threshold of 1.5 was not arbitrary:

in education, standardized mean-difference effects above roughly 0.5 are already considered large, and effects approaching or exceeding 1.0 are highly atypical of rigorous field trials (Kraft, 2020), so a lower-confidence-bound criterion of 1.5 identifies effects well outside the plausible range for universal, school-based interventions.

■ RESULT AND DISCUSSION

The results are organised around the three research questions: the first subsection establishes the pooled effect of the controlled designs (Research Question 1); the moderator, subgroup, and meta-regression analyses then address how design, risk of bias, sample size, and outcome domain condition that effect (Research Question 2); and the funnel, Egger, trim-and-fill, and influence analyses address small-study effects and publication bias (Research Question 3), before the single-group designs are considered as secondary evidence.

Characteristics of the Included Studies

The 18 included studies were conducted in Taiwan, Poland, Portugal, Ukraine, the United States, Germany, England, Finland, Greece, Spain, Indonesia, and, in one multi-country implementation, Türkiye, Poland, Portugal, Romania, and Greece jointly. They covered grades 1 through 6 of mainstream primary education. Interventions ranged from manualized SEL curricula (PATHS, Together at School,

Papilio-6to9, Steps for Life, SPARK, Meetings with Leon) to curriculum-infused and teacher-delivered programs, mindfulness-based practice, computerized socio-emotional training, and SEL integrated into physical education, traditional games, dance, sport education, and a school reading plan. Fifteen of the 18 effects came from controlled between-group contrasts of SEL programming against conventional schooling; eight were cluster- or individually randomized, seven were quasi-experimental, and three were single-group pre-post designs. Outcomes were social-emotional competence measures reported by children, teachers, or parents. The combined sample was approximately 11,900 pupils (11,481 in controlled and 386 in single-group studies). Table 1 summarises the characteristics of the included studies. BG = between-group controlled; WG = single-group pre-post; R = randomized; Q = quasi-experimental. Studies are grouped by the construct measured (emotional competence vs. social competence) and ordered by effect size within each group to support cross-study comparison; RoB = overall ROBINS-I rating. The counts $k = 10$ and $k = 8$ include all 18 studies grouped by construct; the credible controlled subgroup analysis in Figure 3 is based on the eight emotional-competence and four social-competence between-group studies that remain once the three outliers (S4, S9, S12) and the three single-group designs (S16–S18) are excluded.

Table 1. Characteristics of included studies (18 studies)

ID	Study (year)	Grade(s)	Instrument/measure	Modality	Design	N (I/C)	RoB	Hedges' g [95% CI]
Emotional-competence measures ($k = 10$)								
S4	Ishchenko et al. (2024)	1–3	EI integral test	Digital (mobile app)	BG-Q	30/30	Serious	3.86 [3.04, 4.68]
S13	Marin-Marín et al. (2023)	primary	BarOn EQ-i:YV	Reading-plan (infused)	BG-Q	128/203	Moderate	1.26 [1.04, 1.48]
S5	Green et al. (2021)	4–5	CDP total	Mentoring	BG-R	47/47	Low	0.80 [0.38, 1.22]

S17	Candeias et al. (2025)	primary	COMPUSE L SEL comp. (5)	Computational thinking	WG	89	Serious	0.77 [0.59, 0.95]
S11	Resurrección et al. (2021)	ages 8–11	SEC – self-awareness	Curriculum	BG-Q	103/118	Serious	0.50 [0.24, 0.77]
S2	Kolodziejczyk et al. (2025)	K–primary	Emotion-competence tasks	Curriculum	BG-R (teams)	323/305	Moderate	0.27 [0.15, 0.39]
S1	Lin et al. (2025)	3–6	EI scale (self/teacher)	Mindfulness	BG-R (cluster)	123/158	Moderate	0.24 [0.03, 0.46]
S7	Mayer et al. (2025)	Year 3	Socio-emotional cognition test	Computerized	BG-R (cluster)	912/1826	Low	0.13 [0.05, 0.21]
S3	Coelho et al. (2023)	3–4	Socio-emotional skills (7 subscales)	Curriculum-infused	BG-Q	143/65	Moderate	0.03 [-0.18, 0.24]
S8	Kivuruusu et al. (2016)	1–3	MASCS / SDQ	Curriculum	BG-R (cluster)	1940/1594	Low	0.03 [-0.02, 0.08]
Social-competence / social-skills measures (k = 8)								
S9	Calhoun et al. (2020)	1	Social competence (PATHS)	Curriculum	BG-R	150/177	Serious	2.19 [1.92, 2.47]
S12	Santamaría-Villar et al. (2021)	3	MESSY & EQ-i	Curriculum	BG-Q	50/50	Serious	2.07 [1.67, 2.48]
S15	Filippou et al. (2019)	4–6	Social skills (3 factors)	Folk dance	BG-Q	132/74	Serious	1.60 [1.34, 1.85]
S14	Hartanto et al. (2021)	elementary	Social-skills scale	Games / PE	BG-R	45/45	Serious	1.09 [0.65, 1.53]
S18	Nur et al. (2025)	elementary	Social attitudes scale	Sport Education / PE	WG	90	Serious	0.77 [0.58, 0.97]
S6	Lechner et al. (2025)	1–2	SSIS-RS (6 subscales)	Curriculum (Papilio)	BG-R (pilot)	109/115	Moderate	0.46 [0.25, 0.68]
S10	Kourmoussi et al. (2018)	1–2	Personal & social skills (6)	Curriculum	BG-Q	1516/923	Moderate	0.37 [0.31, 0.43]
S16	Moreira et al. (2025)	4	Prosocial behaviour scale	Curriculum	WG	207	Serious	0.21 [0.09, 0.33]

Risk of Bias in the Included Studies

The ROBINS-I appraisal (Table 2) rated four studies at low overall risk of bias, two at moderate risk, and twelve at serious risk. Serious ratings were assigned to the three single-group

pre-post studies, which are confounded by design; to quasi-experiments with residual confounding or with anomalous or incomplete reporting (implausibly uniform gains, missing dispersion statistics, or scoring inconsistencies);

and to one randomized trial whose reported effects lay far outside the range of the remaining randomized evidence. Crucially, the three studies flagged by the pre-specified outlier rule were all rated at serious risk of bias, so the risk-of-bias appraisal and the quantitative influence diagnostics converge on the same conclusion: the credible

estimate rests on the controlled studies rated at low-to-moderate risk.

Individual Effect Sizes

Individual effect sizes are displayed, with their numerical values, in the forest plot in Figure 2. Among the controlled studies, the largest and

Table 2. Risk-of-bias appraisal (ROBINS-I). Shaded rows coincide with the statistically influential effects identified in the sensitivity analyses

ID	Study	Confounding	Selection	Missing data	Outcome	Reporting	Overall
S1	Lin et al. (2025)	Low	Low	Low	Low	Low	Low
S2	Kołodziejczyk et al. (2025)	Moderate	Moderate	Low	Moderate	Low	Moderate
S3	Coelho et al. (2023)	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
S4	Ishchenko et al. (2024)	Serious	Moderate	Low	Moderate	Serious	Serious
S5	Green et al. (2021)	Low	Low	Moderate	Moderate	Low	Low
S6	Lechner et al. (2025)	Serious	Moderate	Serious	Moderate	Moderate	Serious
S7	Mayer et al. (2025)	Low	Low	Low	Low	Low	Low
S8	Kiviruusu et al. (2016)	Low	Low	Moderate	Low	Low	Low
S9	Calhoun et al. (2020)	Moderate	Moderate	Moderate	Serious	Serious	Serious
S10	Kourmoussi et al. (2018)	Serious	Low	Low	Serious	Moderate	Serious
S11	Resurrección et al. (2021)	Moderate	Moderate	Low	Moderate	Serious	Serious
S12	Santamaría-Villar et al. (2021)	Moderate	Moderate	Low	Moderate	Serious	Serious
S13	Marín-Marín et al. (2023)	Serious	Moderate	Low	Moderate	Low	Serious
S14	Hartanto et al. (2021)	Moderate	Moderate	Serious	Moderate	Serious	Serious
S15	Filippou et al. (2019)	Moderate	Moderate	Low	Serious	Moderate	Serious
S16	Moreira et al. (2025)	Serious	Moderate	Low	Moderate	Low	Serious
S17	Candeias et al. (2025)	Serious	Moderate	Moderate	Moderate	Moderate	Serious
S18	Nur et al. (2025)	Serious	Moderate	Low	Moderate	Moderate	Serious

most rigorous trials clustered at the modest end of the distribution: the Finnish Together at School trial and the Portuguese socio-emotional program both yielded a Hedges' g of 0.03, the English SEE+ trial 0.13, the Taiwanese mindfulness trial 0.24, the Polish Meetings with Leon evaluation 0.27, and the Greek Steps for Life evaluation 0.37. A middle band comprised the German Papilio-6to9 pilot (0.46), the Spanish Learning to Be evaluation (0.50), the American SPARK trial (0.80), the Indonesian traditional-games

study (1.09), the Spanish reading-plan study (1.26), and the Greek folk-dance study (1.60). Three studies produced markedly larger estimates the Spanish socio-emotional-competencies quasi-experiment ($g = 2.07$), the American PATHS evaluation ($g = 2.19$), and the Ukrainian emotional-education quasi-experiment ($g = 3.86$) magnitudes far outside the range of the large controlled trials. The three single-group pre-post effects were 0.21, 0.77, and 0.77.

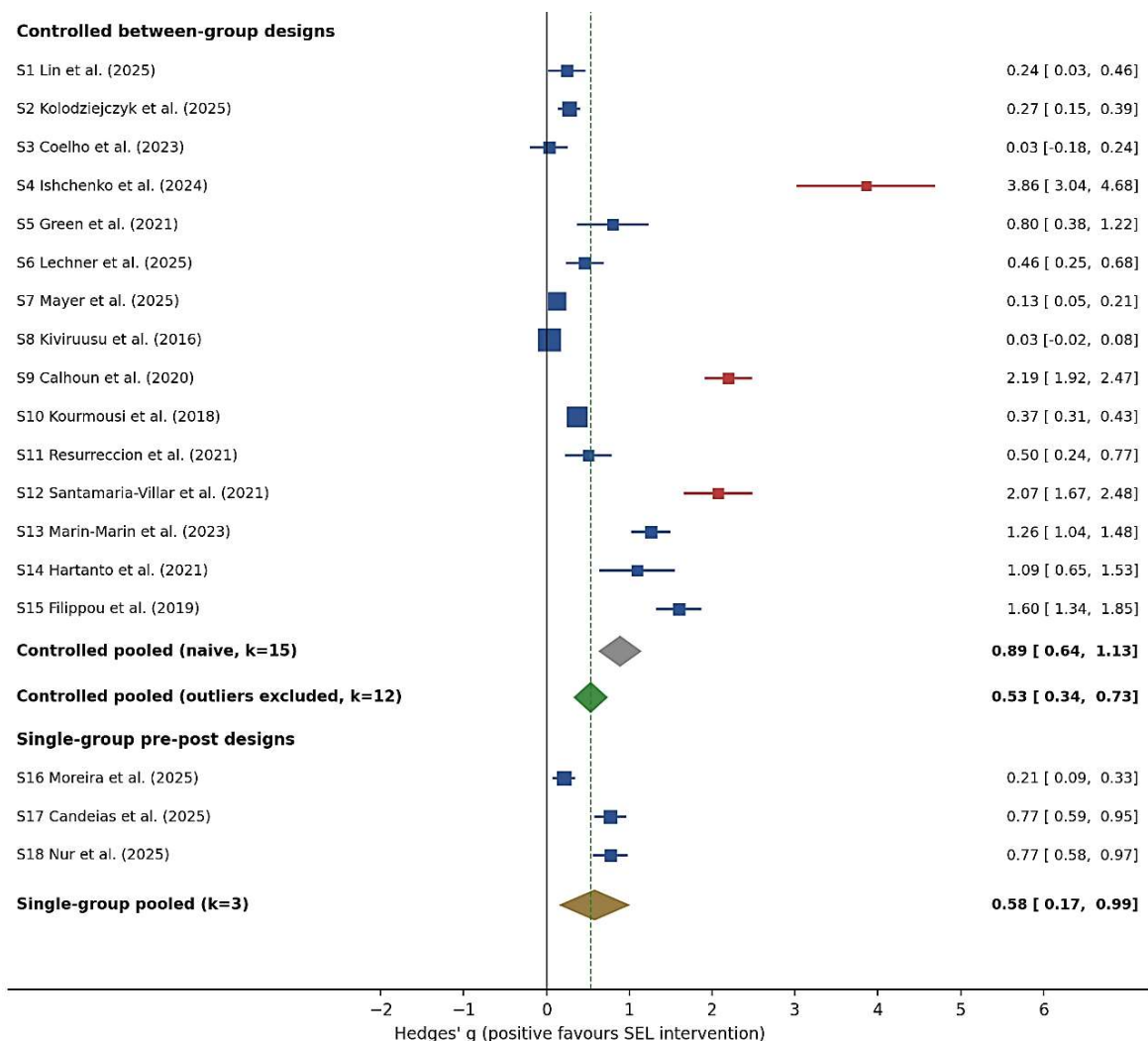


Figure 2. Forest plot of SEL intervention effects on social-emotional competence, grouped by study design. The green diamond is the controlled estimate after excluding the three influential outliers (S4, S9, S12)

Primary Analysis of the Controlled Designs (Research Question 1)

Pooling the 15 controlled effects under a random-effects model produced a large but heterogeneous estimate, with a Hedges' g of 0.89 and a 95% confidence interval from 0.64 to 1.13, a z of 7.08, and a p -value below .001; heterogeneity was extreme, with Q equal to 657.9 on 14 degrees of freedom, an I^2 of 97.9%, and a between-study variance of 0.21, and the prediction interval from "0.14 to 1.92 crossed zero. The forest plot and the influence diagnostics localised a disproportionate share of the heterogeneity to three studies: the Ukrainian quasi-experiment of Ishchenko et al. (2024), whose effect of 3.86 was accompanied by implausibly uniform post-test gains; the PATHS evaluation of Calhoun et al. (2020), whose model-based effects of about 2.2 exceed those of every other randomized PATHS trial in the literature; and the quasi-experiment of Santamaría-Villar et al. (2021), whose interaction statistics imply an effect of 2.07 in a sample of 100 pupils. The pre-specified outlier rule flagged exactly these three effects, and each had been independently rated at serious risk of bias.

With the outliers excluded, SEL interventions were associated with a robust moderate effect, with a Hedges' g of 0.53 and a 95% confidence interval from 0.34 to 0.73, a z of 5.40, and a p -value below .001. Heterogeneity remained substantial, with Q equal to 313.2 on 11 degrees of freedom, an I^2 of 96.5%, and a between-study variance of 0.10, and the prediction interval from "0.22 to 1.28 continued to cross zero, indicating that true effects vary widely across programs and contexts even within

the credible set. The leave-one-out analysis confirmed that the estimate did not depend on any single study, ranging from 0.43 to 0.59. A further sensitivity analysis restricted to the six studies rated at low-to-moderate overall risk of bias yielded a Hedges' g of 0.18 with a 95% confidence interval from 0.06 to 0.30, converging on the estimates of the largest, most rigorous trials. The credible controlled estimate of 0.53 (0.53–0.55 across estimators), bracketed from below by the more conservative low-to-moderate risk-of-bias estimate of 0.18 (0.18–0.21), is regarded as the best-supported answer to the first research question: an educationally meaningful but modest-to-moderate benefit that is considerably smaller than the headline figures advertised in parts of the literature.

Robustness Across Heterogeneity Estimators

To confirm that the credible estimate did not depend on modelling choices, the pooled effect was recomputed under three between-study variance estimators and with the Hartung-Knapp adjustment (Table 3). For the outlier-excluded controlled set the estimate was essentially unchanged across estimators, with a Hedges' g of 0.53 under DerSimonian-Laird and 0.55 under both restricted maximum likelihood and Paule-Mandel, and a Hartung-Knapp confidence interval of 0.21 to 0.86. The same stability held for the low-to-moderate risk-of-bias set (0.18 to 0.21) and for the outcome-domain subgroups. Results were therefore nearly identical across estimators, and the substantive conclusions are not artefacts of a particular random-effects specification.

Table 3. Pooled Hedges' g under three heterogeneity estimators with Hartung-Knapp (HK) confidence intervals. DL = DerSimonian-Laird; REML = restricted maximum likelihood; PM = Paule-Mandel

Analysis	DL g [95% CI]	REML g [95% CI]	PM g [95% CI]	HK 95% CI
Controlled — all (naïve)	0.89 [0.64, 1.13]	0.96 [0.45, 1.47]	0.96 [0.44, 1.49]	[0.36, 1.42]
Controlled — outliers excluded	0.53 [0.34, 0.73]	0.55 [0.27, 0.83]	0.55 [0.26, 0.84]	[0.21, 0.86]

Low-to-moderate risk of bias	0.18 [0.06, 0.30]	0.18 [0.06, 0.30]	0.21 [0.01, 0.40]	[-0.04, 0.39]
Emotional competence (no outlier)	0.38 [0.17, 0.59]	0.39 [0.11, 0.67]	0.39 [0.10, 0.69]	[0.02, 0.74]
Social skills (no outlier)	0.87 [0.31, 1.42]	0.86 [0.35, 1.38]	0.87 [0.30, 1.44]	[-0.06, 1.79]

Moderator Analyses (Research Question 2)

Within the credible set, effects were examined by outcome domain (Figure 3). Because this subgroup analysis is confined to the credible controlled set, it excludes both the three influential outliers (S4, S9, S12) and the three single-group pre-post studies (S16–S18), and therefore does not use the full construct-based counts shown in Table 1: of the ten emotional-competence studies in Table 1, one is an excluded outlier (S4) and one is a single-group design (S17), which leaves eight, whereas of the eight social-competence studies, two are outliers (S9, S12) and two are single-group designs (S16, S18), which leaves four. The eight studies whose primary outcome captured emotional competence emotional intelligence, emotion recognition and regulation, or self-awareness pooled to a Hedges' g of 0.38 with a 95% confidence interval from 0.17 to 0.59, whereas the four studies with social-skills outcomes pooled to 0.87 with a wide interval from 0.31 to 1.42.

The social-skills subgroup, however, was dominated by the movement- and games-based studies with smaller samples and higher risk ratings, and its Hartung-Knapp interval crossed zero, so the apparent domain difference should be read as reflecting study quality and modality at least as much as outcome construct. Univariate random-effects meta-regressions told a consistent story: in the full controlled set, effect size decreased sharply with the logarithm of total sample size ($b = -0.42, p < .001$), and this association persisted within the credible set ($b = -0.21, p = .008$), whereas publication year was unrelated to effect size in both sets ($p = .640$ and $p = .460$). The persistence of the sample-size association within the credible set indicates a residual small-study effect even after the most extreme studies are excluded, addressing the second research question: the estimate depends far more on design quality and sample size than on the substantive characteristics of the interventions.

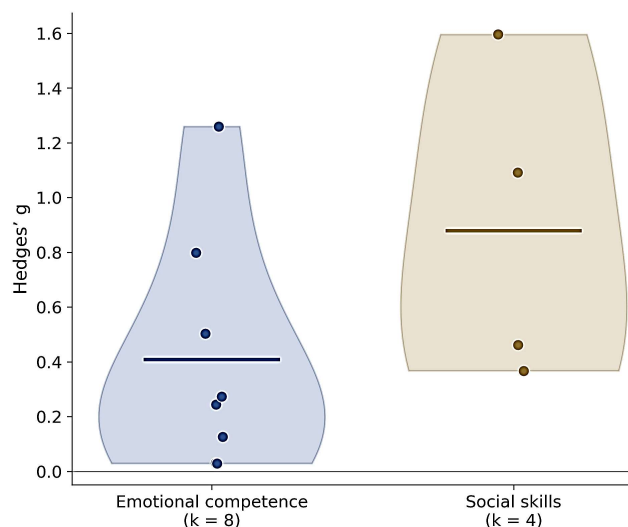


Figure 3. Violin plot of Hedges' g by outcome domain within the credible controlled set (emotional competence, $k = 8$; social skills, $k = 4$). Points are individual studies and the horizontal line marks the subgroup mean

Secondary Analysis of the Single-Group Designs

The three single-group pre-post effects pooled to a Hedges' g of 0.58, with a 95% confidence interval from 0.17 to 0.99 and an I^2 of 94.6%. Although this value is numerically close to the credible controlled estimate, it cannot be interpreted causally: single-group designs attribute to the intervention any improvement that in fact arises from maturation, test familiarity, additional instructional time, or statistical regression, and the small number of such studies precludes a stable estimate. Such designs are informative about within-cohort change but are not valid estimators of the causal effect of SEL programming, and they should not be combined with controlled evidence in a single headline figure.

Small-Study Effects and Publication Bias (Research Question 3)

The funnel plot in Figure 4 was conspicuously asymmetric: the most precise studies clustered between Hedges' g values of 0.03 and 0.37, whereas the least precise studies

fanned out toward values of 1 to 4. Egger's regression confirmed pronounced asymmetry for the full controlled set, with an intercept of 7.57 and a p -value of .002, and, importantly, the asymmetry remained significant after the three outliers were removed, with an intercept of 5.32 and a p -value of .029. The persistent inverse relationship between sample size and effect magnitude is the defining signature of small-study effects, encompassing publication bias, selective reporting, and lower methodological rigor in small studies. Because these mechanisms are observationally equivalent in a funnel plot, the asymmetry documents small-study effects and cannot, by itself, be attributed specifically to publication bias. This pattern addresses the third research question: the apparent grandeur of the effect in parts of the literature is substantially an artifact, and the credible effect is the modest-to-moderate estimate anchored by the larger, controlled trials.

Influence diagnostics corroborated the funnel-plot reading. In the full controlled set, the three flagged studies were the only markedly

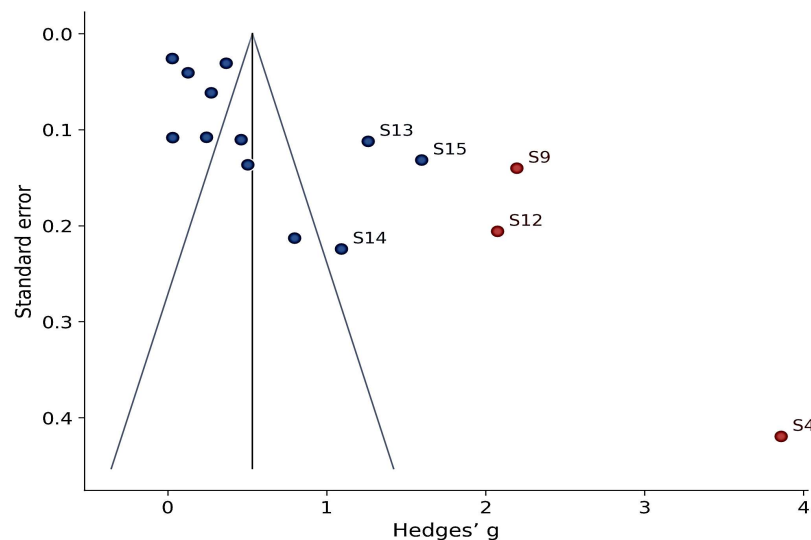


Figure 4. Funnel plot of effect size against standard error for the controlled effects. Asymmetry (Egger $p = .002$, persisting after outlier removal at $p = .029$) indicates strong small-study effects

influential ones: Ishchenko et al. (2024) had a studentized residual of 5.03 and a Cook's distance of 2.69, Calhoun et al. (2020) a

studentized residual of 3.32 and a Cook's distance of 4.00, and Santamaría-Villar et al. (2021) a studentized residual of 2.56 and a

Cook’s distance of 1.84, whereas the large trials all had studentized residuals below 2. These same studies also contributed most to heterogeneity and exerted the greatest influence on the pooled estimate. Applying the trim-and-fill procedure to the credible, outlier-excluded set imputed no additional studies and left the estimate unchanged

at 0.53; the small-study signal in this set is carried by the significant Egger slope and the sample-size meta-regression (Figure 5) rather than by funnel-count asymmetry, and the low-to-moderate risk-of-bias estimate of 0.18 provides the more conservative bound.

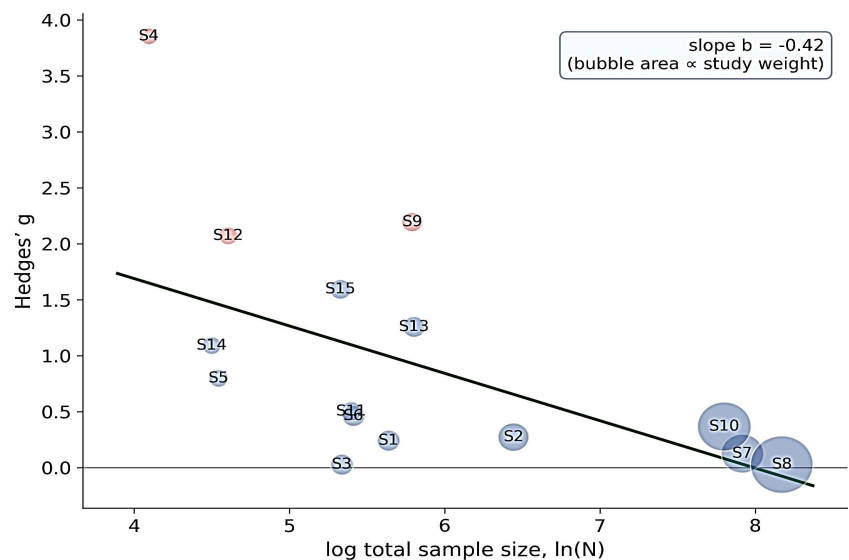


Figure 5. Bubble plot of the random-effects meta-regression of Hedges’ g on log total sample size for the 15 controlled studies. Bubble area is proportional to study weight; the diagonal line is the weighted meta-regression fit ($b = -0.42$). Larger studies report systematically smaller effects.

Summary of Pooled Estimates

Table 4 collates the random-effects estimates for every analysis, including the naive controlled pool, the outlier-adjusted controlled estimate, the risk-of-bias-restricted estimate, the two outcome subgroups, the single-group secondary analysis, and the design-mixed reference pool. The outlier-adjusted controlled estimate, bracketed from below by the low-risk

estimate, is highlighted as the best-supported summary of the effect of SEL interventions on primary-school children’s social-emotional competence.

Principal Findings

This meta-analysis estimated the effect of universal SEL interventions on primary-school children’s social-emotional competence and

Table 4. Random-effects (dersimonian-laird) pooled estimates. PI = 95% prediction interval. The shaded rows bracket the best-supported estimate

Analysis	k	g [95% CI]	z (p)	Q	I ²	τ ² / PI
Controlled — all (primary, naive)	15	0.89 [0.64, 1.13]	7.08 (<.001)	657.9	97.9%	0.21 / [−0.14, 1.92]
Controlled — outliers excluded	12	0.53 [0.34, 0.73]	5.40 (<.001)	313.2	96.5%	0.10 / [−0.22, 1.28]
Low-to-moderate risk of bias	6	0.18 [0.06, 0.30]	2.87 (.004)	28.6	82.5%	0.016 / [−0.21, 0.57]

Emotional competence (no outlier)	8	0.38 [0.17, 0.59]	3.57 (<.001)	141.6	95.1%	0.08 / [-0.36, 1.12]
Social skills (no outlier)	4	0.87 [0.31, 1.42]	3.04 (.002)	91.3	96.7%	0.30 / [-1.81, 3.54]
Single-group pre-post (secondary)	3	0.58 [0.17, 0.99]	2.79 (.005)	37.2	94.6%	0.12
All effects (reference only)	18	0.83 [0.61, 1.04]	7.59 (<.001)	712.2	97.6%	0.19 / [-0.13, 1.79]

interrogated, rather than suppressed, the heterogeneity that pervades this literature. Three findings stand out. First, when the evidence is restricted to adequately controlled studies and the three influential, higher-risk outliers are removed, SEL interventions are associated with a robust, moderate benefit, with a Hedges' g of about 0.53, declining further to about 0.18 when the pool is further restricted to studies at low-to-moderate risk of bias. Second, the estimate depends far more on study design, sample size, and methodological quality than on the substantive characteristics of the programs: the largest randomized trials by sample size yielded effects between 0.03 and 0.37, while small quasi-experiments regularly report effects several times larger. Third, and most consequentially, the wider literature is distorted by small-study effects: uncontrolled designs and a handful of small quasi-experiments report effects two to seven times larger than the credible estimate, and funnel-plot asymmetry remains pronounced even after outlier removal.

Interpretation and Relation to Prior Syntheses

These results reconcile an apparent paradox in the field. Foundational syntheses have advertised mean effects on social-emotional skills as high as 0.57 (Durlak et al., 2011), whereas recent large European trials have reported effects close to zero (Berry et al., 2016; Kiviruusu et al., 2016), and the most recent comprehensive synthesis reports substantially more modest

averages (Cipriano et al., 2023). The present analysis suggests that both readings contain an element of truth: universal SEL programming genuinely helps primary-school children, but the eye-catching magnitudes are largely manufactured by design weakness and selective reporting. The credible estimate of 0.53, bracketed below by the conservative low-risk value of 0.18, sits squarely between the early headline figures and the null results of the largest trials, and it converges with quality-restricted analyses in earlier reviews, which likewise found that effects shrink as methodological rigor increases (Wigelsworth et al., 2016). Practically, stakeholders should expect a solid, worthwhile improvement from well-implemented SEL programming, on the order of one-fifth to one-half of a standard deviation, rather than the transformational gains occasionally claimed. Because heterogeneity remained extreme ($I^2 = 96.5\%$) and the 95% prediction interval ("0.22 to 1.28) ranges from null or negative to large, the pooled g of 0.53 is an average across diverse programs, not the effect any single SEL program will reliably produce; program design and implementation context matter as much as the mean.

Taken together, these patterns suggest a general principle that we term the rigor-inflation trade-off: in this literature the apparent magnitude of SEL benefits is inversely tied to the methodological stringency of the evaluation, so that reducing design bias through randomization, adequate power, and pre-post-control adjustment systematically deflates the estimated

effect toward a stable, credible band around $g = 0.5$. Rather than treating divergent findings as contradictory, this framework reframes the field's heterogeneity as a predictable equilibrium between test validity and reported outcome size. It implies that future syntheses should model study rigor explicitly as a moderator rather than averaging across it.

The mechanism implied by the credible estimate is consistent with developmental theories of social-emotional competence. Sequenced, explicit competence instruction gives children a shared emotional vocabulary and rehearsed interpersonal routines; guided practice in emotion recognition, regulation, and cooperative problem-solving consolidates these into transferable skills; and classroom-level reinforcement sustains them (Bouffard & Rice, 2011; Kumschick et al., 2014; Rojas & Abenavoli, 2021). That effects were detectable across delivery modalities as different as manualized curricula, mindfulness practice, computerized training, and movement-based programs suggests that the active ingredient is the structured, repeated engagement with emotional and social content rather than any particular vehicle. At the same time, the wide prediction intervals indicate that true effects vary substantially across programs and contexts. Hence, a favorable average does not guarantee benefit for any particular implementation dosage, fidelity, and teacher preparation plausibly condition the benefit (Durlak & DuPree, 2008).

The Role of Delivery Modality

Notably, the largest credible effects came from movement- and games-based deliveries the Indonesian traditional-games study (1.09) and the Greek folk-dance study (1.60) and from a curriculum-embedded reading plan (1.26). In contrast, the large manualized-curriculum trials produced the smallest estimates. The present data cannot establish that embodied or curriculum-infused delivery amplifies the benefit, because

modality is confounded with sample size, country, and methodological quality within these studies: the movement-based studies were among the smaller and higher-risk designs, and their effects lie exactly on the small-study gradient identified by the meta-regression. The observation nonetheless identifies a promising and testable hypothesis for future controlled research that manipulates delivery modality while holding program content, dosage, and design quality constant.

The comparatively large effects of movement- and games-based delivery must, however, be weighed against differences in instructional time and cognitive demand across conditions. From the perspective of cognitive load theory (Sweller, 1988), embedding social-emotional practice in physical play may reduce the extraneous load imposed by abstract, language-heavy curricula, freeing working-memory resources for the intrapersonal and interpersonal processing that SEL targets; alternatively, such formats may simply provide more time-on-task or more motivating contexts than the business-as-usual comparison. Because the present data do not equate dosage or engagement across modalities, the apparent superiority of embodied delivery should be read as a hypothesis about load and motivation to be tested directly, not as evidence that physical formats are intrinsically more effective.

Implications for Practice and Policy

For teachers, school leaders, and policymakers, the credible estimate carries a clear and usable message. A benefit of one-fifth to one-half of a standard deviation is large enough to justify the planning and training effort that genuine SEL programming demands it is comparable to the most effective universal academic interventions (Kraft, 2020) yet modest enough that it should be presented to stakeholders as a solid improvement rather than a guaranteed

transformation. The convergence of the largest trials on smaller effects implies that expectations calibrated on early headline figures will be disappointed at scale, and that system-level rollouts should be accompanied by rigorous evaluation rather than assumed efficacy. The inflation evident in small and uncontrolled studies cautions against uncritically importing dramatic success stories, which may reflect design artifacts as much as pedagogical power.

Methodological Implications

Methodologically, the study illustrates why pooling across designs and quality strata is hazardous and why heterogeneity should be modeled as a finding rather than concealed. Had only the naive, design-mixed estimate been reported, the effect of SEL interventions would have been overstated by roughly two-thirds, and the decisive role of study design would have been obscured. The contrast between the inflated naive pool (0.89), the outlier-adjusted controlled estimate (0.53), and the low-risk estimate (0.18) is the central methodological contribution of this synthesis, and it offers a cautionary template for future quantitative reviews of classroom interventions (Sterne et al., 2016; Wigelsworth et al., 2016). Reviewers and editors in educational research might reasonably require that primary studies report cell means, standard deviations, and exact test statistics, so that effect sizes can be recomputed and small-study effects diagnosed.

The disproportionate effects observed in the small quasi-experiments are not merely a statistical artifact but have identifiable pedagogical roots. When intervention and comparison groups are not formed by randomization, they typically differ at baseline in teacher enthusiasm, classroom climate, and pupils' prior disposition toward the targeted skills, and the intervention is then delivered by implementers who are aware of the study's aims and often rate the outcomes

themselves. In social-emotional domains, where outcomes are frequently rated rather than objectively measured, these conditions allow expectancy and demand effects to masquerade as genuine learning, producing the illusion of rapid, uniform gains. This is why the absence of a randomly constituted control group in school-based SEL trials is especially corrosive: it converts ordinary selection and rater bias into effect-size inflation, as seen in studies S4, S9, and S12.

Limitations

Several limitations qualify these conclusions. First, the synthesis rests on a single database (Scopus) restricted to English-language reports, so eligible studies indexed elsewhere or published in other languages were not captured, and grey literature was not searched; because such omissions tend to inflate rather than deflate published effects, they do not threaten the conservative conclusion. This English-only, single-database scope nonetheless constrains the synthesis's external validity: the pooled estimate may not generalize to programs published in other languages or indexed in other databases, or to educational systems under-represented in Scopus. Second, several effect sizes required conversion from reported test statistics rather than raw cell means, and change-score conversions imputed a pre-post correlation of 0.60; both choices are reported transparently, and neither materially affects the headline estimates. Third, where studies reported multiple social-emotional subscales, a composite was formed assuming a within-study correlation of 0.60; alternative aggregation choices shift individual study effects only marginally. Fourth, heterogeneity remained extreme even within the credible set, and the prediction intervals crossed zero, so the pooled estimate should be read as a central tendency across widely varying program-context combinations rather than a guarantee. Fifth, intervention dosage, implementation fidelity, and

informant type could not be modeled as moderators because reporting was inconsistent. Finally, the included studies span 13 countries but derive predominantly from Europe and Asia, with limited representation of low-income settings, and the single-group evidence base was too small for meaningful secondary analysis. The estimate should be updated as further controlled evidence accumulates.

■ CONCLUSION

Universal social-emotional learning interventions are associated with a genuine, modest-to-moderate improvement in primary-school children's social-emotional competence. When the evidence is restricted to adequately controlled studies and three influential, higher-risk outliers are removed, the pooled effect is a Hedges' g of 0.53–0.55 that proved stable across the DerSimonian-Laird, restricted maximum likelihood, and Paule-Mandel estimators and robust to Hartung-Knapp confidence intervals and leave-one-out deletion; restricting the pool to studies at low-to-moderate risk of bias yields a more conservative 0.18, and the largest randomized trials by sample size reported effects between 0.03 and 0.37. The literature as a whole, however, systematically overstates the benefit: small, higher-risk studies report effects up to seven times larger, small-study effects persist after outlier removal, and single-group designs cannot support causal claims. For practice, the message is encouraging but sober well-implemented SEL programming appears to provide meaningful educational benefits, on the order of one-fifth to one-half of a standard deviation; for research, the message is that the field needs fewer small quasi-experiments and more adequately powered, transparently reported, randomized evaluations, particularly in low- and middle-income contexts. Heterogeneity in this literature is not noise around a settled truth but the central empirical fact that future primary

studies and syntheses must explain. With respect to the third research question, the evidence for persistent small-study effects is affirmative rather than incidental: Egger's test remained significant after the three outliers were removed ($p = .029$), the funnel plot was conspicuously asymmetric, and larger studies reported systematically smaller effects. Because publication bias is only one of several possible sources of this asymmetry alongside selective reporting and lower methodological rigor in small studies the current published record is best read as an upper bound on the true benefit of universal SEL programming.

■ DECLARATION OF GENERATIVE AI USAGE IN THE WRITING PROCESS

During the preparation of this manuscript, the author(s) used Claude (Anthropic) for language editing and refining sentence structure. After using this tool, the author(s) reviewed and edited the content as needed and took full responsibility for the article's content.

■ REFERENCES

- Berry, V., Axford, N., Blower, S., Taylor, R. S., Edwards, R. T., Tobin, K., Jones, C., & Bywater, T. (2016). The effectiveness and micro-costing analysis of a universal, school-based, social-emotional learning programme in the UK: A cluster-randomised controlled trial. *School Mental Health*, 8(2), 238–256. <https://doi.org/10.1007/s12310-015-9160-1>
- Bird, K. A., & Sultmann, W. F. (2010). Social and emotional learning: Reporting a system approach to developing relationships, nurturing well-being and invigorating learning. *Educational and Child Psychology*, 27(1), 143–155. <https://doi.org/10.53841/bpsecp.2010.27.1.143>
- Borenstein, M. (2009). Effect sizes for continuous data. In *The Hand. of Res. Synthesis and*

- Meta-Analysis, 2nd Ed.* (pp. 221–235). <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470743386.fmatter>
- Borenstein, M., Hedges, L. V, Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to Meta-Analysis. In *Introduction to Meta-Analysis*. John Wiley and Sons. <https://doi.org/10.1002/9780470743386>
- Borenstein, Michael., Hedges, L. V., Higgins, J. P. T., & Rothstein, Hannah. (2011). *Introduction to meta-analysis*. John Wiley & Sons Ltd.
- Bouffard, J. A., & Rice, S. K. (2011). The influence of the social bond on self-control at the moment of decision: Testing Hirschi's redefinition of self-control. *American Journal of Criminal Justice*, 36(2), 138–157. <https://doi.org/10.1007/s12103-010-9095-x>
- Browne, C. H., & Shapiro, C. J. (2016). Building young children's social-emotional competence at home and in early care and education settings. In *Innovative Approaches to Supporting Families of Young Children* (pp. 87–105). Springer International Publishing. https://doi.org/10.1007/978-3-319-39059-8_5
- Calhoun, B., Williams, J., Greenberg, M., Domitrovich, C., Russell, M. A., & Fishbein, D. H. (2020). Social emotional learning program boosts early social and behavioral skills in low-income urban children. *Frontiers in Psychology*, 11. <https://doi.org/10.3389/fpsyg.2020.561196>
- Candeias, A., Felix, A., Dumitrache, A., Almã'an, B., Gencel, I. E., Zadworna, M., Kossakowska, K., Sakellariou, A., Sakellariou, M., Sahin, C., & Taskin, C. S. (2025). Enhancing socio-emotional learning and mental health through computational thinking: a cross-cultural analysis of the COMPUSEL programme. *Frontiers in Psychology*, 16. <https://doi.org/10.3389/fpsyg.2025.1480731>
- Cherniss, C., Extein, M., Goleman, D., & Weissberg, R. P. (2006). Emotional intelligence: What does the research really indicate? *Educational Psychologist*, 41(4), 239–245. https://doi.org/10.1207/s15326985ep4104_4
- Cipriano, C., Strambler, M. J., Naples, L. H., Ha, C., Kirk, M., Wood, M., Sehgal, K., Zieher, A. K., Eveleigh, A., & McCarthy, M. (2023). The state of evidence for social and emotional learning: A contemporary meta-analysis of universal school-based SEL interventions. *Child Development*, 94(5), 1181–1204. <https://doi.org/10.1111/cdev.13968>
- Close, M., Killingly, C., Gaumer-Erickson, A. S., & Noonan, P. M. (2023). What is social and emotional learning (SEL) and why is it important? In *Inclusive Education for the 21st Century: Theory, Policy, and Practice, Second Edition* (pp. 275–301). Taylor and Francis. <https://doi.org/10.4324/9781003350897-17>
- Coelho, V., Peixoto, C., Azevedo, H., Machado, F., Soares, M., & Espain, A. (2023). Effects of a Portuguese social-emotional learning program on the competencies of elementary school students. *Frontiers in Psychology*, 14. <https://doi.org/10.3389/fpsyg.2023.1195746>
- Cornell, C., Kiernan, N., Kaufman, D., Dobee, P., Frydenberg, E., & Deans, J. (2017). Developing social emotional competence in the early years. In *Social and Emotional Learning in Australia and the Asia-Pacific: Perspectives, Programs and Approaches* (pp. 391–411). Springer Singapore. https://doi.org/10.1007/978-981-10-3394-0_21
- DerSimonian, R., & Laird, N. (2015). Meta-analysis in clinical trials revisited.

- Contemporary Clinical Trials*, 45, 139–145. <https://doi.org/10.1016/j.cct.2015.09.002>
- Dickinson, P., & Peterson, R. (2015). Flourishing schools in Aotearoa, New Zealand: The wellbeing in schools model. In *School Mental Health: Global Challenges and Opportunities* (pp. 170–185). Cambridge University Press. <https://doi.org/10.1017/CBO9781107284241.016>
- Domitrovich, C. E., Durlak, J. A., Staley, K. C., & Weissberg, R. P. (2017). Social-emotional competence: An essential factor for promoting positive adjustment and reducing risk in school children. *Child Development*, 88(2), 408–416. <https://doi.org/10.1111/cdev.12739>
- Durlak, J. A., & DuPree, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 36, 179–194. <https://doi.org/10.1007/s10464-008-9165-0>
- 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>
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, 315(7109), 629–634. <https://doi.org/10.1136/bmj.315.7109.629>
- Ferguson-Patrick, K. (2023). Teacher perceptions of primary-aged students' social capabilities in Australia, England, and Sweden and their importance to inclusive classrooms. In *Contemporary Global Perspectives on Cooperative Learning: Applications Across Educational Contexts* (pp. 82–93). Taylor and Francis. <https://doi.org/10.4324/9781003268192-9>
- Filippou, G., Derri, F., Mavridis, V., & Rokka, G. (2019). Traditional dances as a means of teaching social skills to elementary school students. In *International Journal of Instruction* (Vol. 12, Number 1). www.e-iji.net
- Gómez-Carmona, C. D., Ramírez-Arroyo, A. P., & Rico-González, M. (2025). Interventions to develop preschool-aged children's emotional competence through movement and instrument for its evaluation. In *Physical Education in Early Childhood: Movement and Development from 3 to 6 Years* (pp. 144–154). Taylor and Francis. <https://doi.org/10.4324/9781003615675-13>
- Grant, S., Hamilton, L. S., Wrabel, S. L., Gomez, C. J., Whitaker, A., Leschitz, J. T., Unlu, F., Chaves-Herrerias, E. R., Baker, G., Barret, M., Harris, M., & Ramos, A. (2017). *Social and emotional learning interventions under the every student succeeds*. RAND Corporation.
- Green, A. L., Ferrante, S., Boaz, T. L., Kutash, K., & Wheeldon-Reece, B. (2021). Evaluation of the SPARK Child Mentoring Program: A social and emotional learning curriculum for elementary school students. *Journal of Primary Prevention*, 42(5), 531–547. <https://doi.org/10.1007/s10935-021-00642-3>

- Greenberg, M. T., Domitrovich, C. E., Weissberg, R. P., & Durlak, J. A. (2017). *Social and emotional learning as a public health approach to education* (Vol. 27, Number 1). www.futureofchildren.org
- Hartanto, D., Kusmaedi, N., Ma'mun, A., & Abduljabar, B. (2021). Integrating social skills in traditional games with physical education interventions. *International Journal of Human Movement and Sports Sciences*, 9(5), 921–928. <https://doi.org/10.13189/saj.2021.090513>
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128. <https://doi.org/10.3102/10769986006002107>
- Hubbard, B. (2019). Developing life skills in children: A road map for communicating with parents. *Journal of Higher Education Theory and Practice*, 19(3), 33–54. <https://doi.org/10.33423/jhetp.v19i3.2115>
- 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). <https://doi.org/10.1186/1471-2288-14-25>
- Ishchenko, Y., Hubarieva, D., Soroka, I., Usyk, D., & Chemonina, L. (2024). Emotional education as a means of developing social competence in primary school students in wartime. *International Electronic Journal of Elementary Education*, 17(1), 1–13. <https://doi.org/10.26822/iejee.2024.359>
- Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health*. <https://doi.org/10.2105/AJPH.2015.302630>
- Kiviruusu, O., Björklund, K., Koskinen, H. L., Liski, A., Lindblom, J., Kuoppamäki, H., Alasuovanto, P., Ojala, T., Samposalo, H., Harmes, N., Hemminki, E., Punamäki, R. L., Sund, R., & Santalahti, P. (2016). Short-term effects of the “Together at School” intervention program on children's socio-emotional skills: A cluster randomized controlled trial. *BMC Psychology*, 4(1). <https://doi.org/10.1186/s40359-016-0133-4>
- Kołodziejczyk, J., Kołodziejczyk, J., & Walczak, B. (2025). An evaluation of social and emotional learning-based program for kindergarten and primary schools in Poland. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-08494-1>
- Kourmoussi, N., Markogiannakis, G., Tzavara, C., Kounenou, K., Mandrikas, A., Christopoulou, E., & Koutras, V. (2018). Students' psychosocial empowerment with the ‘Steps for Life’ personal and social skills Greek elementary programme. *International Electronic Journal of Elementary Education*, 10(5), 535–549. <https://doi.org/10.26822/iejee.2018.541303>
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Kumschick, I. R., Beck, L., Eid, M., Witte, G., Klann-Delius, G., Heuser, I., Steinlein, R., & Menninghaus, W. (2014). Reading and feeling: The effects of a literature-based intervention designed to increase emotional competence in second and third graders. *Frontiers in Psychology*, 5(DEC). <https://doi.org/10.3389/fpsyg.2014.01448>

- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Lechner, V., Ortelbach, N., & Scheithauer, H. (2025). Fostering of social, emotional, and cognitive skills in elementary schools: The Papilio-6to9 program. *Frontiers in Psychology*, 16. <https://doi.org/10.3389/fpsyg.2025.1599946>
- Lin, L. J., Lin, Y. H., Yu, S. P., Liu, T. H., & Chen, Y. L. (2025). A cluster randomized controlled trial examining the effects of a four-week mindfulness-based practice on primary school students' interpersonal mindfulness, emotional intelligence, and attentional focus. *Frontiers in Psychology*, 16. <https://doi.org/10.3389/fpsyg.2025.1539962>
- Lohr, A. M., Krause, K. C., McClelland, D. J., Van Gorden, N., Gerald, L. B., Del Casino, V., Wilkinson-Lee, A., & Carvajal, S. C. (2021). The impact of school gardens on youth social and emotional learning: A scoping review. *Journal of Adventure Education and Outdoor Learning*, 21(4), 371–384. <https://doi.org/10.1080/14729679.2020.1838935>
- Marín-Marín, J.-A., López-Belmonte, J., Lampropoulos, G., & Moreno-Guerrero, A.-J. (2023). Influence of the application of a reading plan on motivation, emotional intelligence, fluency and reading comprehension in Spanish primary school students. *Texto Livre*, 16, 1–16. <https://doi.org/10.1590/1983-3652.2023.41548>
- Mayer, S., Dumontheil, I., Wilkinson, H. R., Porayska-Pomsta, K., Farran, E. K., Tolmie, A. K., & Mareschal, D. (2025). SEE+ computerized classroom-based training enhances 7- to 10-year-olds' socio-emotional cognition through observation and inference. *PLOS ONE*, 20(9 September). <https://doi.org/10.1371/journal.pone.0330934>
- Moreira, S., Coelho, V., Peixoto, C., Espain, A., & Azevedo, H. (2025). A universal socio-emotional learning programme for child adjustment and peer relationships. *International Journal of Emotional Education*, 17(1), 111–115. <https://doi.org/10.56300/OYEX7238>
- Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, 11(2). <https://doi.org/10.1177/1094428106291059>
- Mosley, J., & Niwano, Z. (2012). Promoting social, emotional and behavioural skills through circle time and circles of support. In *The Routledge International Companion to Emotional and Behavioural Difficulties* (pp. 272–279). Taylor and Francis. <https://doi.org/10.4324/9780203117378>
- Nur, L., Stephani, M. R., Kusdinar, Y., Hidayat, T., Purwati, Wijaya, A., Dadi, Suhendra, D. I., Malik, A. A., & Nilan, F. (2025). The effectiveness of the Sport Education Model with local wisdom in shaping social attitudes among elementary school students. *International Journal of Human Movement and Sports Sciences*, 13(4), 901–908. <https://doi.org/10.13189/saj.2025.130426>
- Oberle, E., Domitrovich, C. E., Meyers, D. C., & Weissberg, R. P. (2016). Establishing systemic social and emotional learning approaches in schools: A framework for schoolwide implementation. *Cambridge Journal of Education*, 46(3), 277–297. <https://doi.org/10.1080/0305764X.2015.1125450>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E.

- A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLoS Medicine*, 18(3). <https://doi.org/10.1371/JOURNAL.PMED.1003583>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2022). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Revista Panamericana de Salud Publica/Pan American Journal of Public Health*, 46. <https://doi.org/10.26633/RPSP.2022.112>
- Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., Shamseer, L., Tetzlaff, J., Akl, E., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2024). PRISMA 2020 statement: An updated guideline for publishing systematic reviews. *Germinare*, 4, 1 – 19. <https://doi.org/10.5281/zenodo.13271469>
- Resurrección, D. M., Jiménez, Ó., Menor, E., & Ruiz-Aranda, D. (2021). The Learning to Be Project: An intervention for Spanish students in primary education. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.632617>
- Rojas, N. M., & Abenavoli, R. M. (2021). Bidirectionality in behavioral regulation, emotional competence, and expressive vocabulary skills: Moderation by preschool social-emotional programs. *Journal of Applied Developmental Psychology*, 76. <https://doi.org/10.1016/j.appdev.2021.101316>
- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. SAGE Publications, Inc. <https://doi.org/10.4135/9781412984997>
- Salovey, P., & Mayer, J. D. (1990). Emotional Intelligence. *Imagination, Cognition and Personality*, 9(3). <https://doi.org/10.2190/DUGG-P24E-52WK-6CDG>
- Sancassiani, F., Pintus, E., Holte, A., Paulus, P., Moro, M. F., Cossu, G., Angermeyer, M. C., Carta, M. G., & Lindert, J. (2015). Enhancing the emotional and social skills of the youth to promote their wellbeing and positive development: A systematic review of universal school-based randomized controlled trials. *Clinical Practice and Epidemiology in Mental Health*, 11, 21–40. <https://doi.org/10.2174/1745017901511010021>
- Santamaría-Villar, M. B., Gilar-Corbi, R., Pozo-Rico, T., & Castejón, J. L. (2021). Teaching socio-emotional competencies among primary school students: Improving conflict resolution and promoting democratic co-existence in schools. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.659348>
- Sklad, M., Diekstra, R., Ritter, M. D., Ben, J., & Gravesteyn, C. (2012). Effectiveness of school-based universal social, emotional, and behavioral programs: Do they enhance students' development in the area of skill, behavior, and adjustment? *Psychology in the Schools*, 49(9), 892–909. <https://doi.org/10.1002/pits.21641>
- Sterne, J. A., Hernán, M. A., Reeves, B. C., Savoviæ, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A. W., Churchill, R., Deeks,

- J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., ... Higgins, J. P. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ (Online)*, 355. <https://doi.org/10.1136/bmj.i4919>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. [https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7)
- Swisher, J. L., & Doyle, L. B. (2024). The CASEL framework and Christian schools: Using a hermeneutical tool to determine worldview alignment. *Journal of Research on Christian Education*, 33(1–2), 40–57. <https://doi.org/10.1080/10656219.2024.2345634>
- 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>
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1), 55–79. <https://doi.org/10.1002/jrsm.1164>
- Durlak, J. A., Domitrovich, C. E., Weissberg, R. P., & Gullotta, T. P. (2015). Handbook of social and emotional learning: Research and practice. *Guilford Publications*.
- Wigelsworth, M., Lendrum, A., Oldfield, J., Scott, A., ten Bokkel, I., Tate, K., & Emery, C. (2016). The impact of trial stage, developer involvement and international transferability on universal social and emotional learning programme outcomes: A meta-analysis. *Cambridge Journal of Education*, 46(3), 347–376. <https://doi.org/10.1080/0305764X.2016.1195791>