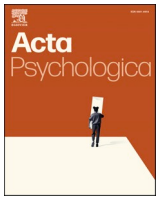
Acta Psychologica 271 (2026) 108128



Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



How effective is Waldorf education? A systematic review of empirical research

Mátyás Turós

Eszterházy Károly Catholic University, Eszterházy tér 1, H-3300, Eger, Hungary

ARTICLE INFO

Keywords:
Steiner education
Systematic review
Socioeconomic status
Selection bias
Academic achievement

ABSTRACT

Waldorf education operates in more than 3000 schools and kindergartens, yet no synthesis has examined academic, socio-emotional and creative outcomes together across compulsory schooling. Following PRISMA 2020, we searched six databases without language or temporal restriction. Of 1521 records screened, 38 reports of 37 studies from 19 countries were included; 35 contributed coded results. Risk of bias was appraised with ROBINS-I, NOS, RoB 2 and CASP, and certainty with GRADE. Each study contributed two classifications per domain, one based on significance and one on the direction of effect; conventional-school contrasts formed the primary synthesis. Academic achievement showed no consistent direction (3 of 18 studies Waldorf-favouring, 5 comparator-favouring, 5 mixed; direction of effect 6 vs 10). Comparator-favouring academic findings were more frequent among studies adjusting for socioeconomic status (SES; 3 of 7) than among others (2 of 11), and Waldorf-favouring findings were infrequent in both (1 of 7, 2 of 11), a descriptive between-study pattern; within studies, adjustment did not change findings uniformly. Socio-emotional findings favoured Waldorf in 8 of 10 studies and creativity findings in 5 of 10, none favouring the comparator, but only two socio-emotional studies and no creativity study adjusted for SES. Motivation showed no clear pattern. Certainty was very low in all domains: 34 of 37 studies carried a high risk of bias and 28 applied no statistical SES control. The evidence documents descriptive differences between school groups but does not establish causal effects; the academic findings are compatible with residual socioeconomic confounding of unknown magnitude.

1. Introduction

1.1. Theoretical framework: holistic education and intrinsic motivation

Several lines of research are relevant to the "whole-child" aims of holistic education, although none of them concerns a specific educational model. Some findings suggest that learning is closely bound to social and emotional experience, and that learning divorced from its emotional context may be less effective (Immordino-Yang et al., 2019). Social-emotional learning programmes have been associated with improved academic performance (Shi & Cheung, 2024) and with sustained achievement (Martinsone et al., 2025), and comparative studies indicate that academic achievement and subjective well-being can be mutually reinforcing (Su & Lee, 2023). Self-determination theory (SDT), supported by meta-analytic, review and intervention evidence, offers an account of sustained engagement based on autonomy support (Howard & Slemp, 2025; Matos et al., 2025; Zixia et al., 2025). These findings concern general educational principles; they do not, in themselves, constitute evidence for the effectiveness of Waldorf pedagogy, which must be assessed on the basis of studies of Waldorf settings.

One of the most widely implemented, yet highly debated, pedagogical approaches that draws on such holistic principles is the Waldorf method (frequently referred to as Steiner education or Steiner Waldorf education in the British and Scandinavian scholarly literature; hereafter the two terms are treated as synonymous, and the terminology of the original authors is followed when describing specific studies), which treats cognitive, emotional, and social development as a unified process, aligning the curriculum with the child's developmental rhythm. Specific elements of the method, such as developmental readiness and delayed reading instruction, have been subjected to empirical investigation.

Suggate's (2010) meta-analysis indicates that while early phonics instruction is effective through the first grade, a rich linguistic background and reading comprehension subsequently become the primary drivers of reading development. These findings concern the timing and type of reading interventions rather than the age at which formal instruction begins, and do not in themselves support a later start. They have nonetheless been read as consistent with the Waldorf strategy of building on oral foundations (storytelling, vocabulary), on which

E-mail address: turos.matyas@uni-eszterhazy.hu.

https://doi.org/10.1016/j.actpsy.2026.108128
Received 12 April 2026; Received in revised form 3 October 2026; Accepted 6 November 2026
Available online 11 November 2026
0001-6918/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

account Waldorf students who begin formal reading instruction later would be expected to reach the reading level of their peers within a few years; whether such convergence occurs is an empirical question examined in Section 3.3.6. The data also warrant caution. Lovett et al. (2017) demonstrated that for children at risk of reading difficulties, the first grade is a critical intervention window in which intensive and explicit phonological instruction yields the best results. The slower, indirect approach of the Waldorf curriculum may therefore disadvantage students at risk of dyslexia unless supplemented by targeted support (Lovett et al., 2017).

This systematic review treats Waldorf education as a test case for the developmental readiness hypothesis, which posits that delaying formal instruction in favour of age-appropriate, play-based activities leads to superior long-term cognitive and motivational outcomes. Because families choose Waldorf schools and differ systematically from other families, the review also examines selection: by synthesising studies with and without adjustment for socioeconomic status (SES), it asks whether reported Waldorf/non-Waldorf differences vary according to such adjustment. Statistical adjustment for measured family characteristics cannot isolate a pedagogical effect, and the review does not attempt to do so.

1.2. The evidence gap in Waldorf research

Waldorf education has a history of over a century and is currently implemented in 1281 schools and 1968 kindergartens worldwide (Waldorf World List, 2025, p. 12); however, the systematic synthesis of empirical evidence remains incomplete and methodologically fragmented. While the effectiveness of the Montessori method has been summarised by systematic reviews and meta-analyses (Demangeon et al., 2023; Randolph et al., 2023), comprehensive syntheses for Waldorf pedagogy are either lacking or restricted to narrow sub-fields. Guerrero et al. (2024) identified twice as much relevant empirical data pertaining to Montessori schools as to Waldorf schools. Previous attempts to synthesise Waldorf research are hampered by three limitations.

First, most comprehensive works (Attfield, 2024; Goldshmidt, 2017; Rawson, 2024; Stehlik, 2019) are theoretical or narrative and do not follow a systematic search protocol. Although Tyson (2024) identified nearly 270 studies, his work is a structured survey rather than a systematic review, since it uses neither a pre-registered protocol nor a formal appraisal of risk of bias. Second, existing systematic reviews focus on a single theoretical framework (Mavrelos & Daradoumis, 2020) or on specific sub-topics (Deluca et al., 2019; Tsortanidou et al., 2022), and Randoll and Peters' (2015) review of nearly 100 studies from the German-speaking region is narrative, without formal quality criteria. Third, the methodologically rigorous review by Guerrero et al. (2024) covered several types of alternative schools, included only eight Waldorf studies and was restricted to English and Spanish; the German-language empirical tradition therefore remained open for a Waldorf-focused synthesis.

To date, no PRISMA-based synthesis has examined academic achievement, socio-emotional development, and creativity in Waldorf education together across compulsory schooling (Table 1).

The present review addresses these gaps. It (a) integrates English- and German-language scholarship; (b) applies PRISMA 2020 with a critical appraisal of the included studies (ROBINS-I, RoB 2, NOS, CASP) and a rating of the certainty of evidence (GRADE); (c) pays particular attention to confounding by socioeconomic background, which Dahlin (2017) identified as one of the field's greatest weaknesses, reporting which SES indicators each study adjusted for and comparing unadjusted and adjusted estimates where both are available; and (d) describes which Waldorf-specific elements (such as delayed reading instruction, arts integration, and the epoch system) have been examined and at which ages.

The research questions are as follows:

Q1: What levels of performance do Waldorf students achieve on conventional standardised academic measures (e.g., PISA, national competence assessments) and on domain-specific academic and developmental measures, compared with non-Waldorf control groups?

Q2: What differences are observable in socio-emotional development, creativity, and motivation?

Q3: Which specific elements of Waldorf pedagogy have been examined in the empirical literature, and which show a plausible association with the observed outcomes?

Q4: How does achievement evolve over the course of the school career (early vs. late phases)?

Q5: What methodological challenges and risks of bias characterise Waldorf research, to what extent do these studies adjust for socioeconomic background, and do reported differences vary according to such adjustment?

2. Methods

2.1. Protocol and registration

The protocol was prospectively registered in PROSPERO (CRD420261288126). During implementation, the search was expanded from the four databases originally planned to six (adding APA PsycINFO and FIS Bildung) to capture psychological outcomes and German-language evidence. The review follows the PRISMA 2020 guidelines (Page et al., 2021). Departures from the registered protocol are declared where they occur (Sections 2.2, 2.5, 2.7 and 2.8).

Table 1
Comparison of the present systematic review with previous overview works.

Criterion	Randoll and Peters (2015)	Dahlin (2017)	Tyson (2024)	Guerrero et al. (2024)	Present study (2026)
Methodology	Narrative	Narrative	Narrative	Systematic (PRISMA 2020)	Systematic (PRISMA 2020)
Number of databases	Not specified	Limited	Mixed	4	6
Language coverage	DE (minimal EN, SV)	DE, EN, SV	EN, DE, Nordic	EN, ES	Unrestricted
Temporal restriction	1990–2015	Unrestricted	1990–2024	2010–2022	Unrestricted
Included studies	~100	~30	~270	8	37 (36 in synthesis; 35 with coded results)
Risk-of-bias assessment	None	None	None	None	ROBINS-I, RoB 2, NOS, CASP
Certainty of evidence	None	None	None	None	GRADE
Open Science	None	None	None	None	Yes
Dual independent screening	None	None	None	Yes	Yes
SES control	None	None	None	Partial	Primary focus: adjustment sets recorded per study; unadjusted vs adjusted estimates compared

2.2. Eligibility criteria

The research questions were operationalised using the PICO framework: (a) Population: students aged 3–18 years (from pre-school through upper secondary education); (b) Intervention: Waldorf/Steiner educational settings (any geographical location); (c) Comparator: non-Waldorf educational settings (state schools or other alternative pedagogies); studies without a concurrent comparator were eligible only where within-group longitudinal change served as the contrast; (d) Outcomes: academic achievement, socio-emotional development, creativity, motivation, and long-term life-course outcomes.

No language or temporal restrictions were applied. For publications in languages other than English or German, AI-based translation (Claude) was used for title and abstract screening and, where necessary, full-text assessment, validated by a three-tier protocol (specialist terms cross-referenced against international Waldorf terminology; native-speaker or expert consultation where eligibility depended on nuanced interpretation; external validation of all potentially eligible non-English/German full texts).

Journal articles and peer-reviewed book chapters were eligible; monographs, dissertations, conference proceedings, editorials and institutional reports were excluded, although findings from dissertation research subsequently published as a peer-reviewed chapter were eligible under that source type. Only empirical studies reporting primary data entered the synthesis. Inclusion required a Waldorf/Steiner context, a relevant outcome, a clear description of methods, a comparative or longitudinal design (quantitative or qualitative), and participants aged 3–18 years. Exclusion criteria were: works without empirical data; general introductions to anthroposophy; publications in which the Waldorf programme was not clearly defined; editorials, letters and reviews without primary data; and results that could not be attributed to Waldorf students specifically (for example, results reported only for a pooled group of alternative schools whose Waldorf share was not reported). Where a study reported both pooled and Waldorf-specific results, only the latter were extracted. A pooled result with a reported Waldorf share was retained and excluded in a sensitivity analysis (Bilde et al., 2013), a rule not specified in the registered protocol. Duplicate reports of the same sample were merged into a single study unit rather than excluded. Samples of final-year secondary students were eligible even where some participants were 18 or 19 years old, because the population criterion refers to the period of schooling rather than to an exact age limit.

Cross-sectional retrospective studies of adults were excluded because of recall bias, non-response bias and the difficulty of linking adult outcomes to schooling (Guyatt et al., 2011); prospective longitudinal studies following participants from school age into adulthood were eligible. Parent- and teacher-proxy reports were eligible when they referred to the child's current state and were treated as carrying an elevated risk of subjective bias.

2.3. Information sources and search strategy

Six databases were searched: Scopus, Web of Science Core Collection, ERIC and Education Source (19 January 2026), FIS Bildung (20 January 2026) and APA PsycINFO (21 January 2026). In FIS Bildung the bilingual query extended to all source bases of the portal (German Education Index, BASE, Online Contents, Library of Congress collections). The query covered the designations under which Waldorf education appears (e.g., Waldorf, Steiner) and was not combined with outcome terms, to avoid excluding relevant studies; because “Steiner” retrieved many non-pedagogical documents (e.g., Wiedemann–Steiner syndrome, Steiner tree algorithm), results were restricted with subject-category filters. The search identified 1955 records (Supplementary Table S1); the search strings are available in the study repository (see Data Availability Statement).

2.4. Study selection

After removal of 434 duplicates in Rayyan, 1521 records were screened. Of these, 361 publications of Erziehungskunst, the magazine of the Association of Free Waldorf Schools in Germany, which publishes essays and practical reports without independent peer review, were batch-excluded as incompatible with the PICO criteria by format. Two reviewers independently screened the remaining 1160 records, blinded to each other's decisions; agreement was excellent (Gwet's AC1 = 0.863, 95% CI: 0.841–0.886; Byrt et al., 1993; Gwet, 2008). Discordant items were resolved by consensus, and grey literature was excluded at this stage. Seventy-five reports proceeded to full-text assessment, again by two independent reviewers with consensus resolution. Four reports that passed full-text screening proved at data extraction not to meet all PICO elements (the Waldorf school served only as a recruitment site, or no Waldorf-specific outcome was measured). Thirty-eight reports of 37 studies were included; one study retained in the review was excluded from the synthesis because of a critical risk of bias (Section 2.5), and one is described narratively because its comparator pooled traditional and democratic schools that could not be separated (Dorfman & Fortus, 2019). Details are given in Supplementary Text S1.3 and Fig. 1; the screening records are available in the study repository.

2.5. Quality assessment (risk of bias)

Risk of bias was appraised with tools appropriate for each design: (a) RoB 2 for randomised and cluster-randomised designs, using its variant for cluster-randomised trials where allocation was by cluster (Sterne et al., 2019); (b) ROBINS-I for non-randomised studies of interventions (Sterne et al., 2016); (c) the Newcastle–Ottawa Scale (NOS) for observational comparisons (Wells et al., 2000); and (d) the CASP Qualitative Checklist for qualitative studies (Critical Appraisal Skills Programme, 2018). Where a study could fall under more than one rule, the design as described by its authors governed: a study reporting naturalistic observation and interpretive analysis without inferential statistics was appraised with CASP; among cross-sectional designs, ROBINS-I was applied where the authors framed Waldorf attendance as an exposure whose effect was being estimated, and the NOS where the study was presented as a descriptive group comparison. For ROBINS-I, the effect of interest was the effect of attending a Waldorf school; domain D2 was judged on selection after the start of exposure (parental self-selection into school type being assessed under D1), and domain D4 on co-interventions and implementation. All NOS-appraised studies were cross-sectional, and items S4 and O2 were omitted (maximum 7 stars). Unit-of-analysis errors in cluster-allocated designs were recorded separately and considered under imprecision in GRADE. Decision rules for individual domains are given in Supplementary Text S1.2.

To summarise the appraisals across tools, the original categories were mapped onto a three-level scale (Table 2). Because the tools assess partly different dimensions and their categories are not interchangeable, this harmonisation is a pragmatic, descriptive summary, not an equivalent metric of methodological quality across designs; tool-specific judgements are reported alongside it (Supplementary Table S2). Two conversion rules depart from the registered protocol. First, a NOS-appraised study without a comparability star (C1 = 0) was rated High irrespective of its total score, which aligns the NOS with ROBINS-I, where the absence of adequate control for confounding yields a Serious judgement in D1. A comparability star required SES or baseline achievement to be addressed in a way that equates the groups (Supplementary Text S1.2). Second, a ROBINS-I judgement of Critical was kept as a separate category, and, following ROBINS-I guidance that such results are too problematic to provide useful evidence, the one study so rated was retained in the review but excluded from the synthesis. SES control is recorded and analysed separately from risk of bias throughout, and the two are not treated as interchangeable.

The appraisal tools primarily assess confounder control and

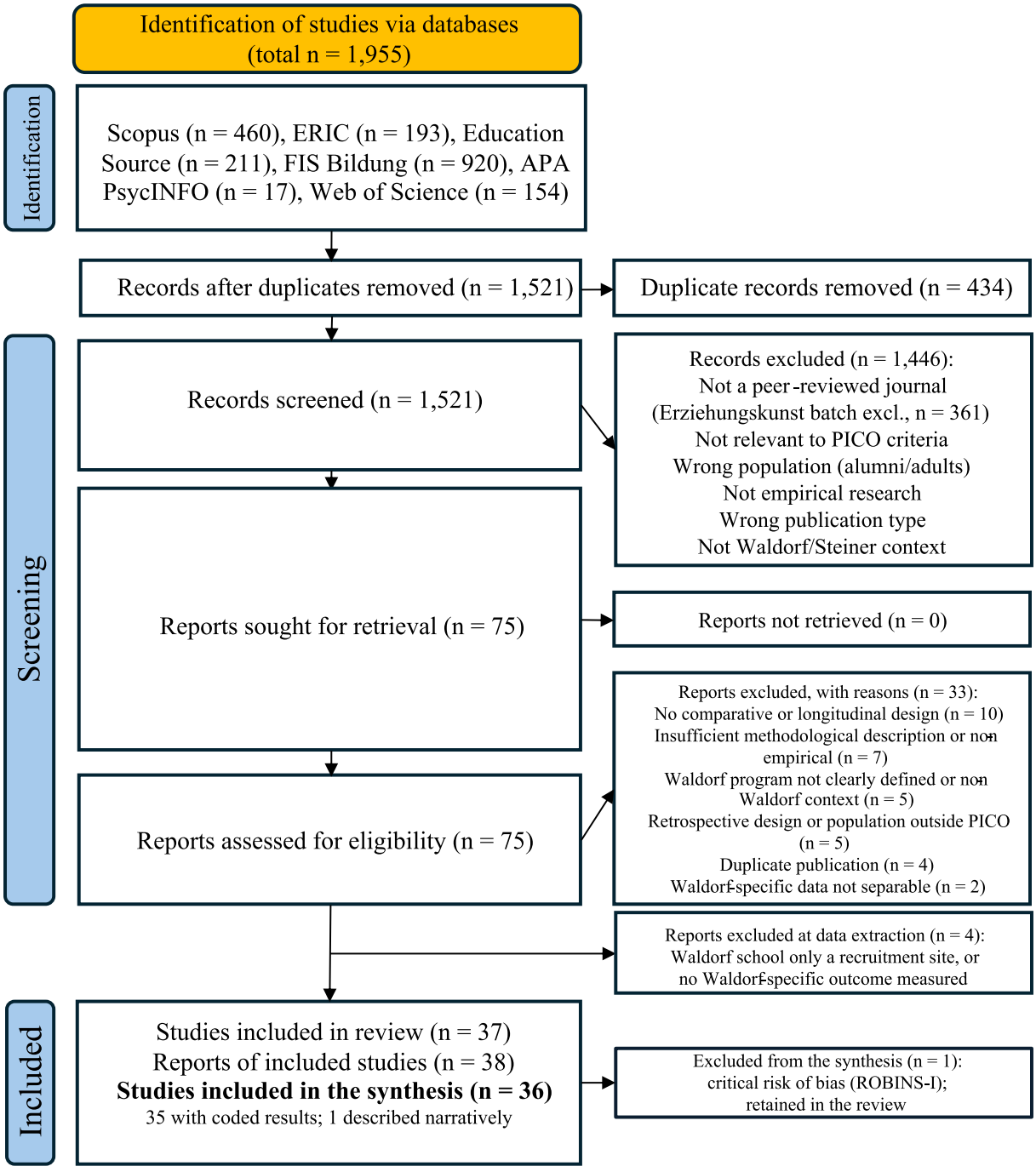


Fig. 1. PRISMA 2020 flow diagram of the study selection process.

statistical rigour; they do not directly evaluate the construct and ecological validity of the measures used. Where relevant, the synthesis notes that a standardised measure such as PISA is confined to one time point and subject domain and does not describe the pedagogical process as a whole.

2.6. Data extraction

Data were extracted using a standardised form recording study identification and design, sample characteristics, funding sources and author affiliations, intervention and comparator features, outcome variables and instruments, the handling of confounders, and the source database (Supplementary Text S1.4). Data were extracted with two large language models (Claude Sonnet 4.6 and Gemini 3.1 Pro) under a structured protocol, checked against the full texts by a third model (Gemini 2.5 Flash) and then verified by the author and a second reviewer (Supplementary Text S1.1).

Waldorf exposure was defined by the school or kindergarten type reported in each study. Accreditation by national Waldorf associations was not assumed to establish exposure to Waldorf pedagogy, and implementation fidelity was not independently verified by the review and was not measured consistently across the included studies, although some reported implementation checks, such as teacher interviews and classroom observation or the screening of schools for implementation of the pedagogical programme. Implementation is likely to vary across schools, teachers, countries and periods, and seven study units

Table 2
Harmonisation of risk-of-bias categories (pragmatic, descriptive summary).

Tool	Tool-specific category	Harmonised rating	Study units (n)
RoB 2	Low risk/Some concerns/ High risk	Low/Moderate/High	0/0/1
ROBINS-I	Low/Moderate	Low/Moderate	0/1
ROBINS-I	Serious/No information	High	23/0
ROBINS-I	Critical	Critical (excluded from synthesis)	1
NOS	Good (7★)	Low	0
NOS	Fair (4–6★), comparability star (C1 ≥ 1★)	Moderate	1
NOS	Fair (4–6★), no comparability star (C1 = 0)	High	4
NOS	Poor (0–3★)	High	6
CASP	Qualitative appraisal	Not harmonised	1 report
Total		Low 0/Moderate 2/High 34/Critical 1	37

Note. The tools assess partly different dimensions, and the harmonised scale is a pragmatic descriptive summary rather than an equivalent metric of quality across designs. All NOS-appraised studies were cross-sectional; items S4 and O2 were omitted, giving a maximum of 7 stars. A NOS-appraised study without a comparability star (C1 = 0) is rated High irrespective of its total score, in line with ROBINS-I domain D1. The CASP-appraised report (Kodsi, 2022) belongs to a study unit whose rating derives from the NOS appraisal of its 2021 report. Tool-specific judgements are given in Supplementary Table S2.

concerned Waldorf-oriented classes or Waldorf-inspired schools within the public system, a short Waldorf-inspired intervention, a pooled Waldorf and Freinet group, or a Steiner subgroup within a wider sample, rather than accredited Waldorf schools. Variation in implementation is therefore treated as a source of heterogeneity (Section 5).

Each outcome was extracted as a separate entry with its estimate, test statistic and significance. Standardised mean differences were calculated only from simple group comparisons; adjusted parameters (β, B, path coefficients) were recorded separately, so that unadjusted and adjusted differences were not conflated. For studies with statistical SES control, the fully adjusted estimate was the primary estimate, and the unadjusted estimate was recorded alongside it where reported. For every entry the numerical direction of the difference was recorded independently of its significance, or coded as not reported where the study gave no group values; which direction counts as favourable followed the original authors' interpretation of the specific construct (e.g., lower anxiety, burnout or competitive behaviour), or otherwise the established theoretical meaning of the construct; outcomes without evaluative meaning (e.g., drawing size) were not coded. Of the 130 non-significant entries with a conventional comparator, the numerical direction was known for 118. Where a reported p value referred to an omnibus test of more than two groups, it was not taken as evidence of a significant pairwise Waldorf–comparator difference; where no test was reported, the entry was coded Uncertain. Internally inconsistent source statistics were recorded as reported and flagged. Entry-level codes were assigned by the author and checked against the source texts by a second reviewer; discrepancies were resolved by consensus. This was a verification of the author's codes rather than independent double coding. Study-level classifications were then derived from the entry-level codes by the explicit rules set out in Section 2.7, implemented as formulas in the deposited database, so that this step involved no further judgement. No agreement statistic was computed for the coding stage.

All reported Waldorf–comparator contrasts within the four outcome domains were extracted; decision rules for comparator selection, age-specific results, multi-scale instruments, pooled groups, separately analysed Waldorf schools and additional cohorts are given in Supplementary Text S1.2. The qualitative Kodsi (2022) report was not entered in the coded synthesis, nor was Dorfman and Fortus (2019), whose comparator pooled traditional and democratic schools and which reported no Waldorf–comparator test; it is described narratively.

2.7. Data synthesis

A meta-analysis was not feasible because comparable effect estimates with sampling variances were too sparse. When the coded outcomes were grouped into conceptually comparable clusters within each domain, no cluster contained more than three studies reporting data from which a standardised effect size and its variance could be derived, and the studies within each cluster differed in instrument, age, design and type of estimate (Supplementary Table S10). As a narrower descriptive indicator, 17 of the 35 studies with coded results reported group means, standard deviations and sample sizes, or a regression coefficient with its standard error, for at least one outcome (16 for an outcome entered in the coded synthesis). We therefore performed a narrative synthesis following the Synthesis Without Meta-analysis (SWiM) guidelines (Campbell et al., 2020). Outcomes were grouped into four domains: (a) academic achievement (including reading, science and long-term educational attainment, the only life-course outcome available; Schindler Rangvid, 2008); (b) socio-emotional development (including well-being); (c) creativity (including motor and artistic development); and (d) motivation (including self-concept).

The unit of analysis is the study, not the outcome entry; each study contributes one classification per domain it measured, so that studies reporting many correlated outcomes do not carry greater weight. Two complementary classifications are reported. The significance-based classification assigns Mixed where statistically significant results in opposing directions occur within the domain; otherwise the single direction that reached significance (W+ or C+); ns where no result reached significance; and Uncertain where direction or significance was not reported. The direction-of-effect classification assigns the direction held by more than half of a study's entries whose numerical direction is known, irrespective of significance, and a two-sided exact sign test compares the number of Waldorf-favouring and comparator-favouring studies in each domain (McKenzie & Brennan, 2019). Because overlapping samples violate the independence assumption of the test, the result restricted to independent samples is treated as the primary inferential result and the distribution across all studies as descriptive; no test is applied to the significance-based counts. This reduces the dependence of the synthesis on the statistical power of individual studies.

The primary synthesis uses contrasts with conventional comparators. Contrasts with an alternative pedagogy address a different question, how Waldorf schools compare with other child-centred school models, and are synthesised separately (Section 3.3.8; Supplementary Table S9); this separation was not specified in the registered protocol.

Because the significance-based rule can classify a study on one significant result among several, and Mixed classifications become more likely with more analyses, a majority rule (the code held by more than half of a study's classifiable entries) is reported as a sensitivity analysis, together with analyses restricted to independent samples and excluding pooled Waldorf groups (Supplementary Table S5). Both classifications give equal weight to studies of very different size ($n = 28$ to $n = 459{,}084$) and are therefore descriptive. A non-significant result denotes the absence of a statistically significant difference, not the equivalence of the groups, since no included study tested equivalence. The move from entry-level to study-level counting is a departure from the registered protocol, which specified entry-level vote counting; entry-level counts are reported in Supplementary Table S3.

Two Austrian reports analyse the same PISA 2015 census of 149 Waldorf pupils (Salchegger et al., 2021; Wallner-Paschon, 2018), whereas Wallner-Paschon and Wiesinger (2021) analyse the PISA 2018 cohort. Rose and Jolley (2019) is probably a subsample of Rose and Jolley (2016), and Kikas (1998) and Kikas (2000) probably share their Waldorf participants. Overlapping units were retained where they

report different analyses or outcomes; a sensitivity analysis excludes one member of each overlapping pair (Supplementary Tables S5 and S8).

The role of socioeconomic background was examined in two ways, which answer different questions. First, studies were stratified by whether they adjusted statistically for SES. The method, level and indicators of SES handling were recorded for each study (Supplementary Table S6), and studies were classified as Full (SES and baseline achievement, or propensity score matching), Stat. (SES covariate without baseline achievement), Partial (design-level control or pre-test equivalence only) or No; a study-domain was assigned to the adjusted stratum only where its coded estimates were SES-adjusted. The comparison between strata is a descriptive between-study pattern: the strata differ in country, design, age, outcomes, comparators and sample size, SES was not tested as a moderator, and differences between strata are not an estimate of SES attenuation. Second, where a study reported the same contrast at two levels of adjustment (unadjusted and SES-adjusted, or in one study a less- and a more-adjusted model; Suggate et al., 2011), the two estimates were compared directly and classified as persisting, strengthening, attenuating, eliminated, emerging, reversed or unchanged (Supplementary Table S7). All percentages are reported with their numerators and denominators.

Age and comparator type were examined narratively as sources of heterogeneity; geographic region and length of exposure were reported too inconsistently for subgroup analysis. To address Q3, we focused on studies that explicitly analysed particular pedagogical elements. Publication bias was assessed qualitatively from the distribution of findings, sample sizes and the exclusion of grey literature. Funding and institutional affiliation were appraised separately as source bias and were not treated as indicators of publication bias.

2.8. Certainty of evidence

Certainty was appraised for each domain with the GRADE approach adapted for narrative synthesis (Murad et al., 2017), considering the starting level, risk of bias, inconsistency, indirectness, imprecision and publication bias, and rated on the four-level GRADE scale. Study design and imprecision were rated as separate domains. Non-randomised evidence appraised with ROBINS-I starts at high certainty and is rated down for risk of bias (Schünemann et al., 2019); NOS-appraised studies were considered within the risk-of-bias judgement; ten of the 11 were at high risk of bias, and starting the non-randomised body at low certainty instead would not have changed the rating in any domain. Starting the non-randomised evidence at high certainty departs from the registered protocol. Randomised and non-randomised evidence were rated separately: the single cluster-randomised trial with one cluster per arm (Arslan & Kartal, 2022) was rated in its own row, with its unit-of-analysis error considered under imprecision, and excluded from the non-randomised academic body. Source-related concerns (funding, author affiliation, framing) were reported descriptively and did not contribute to downgrading. The Summary of Findings is given in Table 5 and the domain-level justification in Supplementary Table S4.

3. Results

3.1. Characteristics of included studies

The review included 38 reports of 37 studies published between 1998 and 2024 from 19 countries, most frequently the United Kingdom ($n = 7$), followed by Austria, Russia and Turkey ($n = 3$ each), and Estonia, France, Hungary, Israel, Ukraine and the USA ($n = 2$ each); nine countries contributed one study each. Of the 38 reports, 27 were in English, four in Russian, three in German, and one each in Arabic, Hungarian, Norwegian and Turkish; these 11 non-English reports (29%) would have been lost to an English-only search. Reports were identified mainly through Scopus ($n = 15$) and ERIC ($n = 11$), supplemented by APA PsycINFO ($n = 6$), FIS Bildung ($n = 4$) and Web of Science ($n = 2$).

By study unit, 25 designs were cross-sectional comparisons, four quasi-experimental, four longitudinal cohorts, two mixed longitudinal–cross-sectional, one longitudinal quasi-experimental and one cluster-randomised with one cluster per arm. Twenty-eight studies compared Waldorf students with conventional schools only; nine also included an alternative-pedagogy comparator (Montessori in six, Elkonin–Davydov developmental education in one, a pooled Montessori and "New schools" group in one, and democratic schools not separated from traditional schools in one). Eight of these contribute coded contrasts to the separate synthesis (Section 3.3.8). No study compared Waldorf exclusively with another alternative pedagogy. Sample sizes ranged from $n = 28$ to $n = 459{,}084$, and participants ranged from 48 months of age to the final year of secondary school (some participants were 18 or 19 years old), with one register study following pupils into adulthood. Study characteristics are summarised in Supplementary Table S2 and study-level findings in the evidence map (Fig. 2).

3.2. Methodological quality of studies

Under the harmonised summary, 34 of the 37 studies (92%) were at high risk of bias, two at moderate risk and one at critical risk; none was at low risk. Among the 25 studies appraised with ROBINS-I, confounding (D1) was the main weakness (23 Serious, two Moderate; Moderate required statistical control of SES together with baseline achievement, or propensity score matching). Selection after the start of exposure (D2) was Moderate in 18 studies, mostly samples of pupils already enrolled, and Serious in five; deviations from the intended intervention (D4) could not be judged in 19; and outcome measurement (D6) was Serious in three (a pretence task scored live by an unblinded experimenter, non-validated instruments, and register data that do not record Waldorf Grades 11–12). Ionova et al. (2018) was rated Critical (no test statistics or raw data) and excluded from the synthesis. The 11 NOS-appraised studies scored one to five of seven stars (five Fair, six Poor); ten received no comparability star and are rated High, and one (Rose & Jolley, 2019) Moderate. The cluster-randomised study of Arslan and Kartal (2022) was rated High with the cluster-randomised variant of RoB 2: allocation was by one cluster per arm, so that the intervention is confounded with group and teacher, sequence generation and allocation concealment were not described, and no analysis plan was pre-registered; its reported t-values also imply that each child was counted twice. The qualitative Kodsi (2022) report was appraised with CASP and is not harmonised.

3.2.1. Adjustment for socioeconomic background

Nine studies (24%) adjusted statistically for SES: two with SES and baseline achievement or propensity score matching (Bilde et al., 2013; Salchegger et al., 2021) and seven with SES covariates only. Adjustment ranged from extensive individual register data (Schindler Rangvid, 2008) and composite SES indices to a single school-level indicator (Telfer-Radzat, 2024) and indicators entered one at a time for one outcome (van Schie & Vedder, 2023; Supplementary Table S6). Eight studies addressed SES only at the design stage, for example by selecting schools in areas of similar geodemographic profile, one checked only baseline equivalence by pre-test (Arslan & Kartal, 2022), and 19 did not address it. Twenty-eight studies (76%) therefore applied no statistical SES control, which corroborates Dahlin's (2017) assessment of the field. No adjustment set could account for unmeasured selection factors such as parental educational values or child motivation.

3.2.2. Source bias

Thirty-four studies (92%) carried a low risk of source bias. Three were rated high (Abdulaeva, 2022; Dahlin, 2010; Diefenbach et al., 2020); two reported a Waldorf advantage, and one did not compare school types directly. Source bias is reported descriptively and did not contribute to GRADE downgrading. Statistical SES control was applied in four of the 21 studies published before 2019 and in five of the 16

Study	Comparator	SES control	Risk of bias	Sample overlap	Academic	Socio-emotional	Creativity	Motivation
Salchegger et al. (2021)	State	Full	Moderate	Confirmed	C+			W+
Bilde et al. (2013)	State	Full	High	–				C+
Solhaug (2007)	State	Stat.	High	–	C+	W+		
Schindler Rangvid (2008)	State	Stat.	High	–	C+			
Suggate et al. (2011)	State	Stat.	High	–	ns			
Diefenbach et al. (2020)	State	Stat.	High	–	Uncertain			
Turós (2023)	State	Stat.	High	–		ns		
van Schie and Vedder (2023)	State	Stat.	High	–	W+			ns
Telfer-Radzat (2024)	State	Stat.	High	–	Mixed			
Rose and Jolley (2019)	State	Partial	Moderate	Probable			ns	
Tsukerman (2005)	State+Alt	Partial	High	–	ns			
Cunningham and Carroll (2011)	State	Partial	High	–	Mixed			
Rose et al. (2012)	State+Alt	Partial	High	Possible			Mixed	
Larrison et al. (2012)	State	Partial	High	–	Mixed			
Rose and Jolley (2016)	State	Partial	High	Probable			Mixed	
Kodsi; Kodsi (2021, 2022)	State	Partial	High	Merged		Not coded	W+	
Arslan and Kartal (2022)	State	Partial	High	–	W+			
Korkmaz and Öztürk Samur (2022)	State	Partial	High	–	ns		W+	
Kikas (1998)	State	No	High	Probable	C+			
(Cox & Rowlands)	State+Alt	No	High	–			W+	
Kikas (2000)	State	No	High	Probable	Mixed			
Burkitt et al. (2005)	State	No	High	–			ns	
Yevsikova (2008)	State	No	High	–				C+
Shankland et al. (2009)	State+Alt	No	High	–		W+		
Dahlin (2010)	State	No	High	–		W+		
Mather (2012)	State+Alt	No	High	–	C+			
Besançon et al. (2015)	State	No	High	–		W+	ns	Mixed
Kirkham and Kidd (2017)	State+Alt	No	High	–		ns	W+	
Wallner-Paschon (2018)	State	No	High	Confirmed	Mixed			W+
Borodulkina (2019)	State	No	High	–			W+	
Dorfman and Fortus (2019)	State+Alt	No	High	–				Not coded
Wallner-Paschon and Wiesinger (2021)	State	No	High	–	W+			W+
Abdulaeva (2022)	State	No	High	–		W+		
Kvintova et al. (2022)	State+Alt	No	High	–		W+		W+
Kóródi et al. (2023)	State	No	High	–	Uncertain	W+		
Özçelik and Sapsağlam (2023)	State+Alt	No	High	–		W+		
Ionova et al. (2018) ‡	State	No	Critical	–	Excl.			Excl.

(caption on next column)

Fig. 2. Evidence map of included studies ($N = 37$; study-level direction by outcome domain).

Note. Cells give the significance-based study-level code per domain from contrasts with the conventional comparator; an empty cell means the domain was not measured. W+ = Waldorf advantage; ns = no significant difference; C+ = comparator advantage; Mixed = significant results in opposing directions; Uncertain = direction or significance not reported; Not coded = not entered in the coded synthesis (Kodsi, 2022: qualitative report; Dorfman & Fortus, 2019: comparator not separable). ‡ = retained in the review but excluded from the synthesis (ROBINS-I Critical). SES control: Full = SES and baseline, or PSM; Stat. = statistical SES covariate; Partial = design-level or baseline only; No = none. Comparator: State+Alt = also an alternative-pedagogy comparator, synthesised separately (Supplementary Table S9). Sample overlap: Confirmed = same sample as another included study; Probable/Possible = likely or possible overlap; Merged = two reports of one study. Sorted by SES control, risk of bias and year.

published from 2019 onwards; the subgroups are too small to indicate a trend. Fig. 2 summarises the study-level findings together with comparator type, SES control, harmonised risk of bias and sample overlap.

3.3. Synthesis of findings

Table 3 gives the study-level distribution by domain under both classifications. Direction of effect could be determined for 41 of the 46 study-domain classifications. The two classifications agree in direction: no consistent direction in academic achievement and motivation, and predominantly Waldorf-favouring findings in socio-emotional development and creativity. Under the majority rule the overall picture was similar, but the Waldorf-favouring pattern appeared less consistent (Waldorf- vs comparator-favouring studies: academic 3 vs 4; socio-emotional 5 vs 0, with two studies without a majority; creativity 5 vs 0, with five studies ns; motivation 2 vs 1). Restricting the analysis to independent samples or excluding the pooled Waldorf and Freinet group did not change the distribution materially (Supplementary Table S5).

3.3.1. Academic achievement

Of 18 studies, three were Waldorf-favouring, five comparator-favouring, five Mixed, three ns and two Uncertain; by direction of effect, six favoured Waldorf and ten the comparator (sign test $p = .45$; independent samples, 5 vs 9, $p = .42$).

3.3.1.1. Reading and literacy. Six studies examined reading or writing outcomes in settings with delayed formal instruction (Cunningham & Carroll, 2011; Larrison et al., 2012; Mather, 2012; Suggate et al., 2011; Telfer-Radzat, 2024; van Schie & Vedder, 2023); several report age- or grade-related differences compatible with later convergence, but not a common longitudinal catch-up trajectory (Section 3.3.6). Suggate et al. (2011) found 7-year-old Waldorf entrants behind age-matched state pupils in phonemic segmentation ($d = -1.52$) and nonword reading ($d = -2.09$) and ahead in narrative evaluations ($d = 0.77$) in unadjusted pairwise comparisons; in the authors' regressions, which also control maternal education, ethnicity, vocabulary and home literacy, the Waldorf indicator was not significant for any outcome. Cunningham and Carroll (2011), comparing older Steiner and younger standard-educated children over their first two years of formal instruction, reported an initial deficit in letter-sound knowledge and a deficit in spelling that remained at the end of the second year, alongside an early advantage in phonological awareness; word reading, reading accuracy and comprehension did not differ significantly. Van Schie and Vedder (2023) found a Waldorf advantage in English spelling at Grades 6 and 8, and Mather (2012) more left-sided mirror-writing among Waldorf pupils.

Table 3
Direction of findings by outcome domain (study level; 35 studies with coded results; conventional-comparator contrasts).

Outcome domain	Studies (n)	W+	ns	C+	Mixed	Uncertain	Direction of effect (W/C)	Sign test p (all/independent samples)
Academic achievement	18	3	3	5	5	2	6/10	0.45/0.42
Socio-emotional development	10	8	2	0	0	0	8/1	0.039/0.039
Creativity (incl. Motor and arts)	10	5	3	0	2	0	9/0	0.004/0.008
Motivation and self-concept	8	4	1	2	1	0	4/3	1.00/1.00

Note. Unit of analysis: the study; each study contributes one classification per domain it measured. W+ = at least one significant result favouring Waldorf and none favouring the comparator; C+ = the reverse; Mixed = significant results in opposite directions; ns = no significant result (non-significance, not equivalence); Uncertain = direction or significance not reported. For studies with statistical SES control, fully adjusted estimates are coded. Direction of effect: studies in which most entries with known numerical direction favour Waldorf (W) or the comparator (C), irrespective of significance; studies without a majority or with undeterminable direction are omitted. Sign test: exact two-sided binomial test of W vs C studies, for all studies and for independent samples only (excluding Wallner-Paschon, 2018; Rose & Jolley, 2019; Kikas, 2000); because overlapping samples violate the independence assumption of the test, the independent-samples result is the primary inferential result. No test is applied to the significance-based counts. The direction of effect conveys neither the size nor the precision of the differences. Counts give equal weight to studies of very different size and are descriptive.

3.3.1.2. Mathematics. Results tended to favour the control group, with a similar temporal structure. The Austrian Waldorf mean was reported as significantly below the national mean in PISA 2015 (489 vs 497; Wallner-Paschon, 2018) and described as similar to it in 2018, without a significance statement (Wallner-Paschon & Wiesinger, 2021). Telfer-Radzat (2024) found a significant disadvantage at Grade 3 ($B = -14.90$) and an advantage at Grade 8 ($B = 9.66$), in different cohorts. Arslan and Kartal (2022), the only study with random allocation, found higher early number skills after a four-week intervention combining Waldorf materials with cooperative learning; the child-level analyses did not account for clustering, and with one cluster per arm no valid cluster-adjusted significance test is possible.

3.3.1.3. Scientific understanding. Salchegger et al. (2021) and Wallner-Paschon (2018) analyse the same PISA 2015 sample of 149 Austrian Waldorf pupils. Unmatched, Waldorf pupils scored numerically higher in science (503 vs 495; $d = 0.09$; reported as significant by Wallner-Paschon, 2018, who treated the Waldorf census mean as having no sampling error). After propensity score matching, the difference reversed to a significant control-group advantage ($d = -0.32$; Salchegger et al., 2021), a within-sample comparison of unadjusted and adjusted estimates. In Kikas (1998), Waldorf discussion groups referred to scientific rules less often and relied more often on the naive distance theory of the seasons; in Kikas (2000), Waldorf pupils gave more descriptive, lower-level explanations and fewer memorised ones, and none gave a scientific explanation of the seasons.

3.3.1.4. Long-term and general cognitive outcomes. Schindler Rangvid (2008) found fewer years of education and a lower probability of completing upper secondary education among Waldorf students in Danish register data (adjusted differences −0.79 years and −0.24), but no adjusted difference in tertiary completion. Solhaug (2007) found lower political knowledge after adjustment in one of its two Steiner schools, and Kóródi et al. (2023) a lower GPA without a statistical test. Tsukerman's (2005) reported advantage in "learning to learn" rested on lenient scoring applied to the Waldorf group only and was not significant with standard scoring; Diefenbach et al. (2020) did not compare school abilities between school types directly; and Korkmaz and Öztürk Samur (2022) found no difference on Raven's matrices.

3.3.1.5. Summary pattern. No overall Waldorf advantage or disadvantage emerges in this domain. Where delayed instruction is involved, early deficits are followed in later grades by non-significant differences or Waldorf advantages, mostly in repeated cross-sectional data rather than in tracked cohorts (Section 3.3.6). The relation of these findings to socioeconomic adjustment is examined in Section 3.3.7.

3.3.2. Socio-emotional development

Of ten studies, eight were Waldorf-favouring and two ns, with none favouring the comparator; by direction of effect, eight favoured Waldorf and one the comparator (sign test $p = .039$, identical for independent samples). All ten studies carried a high risk of bias, outcomes were mostly unblinded self-reports or proxy ratings, and only two studies adjusted statistically for SES.

Kóródi et al. (2023) found lower burnout among Hungarian Waldorf students, and Besançon et al. (2015) higher general life satisfaction ($\eta^2 = 0.06$). Shankland et al. (2009) found more problem-focused coping among final-year Steiner students (OR = 4.73, $p = .02$). Özçelik and Sapsağlam (2023) found less anxiety-withdrawal in Waldorf than in state preschoolers ($p = .004$). Dahlin (2010) reported higher moral reasoning and sensitivity and a stronger sense of social and moral responsibility, without SES control and with a high risk of source bias. Kvintova et al. (2022) found better teacher-rated executive functioning among Waldorf preschoolers ($\eta^2 = 0.02$), though not on all scales, and Abdulaeva (2022) higher self-organisation in a productive activity and a rule game. Kirkham and Kidd (2017) found no difference from conventional pupils in pretence.

3.3.2.1. Studies with socioeconomic adjustment. Solhaug (2007), adjusting for books at home and parental education, found higher tolerance in both Steiner schools ($\beta = 0.12$ and 0.091) and higher civic engagement in one. Turós (2023), adjusting for parental education, finances, marital status and child-rearing attitudes, found no significant effect of Waldorf schooling on self-control or trustworthiness, and showed that Waldorf parents used a less restrictive parenting style, a direct indication of selection.

3.3.2.2. Summary pattern. The Waldorf-favouring direction is consistent across well-being, coping, executive functions and self-organisation, civic attitudes, moral reasoning and preschool anxiety-withdrawal. It rests mostly on unblinded subjective instruments in studies at high risk of bias, and the evidence is compatible with a contribution of family background but cannot establish or exclude it.

3.3.3. Creativity

Of ten studies, five were Waldorf-favouring, three ns and two Mixed, with none favouring the comparator; by direction of effect, nine favoured Waldorf and none the comparator (sign test $p = .004$; independent samples, 8 vs 0, $p = .008$). Samples were small (about 15–100 per group) and effect sizes heterogeneous.

Rose and Jolley (2016) found more realistic representational drawings among Waldorf students (mannequin $\eta^2 = 0.18$) and more expressive colour use in happy drawings at age 14, but less expressive composition at age 10 (Mixed). Cox and Rowlands (2000) reported an advantage in free, scene and observational drawing and in colour-use quality. Rose et al. (2012) found Waldorf advantages in expressive drawing at age 9 and representational drawing at age 7, but a disadvantage at age 5. Burkitt et al. (2005) found no difference in rated drawing skill; the larger figures drawn by conventional-school children

are treated by the authors as a convention, not quality, and were not coded. Korkmaz and Öztürk Samur (2022) found an advantage in overall creativity (three-group $\eta^2 = 0.09$), Kirkham and Kidd (2017) a TCT-DP advantage (three-group $\eta^2 = 0.17$), and Borodulkina (2019) more scenic and contextual drawings. Besançon et al. (2015) found no difference in divergent thinking, and Rose and Jolley (2019) none in the sources of drawing ideas. Kodsi (2021) found more complex constructive play in Waldorf kindergartens, in an analysis that treats constructions nested within children as independent.

3.3.3.1. Summary pattern. The Waldorf-favouring direction is concentrated in rated drawing quality and divergent thinking, while comparator-favouring results are age-specific within two studies. No creativity study adjusted statistically for SES, so the sensitivity of these findings to socioeconomic adjustment cannot be assessed, and residual confounding by parental cultural engagement is as plausible here as elsewhere. Because the Waldorf curriculum allocates substantial time to drawing, performance on drawing tasks may also reflect domain-specific practice rather than domain-general creativity.

3.3.4. Motivation

Of eight studies with coded results, four were Waldorf-favouring, two comparator-favouring, one Mixed and one ns; by direction of effect, four favoured Waldorf and three the comparator (sign test $p = 1.00$, also for independent samples). Two of the Waldorf-favouring studies are unadjusted comparisons with the Austrian national mean (Wallner-Paschon, 2018; Wallner-Paschon & Wiesinger, 2021). In a propensity-matched sample, Salchegger et al. (2021) found higher interest in science among Waldorf pupils ($d = 0.30$), whereas the difference in enjoyment ($d = 0.24$) was not significant; both were smaller than the unmatched differences ($d = 0.48$) and largely mediated by inquiry-based science education. Kvintova et al. (2022) found more positive attitudes towards preschool, and van Schie and Vedder (2023) no difference in school attitudes. Bilde et al. (2013), adjusting for SES and prior achievement, found less independent participation in a pooled Waldorf and Freinet group ($B = -0.25$).

3.3.4.1. Self-concept findings conflict. In Yevsikova (2008), all eight significant differences on nine self-attitude scales favoured the traditional school. Besançon et al. (2015) found lower introjected regulation among Waldorf students but higher identified regulation among conventional students. Dorfman and Fortus (2019), described narratively, pooled traditional and democratic schools in the comparator and reported no Waldorf–comparator test; within the Waldorf schools, science self-efficacy peaked in Grade 6 and fell in Grade 7, which the authors attributed to the introduction of grades and examinations. No clear direction emerges.

3.3.5. Specific Waldorf methods

Of the 37 studies, 27 named specific pedagogical elements. Most studies compare school types rather than manipulating components, so the associations below are hypotheses, not mechanisms established by this review. Delayed reading instruction has attracted the most attention (six studies; Sections 3.3.1 and 3.3.6); even documented convergence would not establish that the practice is without risk for children with reading difficulties (Lovett et al., 2017). Inquiry-based science education is the only element formally tested as a mediator: it largely mediated the motivational differences, while the achievement difference remained negative (Salchegger et al., 2021). Arts integration is compatible with the advantages in rated drawing quality and creative thinking, but equally with domain-specific practice and family background. No isolated evidence exists for epochal teaching, narrative assessment or eurythmy.

3.3.6. Developmental trajectory

Several studies report age- or grade-related differences compatible with later convergence, but the evidence does not establish a common longitudinal catch-up trajectory, because the four relevant studies differ in design. Telfer-Radzat (2024) and Larrison et al. (2012) compare school-level results across grades: Waldorf-inspired or Waldorf-methods public schools scored lower in the early grades and similarly or higher by Grades 7–8, but mostly in repeated cross-sections of different cohorts rather than individual trajectories. Cunningham and Carroll (2011) followed older Steiner (7–9 years) and younger standard-educated children (4–6 years) through the same two years of formal instruction; reading did not differ significantly, but the standard group remained ahead in spelling. Suggate et al. (2011) is cross-sectional at school entry, with no later measurement. A comparable age-dependence appears in drawing (Rose et al., 2012).

The grade-related pattern is consistent with the developmental readiness logic of the curriculum, but it is among the least securely established findings. Differences across grades in school-level data can also reflect changes in cohort composition and selective school leaving, adjustment for family background was limited, and none of the four studies controlled for baseline cognitive ability or out-of-school tutoring. That the resources of families who choose Waldorf schools compensate for the later start of formal instruction cannot be distinguished from a pedagogical effect with these designs.

3.3.7. Socioeconomic adjustment: between-study pattern and within-study comparisons

3.3.7.1. Between-study pattern. In academic achievement, comparator-favouring findings were more frequent among the seven studies with statistical SES control than among the 11 without (3/7 vs 2/11), whereas Waldorf-favouring findings were infrequent in both strata (1/7 vs 2/11); the difference rests on one or two studies (Table 4). The single Waldorf-favouring adjusted study (van Schie & Vedder, 2023) is classified as adjusted on the authors' statement that entering SES indicators did not change the result; adjusted estimates are not reported. The socio-emotional and motivation strata contain only two adjusted studies each, and no creativity study adjusted statistically for SES. The strata differ in country, design, age, outcomes and comparators, so this is a descriptive pattern, not evidence that adjustment causes the differences.

3.3.7.2. Within-study comparisons. Six studies reported the same contrast at two levels of adjustment (Supplementary Table S7): five gave unadjusted and SES-adjusted estimates, and one (Suggate et al., 2011) a less- and a more-adjusted model; a further study described the result of adjustment without reporting adjusted estimates (van Schie & Vedder, 2023). Adjustment did not act in one direction. In academic achievement, the one significant unadjusted Waldorf advantage with a reported adjusted estimate was no longer significant after adjustment (Schindler Rangvid, 2008), and the numerical Waldorf lead in PISA science reversed to a significant disadvantage after propensity matching (Salchegger et al., 2021); a comparator advantage was no longer significant after further adjustment in one study (Suggate et al., 2011) but persisted or grew in another (Schindler Rangvid, 2008). Solhaug's (2007) socio-emotional advantages persisted, disappeared or emerged depending on outcome and school. In motivation, a pooled Waldorf and Freinet group showed a significant disadvantage in independent participation only after adjustment (Bilde et al., 2013), and the science-motivation advantages were smaller after matching (Salchegger et al., 2021). This evidence is limited, but it shows that adjustment can change findings in either direction without permitting an estimate of the overall magnitude of socioeconomic confounding. Risk of bias is not used as a stratifier, because the harmonised rating summarises tools that assess different dimensions.

Table 4
Direction of findings by statistical adjustment for socioeconomic status (study level; descriptive between-study comparison).

Outcome domain	Statistical SES control	Studies (n)	W+	ns	C+	Mixed	Uncertain	W+ n/N	C+ n/N
Academic achievement	Yes	7	1	1	3	1	1	1/7	3/7
	No	11	2	2	2	4	1	2/11	2/11
Socio-emotional development	Yes	2	1	1	0	0	0	1/2	0/2
	No	8	7	1	0	0	0	7/8	0/8
Creativity (incl. motor and arts)	Yes	0	–	–	–	–	–	–	–
	No	10	5	3	0	2	0	5/10	0/10
Motivation and self-concept	Yes	2	1	0	1	0	0	1/2	1/2
	No	6	3	1	1	1	0	3/6	1/6

Note. Statistical SES control: Full or Stat. for the coded estimates of the study-domain; No: design-level control only, pre-test equivalence, or none. The comparison is descriptive; SES was not tested as a moderator, and differences between strata do not estimate SES attenuation. Adjustment sets differ substantially in comprehensiveness (Supplementary Table S6).

3.3.8. Comparisons between alternative school models

Nine studies also included a comparator from another alternative pedagogy; eight contribute coded contrasts (Supplementary Table S9), since the democratic schools in Dorfman and Fortus (2019) are pooled with traditional schools. Against Montessori comparators, Waldorf-favouring results predominated in socio-emotional development (Kirkham & Kidd, 2017; Kvintova et al., 2022; Özçelik & Sapsağlam, 2023) and creativity (Cox & Rowlands, 2000; Kirkham & Kidd, 2017), whereas Rose et al. (2012) found Montessori children ahead at age 5 and Steiner children ahead on several expressive measures at ages 7 and 9. Mather (2012), Kvintova et al. (2022, motivation) and Shankland et al. (2009) found no significant differences. In the only comparison with another developmental pedagogy, Elkonin–Davydov pupils handled contradictory information better (Tsukerman, 2005). These contrasts are few, small and unadjusted for SES, and they compare school models rather than isolating Waldorf components or mechanisms.

3.4. Summary of findings (GRADE)

Certainty was very low in all four domains (Table 5). The non-randomised evidence was rated down two levels for risk of bias in every domain; further downgrades were for inconsistency and imprecision in academic achievement (inconsistency alone already yields very low certainty), for imprecision and suspected publication bias in socio-emotional development and creativity, and for inconsistency and imprecision in motivation. The cluster-randomised trial was rated separately and downgraded for high risk of bias, indirectness and very serious imprecision, including its unit-of-analysis error. Upgrading was not applied (Supplementary Table S4).

4. Discussion

4.1. Summary of evidence

Based on the 35 studies with coded results across four outcome domains, the GRADE assessment yielded very low certainty throughout. This indicates an absence of convincing evidence, not evidence that Waldorf education has no effect.

The evidence answers three questions to different degrees. First, descriptive differences between school groups are well documented, in directions that are mixed for academic achievement and motivation and predominantly Waldorf-favouring for socio-emotional development and creativity. Second, associations that remain after statistical adjustment are documented only sparsely: nine studies adjusted statistically for SES,

Table 5
Summary of findings: Waldorf education compared with non-Waldorf education in preschool and school-age children (approximately 4–18 years).

Outcome domain	Studies (n)	Significance-based (W+/C+/Mixed/ns/unc.)	Direction of effect (W/C)	Main finding	Certainty (GRADE)
Academic achievement (non-randomised studies)	17	2/5/5/3/2	5/10	No consistent direction. Comparator-favouring findings were more frequent among studies with statistical SES adjustment (3/7 vs 2/10), and Waldorf-favouring findings were infrequent in both (1/7 vs 1/10; descriptive between-study pattern); within studies, adjustment changed findings in both directions. Compatible with residual socioeconomic confounding of unknown magnitude.	Very low ⊕○○○
Academic achievement (cluster-randomised trial)	1	1/0/0/0/0	1/0	Higher early number skills after a four-week Waldorf-inspired intervention combined with cooperative learning ($n = 30$, one cluster per arm); high risk of bias; the child-level analyses ignore clustering, so no valid significance test is possible. Rated separately.	Very low ⊕○○○
Socio-emotional development	10	8/0/0/2/0	8/1	Predominantly Waldorf-favouring (well-being, coping, executive functions and self-organisation, civic attitudes, moral reasoning and responsibility, preschool anxiety-withdrawal); mostly unblinded subjective measures; two studies adjusted for SES, with discordant results.	Very low ⊕○○○
Creativity (incl. motor and arts)	10	5/0/2/3/0	9/0	Predominantly Waldorf-favouring in rated drawing quality and divergent thinking; comparator-favouring results only age-specific within two studies; no study adjusted statistically for SES; possible domain-specific practice effect.	Very low ⊕○○○
Motivation and self-concept	8	4/2/1/1/0	4/3	No consistent direction; two Waldorf-favouring studies are unadjusted comparisons with the national mean; in the propensity-matched study only the advantage in interest in science remained significant, largely mediated by inquiry-based science education; self-concept findings favoured the traditional school.	Very low ⊕○○○

Note. GRADE adapted for narrative synthesis (Murad et al., 2017); non-randomised evidence appraised with ROBINS-I starts at high certainty (Schünemann et al., 2019). Source-related concerns did not contribute to downgrading. Very low certainty expresses the absence of convincing evidence, not evidence of no effect. The academic row for non-randomised studies excludes the cluster-randomised trial, so its stratum counts (3/7 vs 2/10; 1/7 vs 1/10) differ from those in Table 4 and the Abstract, which include it (3/7 vs 2/11; 1/7 vs 2/11). Domain-level justification in Supplementary Table S4.

with adjustment sets of very different comprehensiveness, and six reported estimates of the same contrast at two levels of adjustment. In academic achievement, comparator-favouring findings were more frequent among studies adjusting for SES, while Waldorf-favouring findings were infrequent with or without adjustment. Third, causal effects of Waldorf education are not established by observational comparisons of self-selected school populations. A recurring descriptive pattern is age- or grade-related: early deficits are followed in later grades by smaller or reversed differences, mostly in repeated cross-sectional data. This is compatible with later convergence but does not establish a common longitudinal catch-up trajectory, and its causal status is unresolved.

Two further points qualify these answers. The sign tests for socio-emotional development and creativity ($p = .039$ and .008 for independent samples) show that Waldorf-favouring directions predominate, but they convey neither the size nor the cause of the differences: both domains rest mostly on unblinded reports or on drawing tasks close to the Waldorf curriculum, few studies adjusted for SES, and publication bias is suspected. For the developmental readiness hypothesis that framed this review (Section 1.1), the evidence is inconclusive: no study followed individual children from an early deficit to later convergence with adequate control for family background and baseline ability, and the only long-term register follow-up found less, not more, educational attainment among Waldorf students after adjustment (Schindler Rangvid, 2008).

4.2. Methodological challenges in Waldorf research

Of the 37 studies, 34 (92%) were at high risk of bias, and only one used random allocation (30 children, one cluster per arm). Families choose Waldorf schools for pedagogical, ideological, economic and geographical reasons, so individual randomisation is rarely feasible; the relevant question is whether observational studies address selection and confounding adequately, a standard the present corpus meets only partially.

The most critical weakness is incomplete adjustment for confounding. Twenty-eight studies (76%) applied no statistical SES control, and even comprehensive adjustment cannot account for parental educational values, cultural engagement, prior achievement or child motivation. Residual confounding is therefore plausible in every domain. The Austrian PISA analyses illustrate the point: the same census of Waldorf pupils showed a numerical science advantage before matching and a significant disadvantage after propensity matching (Salchegger et al., 2021; Wallner-Paschon, 2018). This is compatible with socioeconomic confounding in the academic domain but does not permit an estimate of its magnitude across the literature. Comparators are also asymmetric: only nine studies included an alternative-pedagogy comparator, so comparisons between alternative school models, which come closest to the question of what distinguishes Waldorf schooling, are few and small (Section 3.3.8).

4.3. Comparison with other alternative pedagogies

The evidence base of Waldorf pedagogy lags behind that of Montessori education, which has been synthesised by a meta-analysis (Demangeon et al., 2023) and a systematic review (Randolph et al., 2023), both including randomised trials and documenting small-to-moderate positive effects. This gap does not imply that the pedagogy is less effective; it reflects designs that do not permit reliable, generalisable conclusions. Within the present corpus, the few direct comparisons with Montessori schools mostly favoured Waldorf in socio-emotional and creative outcomes (Section 3.3.8), but they are small, unadjusted for SES and confined mostly to preschool and primary ages, and do not alter this assessment.

4.4. Practical and policy implications

Three cautious implications follow. First, catch-up after delayed reading instruction is not securely established, and delayed instruction may carry risks for children with reading difficulties (Lovett et al., 2017); Waldorf institutions should consider evidence-based screening and early intervention. Second, the mediating role of inquiry-based science education in science motivation (Salchegger et al., 2021) suggests, as a hypothesis for testing, that specific didactic elements might be transferable to mainstream schools. Third, claims about the effectiveness of Waldorf schools should be treated with caution until confirmed by adequately controlled, independent research.

4.5. Recommendations for future research

Future Waldorf research should address four requirements: (a) systematic adjustment for SES, baseline cognitive performance and other relevant confounders through propensity score methods or multilevel modelling, with unadjusted and adjusted estimates reported side by side; (b) prospective longitudinal designs that document developmental trajectories; (c) component-level evaluations of distinctive elements (epochal teaching, eurythmy, narrative assessment); and (d) direct assessment of implementation fidelity.

5. Limitations

First, the exclusion of non-peer-reviewed sources may have constrained coverage, and the relative effectiveness of Waldorf education may be overestimated in the published literature. Second, implementation fidelity was not independently verified and was measured inconsistently across studies, and several study units concerned Waldorf-inspired or partial programmes. Third, the synthesis rests on counts that give equal weight to studies of very different size; the significance-based classification depends on statistical power, and the Mixed category on the number of analyses. The direction-of-effect classification and the majority rule offset but cannot replace statistical weighting, and the sign test counts studies rather than effect sizes; a consistent numerical direction conveys neither the size nor the precision of the differences, and the alternative classifications do not give identical results (under the majority rule, for example, five of the ten creativity studies were ns). Fourth, only six studies reported estimates of the same contrast at two levels of adjustment, and the between-study comparison of SES strata is descriptive. Fifth, entry-level codes were verified by a second reviewer rather than coded independently in duplicate, and no agreement statistic was computed. ROBINS-I was applied at study level rather than separately for each outcome domain. Several primary studies reported internally inconsistent statistics; these were recorded as reported and flagged. The move from entry-level to study-level counting, the NOS comparability rule, the exclusion of a study at critical risk of bias, the separate synthesis of alternative-pedagogy contrasts, the pooled-group rule and the GRADE starting level are departures from the registered protocol (Sections 2.2, 2.5, 2.7 and 2.8).

6. Conclusion

This review synthesised 38 reports of 37 studies on Waldorf education (36 in the synthesis, 35 with coded results). Academic achievement showed no consistent direction (Q1). Socio-emotional findings were predominantly Waldorf-favouring and creativity findings Waldorf-favouring in rated drawing quality, while motivation showed no clear pattern (Q2). Delayed reading instruction has received the most empirical attention, and inquiry-based science education is the only element formally tested as a mediator (Q3). Several studies report age- or grade-related differences in academic achievement compatible with later convergence, mostly from repeated cross-sections; no common

longitudinal catch-up trajectory is established, and none of these studies controlled for baseline cognitive ability (Q4). The corpus is methodologically weak: 92% of studies carried a high risk of bias and 76% applied no statistical control for socioeconomic status, and certainty was very low in all domains (Q5).

Comparator-favouring academic findings were more frequent among studies that adjusted for socioeconomic background, whereas Waldorf-favouring findings were infrequent with or without adjustment, and in the one sample analysed both ways a small numerical science advantage became a significant disadvantage after matching. These findings are compatible with residual socioeconomic confounding of unknown magnitude. In the socio-emotional and creative domains, adjustment was rare or absent, so the evidence cannot separate pedagogy from selection. The review therefore indicates substantial uncertainty about the causal effects of Waldorf education rather than evidence that such effects are absent. Establishing these effects will require longitudinal, SES-adjusted and component-level studies by independent investigators that document implementation.

CRediT authorship contribution statement

Mátyás Turós: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Ethical statement

As this study is a systematic literature review based entirely on previously published secondary data, direct institutional review board (IRB) ethical approval and participant informed consent were not required.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used Claude Sonnet 4.6 and Gemini 3.1 Pro to assist with data extraction and quality appraisal across the full corpus, and Gemini 2.5 Flash to cross-check the extracted data and appraisals against the original full texts. Claude Sonnet 4.6 was additionally used to translate records not published in English or German for title-abstract screening and, where necessary, for full-text eligibility assessment; these translations were validated by the three-tier protocol described in Section 2.2. All extracted data points were subsequently verified by the author and a second, independent reviewer against the original texts, and the coding rules introduced during peer review were applied without AI assistance. After using these tools the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The author is employed part-time as an after-school educator at a primary Waldorf school, in a role that involves neither curricular teaching nor research; the author's primary affiliation is Eszterházy Károly Catholic University, which has no connection with Waldorf teacher training, research or advocacy. The second reviewer has no affiliation with any Waldorf institution. The source-bias criteria applied to the included studies are given in Supplementary Text S1.2.

Acknowledgments

The author would like to thank Dr. Péter G. Molnár (independent researcher) for his contribution to the title-abstract screening, full-text eligibility assessment and consensus resolution, and for checking the extracted data and outcome codes against the source texts. Dr. Molnár was compensated for his time. His role did not meet the ICMJE criteria for authorship, as he was not involved in the conception, design or interpretation of the review findings.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2026.108128.

Data availability

The review protocol was prospectively registered in PROSPERO (CRD420261288126). All data and materials supporting this review are openly available in Zenodo at doi:https://doi.org/10.5281/zenodo.23107860 under a CC BY 4.0 licence. The repository contains the search strategies, the title-abstract and full-text screening records, the inter-rater reliability data, the data-extraction prompt and scripts, and the final extraction and appraisal database from which all reported figures and tables were generated. Its README and codebook document the decision rules applied.

References

Abdulaeva, E. A. (2022). From responsiveness to self-organization: A comparative study of approaches to children in Waldorf and "Directive" preschool education. *National Psychological Journal, 47*(3), 77–88. https://doi.org/10.11621/npj.2022.0310

Arslan, A., & Kartal, S. (2022). The effect of the Waldorf approach applied with the collaborative method on the early assessment skills of 48–60 months-old children. *International Online Journal of Education & Teaching, 9*(3), 1096–1112. https://eric.ed.gov/?id=EJ1351905.

Attfield, K. (2024). The humane education of Waldorf. *Frontiers in Education, 9*, Article 1332597. https://doi.org/10.3389/feduc.2024.1332597

Besançon, M., Fenouillet, F., & Shankland, R. (2015). Influence of school environment on adolescents' creative potential, motivation and well-being. *Learning and Individual Differences, 43*, 178–184. https://doi.org/10.1016/j.lindif.2015.08.029

Bilde, J. de, van Damme, J., Lamote, C., & Fraine, B. de (2013). Can alternative education increase children's early school engagement? A longitudinal study from kindergarten to third grade. School Effectiveness and School Improvement, 24(2), 212–233. doi:https://doi.org/10.1080/09243453.2012.749795.

Borodulkina, T. (2019). The empirical study of psychosemantics of children's drawing: Waldorf experiment. *Insight: The Psychological Dimensions of Society, 2*, 44–51. https://doi.org/10.32999/2663-970X/2019-2-6

Burkitt, E., Barrett, M., & Davis, A. (2005). Drawings of emotionally characterised figures by children from different educational backgrounds. *International Journal of Art & Design Education, 24*(1), 71–83. https://doi.org/10.1111/j.1476-8070.2005.00424.x

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology, 46*(5), 423–429. https://doi.org/10.1016/0895-4356(93)90018-V

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V.[. S. V.]., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ, 368*, Article l6890. https://doi.org/10.1136/bmj.l6890

Cox, M. V., & Rowlands, A. (2000). The effect of three different educational approaches on children's drawing ability: Steiner, Montessori and traditional. *British Journal of Educational Psychology, 70*(4), 485–503. https://doi.org/10.1348/000709900158263

Critical Appraisal Skills Programme. (2018). CASP qualitative checklist. https://casp-uk.net/casp-tools-checklists/.

Cunningham, A. J., & Carroll, J. M. (2011). The development of early literacy in Steiner- and standard-educated children. *British Journal of Educational Psychology, 81*, 475–490. https://doi.org/10.1348/000709910X522474

Dahlin, B. (2010). A state-independent education for citizenship? Comparing beliefs and values related to civic and moral issues among students in Swedish mainstream and Steiner Waldorf schools. *Journal of Beliefs & Values, 31*(2), 165–180. https://doi.org/10.1080/13617672.2010.503629

Dahlin, B. (2017). Conclusion: Does it work? Empirical studies of Waldorf education. In B. Dahlin (Ed.), *Springer briefs in education. Rudolf Steiner* (pp. 125–143). Springer International Publishing. https://doi.org/10.1007/978-3-319-58907-7_7.

Deluca, C., Pyle, A., Roy, S., Chalas, A., & Danniels, E. (2019). Perspectives on kindergarten assessment: Toward a common understanding. *Teachers College Record: The Voice of Scholarship in Education, 121*(3), 1–58. https://doi.org/10.1177/016146811912100302

Demangeon, A., Claudel-Valentin, S., Aubry, A., & Tazouti, Y. (2023). A meta-analysis of the effects of Montessori education on five fields of development and learning in preschool and school-age children. *Contemporary Educational Psychology, 73*, Article 102182. https://doi.org/10.1016/j.cedpsych.2023.102182

Diefenbach, C., Schlecht, J., König, J., Martin, D., Patzlaff, R., Schmidt, M. F., & Urschitz, M. S. (2020). Schulische Fähigkeiten von Kindern mit speziellem medizinischem Versorgungsbedarf am Ende des ersten Schuljahres. Ergebnisse zweier Kohortenstudien an Waldorfschulen und traditionellen Schulen. *RoSE, 11*(1). https://www.rosejourn.com/index.php/rose/article/view/546.

Dorfman, B.-S., & Fortus, D. (2019). Students' self-efficacy for science in different school systems. *Journal of Research in Science Teaching, 56*(8), 1037–1059. https://doi.org/10.1002/tea.21542

Goldshmidt, G. (2017). Waldorf education as spiritual education. *Religion & Education, 44*(3), 346–363. https://doi.org/10.1080/15507394.2017.1294400

Guerrero, S., Valenciano-Valcárcel, J., & Rodríguez, A. (2024). Unveiling alternative schools: A systematic review of cognitive and social-emotional development in different educational approaches. *Children and Youth Services Review, 158*, Article 107480. https://doi.org/10.1016/j.childyouth.2024.107480

Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., DeBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P., & Schünemann, H. J. (2011). Grade guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology, 64*(4), 383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *The British Journal of Mathematical and Statistical Psychology, 61*, 29–48. https://doi.org/10.1348/000711006X126600

Howard, J. L., & Slemp, G. R. (2025). Why we need a living meta-analysis of self-determination theory: An illustration examining temporal changes in need supportive education. *Motivation and Emotion, 50*, 193–210. https://doi.org/10.1007/s11031-025-10132-2

Immordino-Yang, M. H., Darling-Hammond, L., & Krone, C. R. (2019). Nurturing nature: How brain development is inherently social and emotional, and what this means for education. *Educational Psychologist, 54*(3), 185–204. https://doi.org/10.1080/00461520.2019.1633924

Ionova, O., Gres, O., Partola, W., & Luparenko, S. (2018). Waldorf approaches to organization of the pedagogical process and their influence on the quality of education of Ukrainian junior schoolchildren. *The New Educational Review, 54*(4), 135–145. https://doi.org/10.15804/tner.2018.54.4.11

Kikas, E. (1998). Pupils' explanations of seasonal changes: Age differences and the influence of teaching. *British Journal of Educational Psychology, 68*(4), 505–516. https://doi.org/10.1111/j.2044-8279.1998.tb01308.x

Kikas, E. (2000). The influence of teaching on students' explanations and illustrations of the day/night cycle and seasonal changes. *European Journal of Psychology of Education, 15*(3), 281–295. https://doi.org/10.1007/BF03173180

Kirkham, J. A., & Kidd, E. (2017). The effect of Steiner, Montessori, and national curriculum education upon children's pretence and creativity. *The Journal of Creative Behavior, 51*(1), 20–34. https://doi.org/10.1002/jocb.83

Kodsi, S. H. (2021). Constructive play and technological intelligence in kindergarten. *Jami'a, 23*(2). https://ojs.qsm.ac.il/index.php/jamiaa/article/view/166.

Kodsi, S. H. (2022). Constructive play in Waldorf and normative preschools in Israel: Technological thinking and design process during free play. *International Journal of Technology and Design Education, 32*(2), 735–748. https://doi.org/10.1007/s10798-020-09623-9

Korkmaz, S., & Öztürk Samur, A. (2022). Waldorf Yaklaşımına Dayalı Eğitim Programının Okul Öncesi Dönemdeki Çocukların Yaratıcılık ve Bilişsel Becerilerine Etkisi (Effect of Waldorf approach-based education program on the creativity and cognitive skills of preschool children). *İnsan ve Toplum Bilimleri Araştırmaları Dergisi, 11*(1), 452–488. https://doi.org/10.15869/itobiad.880321

Kóródi, K., Jagodics, B., Szabó, É., Kunos, N., & Fülöp, M. (2023). Diák kiégés hagyományos és alternatív pedagógiai programú oktatási intézményekben. *Magyar Pszichológiai Szemle, 78*(3), 329–344. https://doi.org/10.1556/0016.2023.00043

Kvintova, J., Kremenkova, L., Cuberek, R., Petrova, J., Stuchlikova, I., Dobesova-Cakirpaloglu, S., Pugnerova, M., Balatova, K., Lemrova, S., Viteckova, M., & Plevova, I. (2022). Preschoolers' attitudes, school motivation, and executive functions in the context of various types of kindergarten. *Frontiers in Psychology, 13*, Article 823980. https://doi.org/10.3389/fpsyg.2022.823980

Larrison, A. L., Daly, A. J., & VanVooren, C. (2012). Twenty years and counting: A look at Waldorf in the public sector using online sources. *Current Issues in Education, 15*(3). http://cie.asu.edu/ojs/index.php/cieatasu/article/view/807.

Lovett, M. W., Frijters, J. C., Wolf, M., Steinbach, K. A., Sevcik, R. A., & Morris, R. D. (2017). Early intervention for children at risk for reading disabilities: The impact of grade at intervention and individual differences on intervention outcomes. *Journal of Educational Psychology, 109*(7), 889–914. https://doi.org/10.1037/edu0000181

Martinsone, B., Simões, C., Camilleri, L., Conte, E., & Lebre, P. (2025). Students' socio-emotional skills and academic outcomes after the PROMEHS program: A longitudinal study in two European countries. *Behavioral Sciences, 15*(11). https://doi.org/10.3390/bs15111529

Mather, D. S. (2012). Ipsilateral printing in children's mirror-writing: A cause of specific learning disabilities? *Canadian Journal of Experimental Psychology = Revue Canadienne De Psychologie Experimentale, 66*(3), 172–180. https://doi.org/10.1037/a0025873

Matos, L., Reeve, J., Gargurevich, R., & Herrera, D. (2025). An autonomy-supportive teaching intervention benefits students in a low-SES school setting. *The Journal of Experimental Education*, 1–16. https://doi.org/10.1080/00220973.2025.2592063

Mavrelos, M., & Daradoumis, T. (2020). Exploring multiple intelligence theory prospects as a vehicle for discovering the relationship of neuroeducation with imaginative/Waldorf pedagogy: A systematic literature review. *Education Sciences, 10*(11), Article 334. https://doi.org/10.3390/educsci10110334

McKenzie, J. E., & Brennan, S. E. (2019). Synthesizing and presenting findings using other methods. In J. P. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 321–347). Wiley. https://doi.org/10.1002/9781119536604.ch12.

Murad, M. H., Mustafa, R. A., Schünemann, H. J., Sultan, S., & Santesso, N. (2017). Rating the certainty in evidence in the absence of a single estimate of effect. *Evidence-Based Medicine, 22*(3), 85–87. https://doi.org/10.1136/ebmed-2017-110668

Özçelik, M., & Sapsağlam, Ö. (2023). Examination of the social competence and types of competition of preschool children in terms of different early childhood education programs. *International E-Journal of Educational Studies, 7*(13), 1–12. https://doi.org/10.31458/iejes.1189921

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. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Randoll, D., & Peters, J. (2015). Empirical research on Waldorf education. *Educar Em Revista, 56*, 33–47. https://doi.org/10.1590/0104-4060.41416

Randolph, J. J., Bryson, A., Menon, L., Henderson, D. K., Kureethara Manuel, A., Michaels, S., Rosenstein, D. L. W., McPherson, W., O'Grady, R., & Lillard, A. S. (2023). Montessori education's impact on academic and nonacademic outcomes: A systematic review. *Campbell Systematic Reviews, 19*(3), Article e1330. https://doi.org/10.1002/cl2.1330

Rawson, M. (2024). Translating, transmitting and transforming Waldorf curricula: One hundred years after the first published curriculum in 1925. *Frontiers in Education, 9*, Article 1306092. https://doi.org/10.3389/feduc.2024.1306092

Rose, S. E., & Jolley, R. P. (2016). Drawing development in mainstream and Waldorf Steiner schools revisited. *Psychology of Aesthetics, Creativity, and the Arts, 10*(4), 447–457. https://doi.org/10.1037/aca0000070

Rose, S. E., & Jolley, R. P. (2019). Children's creative intentions: Where do the ideas for their drawings come from? *The Journal of Creative Behavior, 54*(3), 712–724. https://doi.org/10.1002/jocb.405

Rose, S. E., Jolley, R. P., & Charman, A. (2012). An investigation of the expressive and representational drawing development in National Curriculum, Steiner, and Montessori schools. *Psychology of Aesthetics, Creativity, and the Arts, 6*(1), 83–95. https://doi.org/10.1037/a0024460

Salchegger, S., Wallner-Paschon, C., & Bertsch, C. (2021). Explaining Waldorf students' high motivation but moderate achievement in science: Is inquiry-based science education the key? *Large-Scale Assessments in Education, 9*(1), 14. https://doi.org/10.1186/s40536-021-00107-3

Schindler Rangvid, B. (2008). Private school diversity in Denmark's national voucher system. *Scandinavian Journal of Educational Research, 52*(4), 331–354. https://doi.org/10.1080/00313830802184491

Schünemann, H. J., Cuello, C., Akl, E. A., Mustafa, R. A., Meerpohl, J. J., Thayer, K., Morgan, R. L., Gartlehner, G., Kunz, R., Katikireddi, S. V.[. S. V.]., Sterne, J., Higgins, J. P.[. J. P.]., & Guyatt, G. (2019). Grade guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. *Journal of Clinical Epidemiology, 111*, 105–114. https://doi.org/10.1016/j.jclinepi.2018.01.012

Shankland, R., França, L. R., Genolini, C. M., Guelfi, J.-D., & Ionescu, S. (2009). Preliminary study on the role of alternative educational pathways in promoting the use of problem-focused coping strategies. *European Journal of Psychology of Education, 24*(4), 499–512. https://doi.org/10.1007/BF03178764

Shi, J., & Cheung, A. C. K. (2024). Effective components of social emotional learning programs: A meta-analysis. *Journal of Youth and Adolescence, 53*(4), 755–771. https://doi.org/10.1007/s10964-024-01942-7

Solhaug, T. (2007). Steinerskoler i et demokratisk perspektiv – En sammenligning med offentlige skoler i Norge (Steiner schools in a democratic perspective: A comparison with public schools in Norway). *Nordisk Pedagogik, 27*(2), 150–172. https://www.scup.com/doi/10.18261/ISSN1891-5949-2007-02-05.

Stehlik, T. (2019). A review of the literature on Steiner education and Waldorf schools. In T. Stehlik (Ed.), *Waldorf schools and the history of Steiner education* (pp. 95–122). Springer International Publishing. https://doi.org/10.1007/978-3-030-31631-0_5.

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, 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.[. J. P.]. (2016). Robins-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ, 355*, Article i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). Rob 2: A revised tool for assessing risk of bias in randomised trials. *BMJ, 366*, Article l4898. https://doi.org/10.1136/bmj.l4898

Su, M., & Lee, J. (2023). Eastern versus Western gaps in whole child education: How school climate works for academic achievement and well-being across China, Korea, Finland, and the United States. *Educational Research for Policy and Practice, 22*(1), 117–138. https://doi.org/10.1007/s10671-022-09322-3

Suggate, S. P. (2010). Why what we teach depends on when: Grade and reading intervention modality moderate effect size. *Developmental Psychology, 46*(6), 1556–1579. https://doi.org/10.1037/a0020612

Suggate, S. P., Schaughency, E. A., & Reese, E. (2011). The contribution of age and reading instruction to oral narrative and pre-reading skills. *First Language, 31*(4), 379–403. https://doi.org/10.1177/0142723710395165

Telfer-Radzat, K. (2024). Exploring alternative education: A comparison on 3 levels—Waldorf charters, non-Waldorf charters, and local public schools, as measured by the California assessment of student performance and progress. *Frontiers in Education, 9*, Article 1323535. https://doi.org/10.3389/feduc.2024.1323535

Tsortanidou, X., Daradoumis, T., & Barberá-Gregori, E. (2022). Convergence among imagination, social-emotional learning and media literacy: An integrative literature review. *Early Child Development and Care, 192*(2), 173–186. https://doi.org/10.1080/03004430.2020.1753720

Tsukerman, G. A. (2005). СистеМа Д.Б. Эльконина – В.В. Давыдова как ресурс повышения коМпетентности российских школьников (The Elkonin-Davydov system as a resource for enhancing the competence of Russian schoolchildren). *Voprosy Psikhologii, 4*, 84–95. https://author-club.org/media/files/d8fa367e18c2b517d5fccf9a13da5f38739a2e28.pdf.

Turós, M. (2023). The impact of public, Catholic and Waldorf schools on pupils' moral judgement. *International Journal of Educational Development, 97*, Article 102706. https://doi.org/10.1016/j.ijedudev.2022.102706

Tyson, R. (2024). Waldorf education – A survey of empirical research. *Research on Steiner Education, 15*(1), 1–24.

van Schie, T. J., & Vedder, P. (2023). Different pedagogies, equivalent results: A comparison of language skills and school attitude between Waldorf school students and public school students in the Philippines. *Globalisation, Societies and Education, 24*(2), 490–503. https://doi.org/10.1080/14767724.2023.2248902

Waldorf World List. (2025). *Waldorf world list*. International Council for Steiner Waldorf Education (Hague Circle). https://www.waldorf-international.org/en/waldorf-world-list-2025/.

Wallner-Paschon, C. (2018). Kompetenzen, Motivation und Selbstwahrnehmung der Waldorfschüler/innen im Vergleich. *R&E-Source, 1*(9). https://journal.ph-noe.ac.at/index.php/resource/article/view/548.

Wallner-Paschon, C., & Wiesinger, L. (2021). Waldorf- und Regelschulen in Österreich. *R&E-Source, 16*, 1–11. https://doi.org/10.53349/resource.2021.i16.a994

Wells, G. A., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2000). The Newcastle–Ottawa Scale (NOS) for assessing the quality of non-randomized studies in meta-analysis. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.

Yevsikova, N. I. (2008). Особенности саМоотношения подростков, обучающихся в традиционной и вальдорфской школах (Characteristics of the attitude to self in adolescents attending traditional and Waldorf schools). *Вопросы Психологии (Voprosy Psikhologii), 6*, 45–56. https://elibrary.ru/item.asp?id=15500041.

Zixia, Z., Rahimi, N. H. Z., & Mohamad, N. (2025). Academic engagement and basic psychological needs: A systematic review of research trends in self-determination theory. *International Journal of Learning, Teaching and Educational Research, 24*(7), 813–840. https://doi.org/10.26803/ijlter.24.7.40