

Article

The Effects of Philosophy for Children on Children's Cognitive Development: A Three-Level Meta-Analysis

Caiyun Wei ¹ and Lele Chen ^{2,*}¹ School of Education Science, Nanjing Normal University, Nanjing 210097, China; 170601020@njnu.edu.cn² Faculty of Education, Henan University, Kaifeng 475004, China

* Correspondence: 10030112@vip.henu.edu.cn

Abstract

Amid the rise of the knowledge economy, accelerated informatization, and the emergence of artificial intelligence, Philosophy for Children (P4C) has been promoted as an effective educational project to enhance children's cognitive development, especially higher-order thinking skills. However, empirical evidence regarding its efficacy remains inconclusive. This three-level meta-analysis synthesizes 53 effect sizes derived from 33 experimental and quasi-experimental studies involving 4568 participants to assess P4C's cognitive effects and potential moderators. The results reveal a statistically significant and moderate-to-strong overall effect ($g = 0.59$). Significant and robust effects were specifically observed for reasoning, critical thinking, and creativity. Subgroup and meta-regression analyses identified sample size as a significant moderator: smaller samples tended to report larger effect sizes. Additionally, cultural context and session length showed marginally significant moderating effects. Crucially, P4C's cognitive impact remained consistent across grade levels, research designs, and publication years, demonstrating robustness and stability across diverse implementation conditions. These findings provide updated and nuanced evidence for the effectiveness of P4C, underscoring its cross-contextual robustness and specific value in fostering cognitive abilities. Implications for policymakers, educators, and future researchers aiming to implement or investigate P4C in varied educational settings are discussed.



Received: 26 August 2025

Revised: 4 October 2025

Accepted: 8 October 2025

Published: 13 October 2025

Citation: Wei, Caiyun, and Lele Chen. 2025. The Effects of Philosophy for Children on Children's Cognitive Development: A Three-Level Meta-Analysis. *Journal of Intelligence* 13: 130. <https://doi.org/10.3390/jintelligence13100130>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: cognitive development; Philosophy for Children; three-level meta-analysis; higher-order thinking; community of inquiry

1. Introduction

In the context of the knowledge economy, rapid informatization, and the emergence of artificial intelligence, cultivating students' cognitive competencies, especially higher-order thinking—such as critical thinking, creativity, reasoning skills, problem-solving skills, and meta-cognitive abilities—has become a central goal and shared consensus of global education reform (OECD 2010; Partnership for 21st Century Skills 2009; Gordon et al. 2009; Voogt and Roblin 2012; Reed 2020). These competencies are fundamental to individual academic success and lifelong learning, and are also regarded as essential for building a well-functioning society (Zhang 2016).

Within this context, Philosophy for Children (P4C)—an educational program and pedagogy dedicated to fostering children's thinking skills and cognitive abilities (Lipman 2003; Yan et al. 2018) has garnered increasing global attention (Cam 2017; Gao 2018; De Cesaris 2018; Santi 2019; Michalik 2023). Rather than transmitting philosophical knowledge, P4C creates a

community of inquiry where children engage in open, collaborative dialogue around a shared philosophical question, facilitated by the teacher. Aiming to form reasonable, meaningful judgments, they challenge and build on one another's ideas, offer personal stories, attend to emotions, imagine new possibilities, and self-correct (Gregory 2021). Through Vygotsky's (1978) concept of the "intrapsychical reproduction of the interpsychical", children internalize the inquiry conducted in the community into the realm of their personal thinking (Lipman 2003; Kennedy and Kennedy 2011). P4C thus cultivates children's cognitive abilities, including reasoning skills, critical thinking, creativity, meta-cognitive abilities, problem-posing and problem-solving capacities, among others.

Given its theoretical promise, numerous empirical studies have investigated P4C's effects on cognitive outcomes. However, findings remain inconsistent, reflecting heterogeneity in design, participants, dosage, and measures. While many studies suggested that P4C enhance cognitive development (e.g., Iorio et al. 1984; Rahdar et al. 2018; Zulkifli and Hashim 2020), others reported inconclusive or non-significant results (e.g., Higa 1980; Meyer 1989; Ventista 2019; Lord et al. 2021). The inconsistency highlights the need for meta-analysis—a method that quantitatively synthesizes results from studies (Luo 2013)—to provide systematic evaluation of P4C's cognitive effects and their moderators.

Two prior meta-analyses (García-Morión et al. 2005; Yan et al. 2018) provided preliminary support for its positive impact. However, they included relatively few studies ($n = 18$ and 10 , respectively) and relied on conventional meta-analytic techniques, which treat all effect sizes as independent. Such an approach may ignore the statistical dependence among multiple effect sizes derived from the same sample, potentially biasing variance estimates and significance tests. Moreover, both reviews predate the recent surge of empirical research. Addressing these gaps, the present study conducts a three-level meta-analysis, which explicitly models within- and between-study variance while incorporating the expanded body of available evidence, thereby providing a more rigorous and comprehensive assessment of P4C's cognitive value.

Accordingly, this study aims to offer a comprehensive and methodologically robust assessment of P4C's cognitive impact through three-level meta-analysis. Specifically, we sought to address the following research questions: (1) Does P4C enhance children's cognitive development? (2) What methodological and substantive factors moderate its efficacy?

2. Literature Review and Hypotheses Development

2.1. *Philosophy for Children (P4C)*

P4C is an educational project initiated by American philosopher Matthew Lipman in 1969, with the aim of "teaching children to think for themselves and make informed choices" (Trickey and Topping 2004). In 1974, Lipman and Ann Sharp established the Institute for the Advancement of Philosophy for Children (IAPC) to advance P4C.

Globally, terms like "Philosophy with Children" and "Philosophy in Schools" are used interchangeably with P4C, referring to initiatives that involve youth in philosophical discussions (Gregory 2013). Today, P4C is recognized both as a standalone curriculum and as a pedagogy integrated into other subjects (Kennedy and Kennedy 2011; Lewis and Chandley 2012).

The typical Lipman-Sharp P4C approach involves six steps (Lipman 2003; Oyler 2017): (1) stimulus presentation: such as a picture book, story, or video, often one of the IAPC-developed philosophical novels; (2) question generation: children raise and democratically select questions for inquiry; (3) facilitated dialogue: conducting inquiry dialogues facilitated by the teacher; (4) deepening inquiry: through structured exercises and discussion plans; (5) meta-cognitive reflection: participants evaluate their own and the

community's cognitive, emotional, and moral performance; (6) further response: extending inquiry through art, action projects, etc.

2.2. The Cognitive Value of P4C

Cognitive abilities refer to mental processes involved in the acquisition of knowledge, manipulation of information, and reasoning, as opposed to other types of skills such as motor or social skills. Cognitive abilities encompass a wide spectrum of mental processes, ranging from basic functions such as perception, attention, and memory, to higher-order abilities such as reasoning, critical thinking, creativity, problem-solving, decision-making, and meta-cognition (Carroll 1993; Kiely 2014; Gu and Hu 2018).

P4C aims to develop children's critical, creative, and caring thinking (Lipman 2003), as well as their reasoning, moral understanding, and meaning-making capacities (Lipman et al. 1980). Over the past half-century, scholars worldwide have empirically tested P4C's efficacy on reasoning (Säre et al. 2016), critical thinking (Cooke 2015), creativity (Kanani Harandi et al. 2021), problem-posing (Zulkifli and Hashim 2019), problem-solving ability (Işıklar and Öztürk 2022), as well as domain-specific academic competencies such as reading comprehension (Imani et al. 2016), listening comprehension (Boyraz and Ayday 2024), and math learning (Sabbagh Hasanazadeh 2024). Thus, P4C targets multiple dimensions of cognitive development.

Previous meta-analyses also confirmed P4C's effectiveness in cognitive skills. García-Moriyón et al. (2005) synthesized 18 studies, reporting a medium effect size on reasoning. Yan et al. (2018) reviewed 10 studies, finding an overall moderate positive effect on cognition. These findings suggest that P4C exerts a robust and significant influence on children's cognitive abilities. Accordingly, we hypothesize the following:

H1. *P4C will significantly and positively affect children's cognitive abilities.*

2.3. Possible Moderators of P4C's Cognitive Effects

To comprehensively understand the effectiveness of P4C intervention, we examine potential moderators from two complementary perspectives: (1) the nature of the outcome (i.e., which specific cognitive abilities are most affected), and (2) the nature of the intervention (i.e., how variations in its delivery influence its efficacy). While these two types of moderators pertain to different parts of the causal chain (output vs. input), they are both essential for mapping the boundary conditions of P4C's effects.

2.3.1. Cognitive Ability Type

While P4C aims to cultivate a range of cognitive skills, it remains unclear whether its effects differ across cognitive domains. Systematic reviews emphasize P4C's particular strength in promoting higher-order thinking skills—such as reasoning, critical thinking, creativity, and question-posing (Ab Wahab et al. 2022). Many studies have reported significant gains in reasoning (e.g., Lipman 1976), critical thinking (e.g., Cooke 2015), and creativity (e.g., Pourtaghi et al. 2014). However, effects on domain-specific academic skills—such as reading or math achievement—are mixed. For instance, Imani et al. (2016) and Boyraz and Ayday (2024) found improvements in reading and listening comprehension, whereas Higa (1980) and Lord et al. (2021) reported no significant academic gains. Meta-analytic evidence further supports this variability across cognition domains. Yan et al. (2018) found that P4C produced a large effect on reasoning ($d = 1.06$), moderate effects on general cognition ($d = 0.40$), and smaller effects on reading comprehension ($d = 0.28$), with significant between-group differences ($Q = 15.44, p < 0.001$). Thus, the type of cognitive ability may moderate P4C's effects. Therefore, we propose:

H2. *The type of cognitive ability will significantly moderate the cognitive effects of P4C.*

2.3.2. Instructional Dosage

In the context of P4C, the intensity and duration of programs vary considerably, raising the question of whether these differences influence cognitive outcomes. Ventista (2018) notes that most P4C studies last less than one school year. Some studies suggest that short-term P4C interventions can generate cognitive gains (Lipman 1976; Pourtaghi et al. 2014; Cooke 2015; Imani et al. 2016; Işıklar and Öztürk 2022), confirmed by Yan et al. (2018), who found 5–20 h programs produce significant positive effects on students' cognitive outcomes. However, Fair et al. (2015b) found that a 22–26-week intervention could lead to significant cognitive gains, whereas a 4–10-week program did not, recommending implementing one 60-min P4C session per week across a single semester as an effective instructional dosage. These findings suggest that dosage-related variables (session length, frequency, and overall duration) may moderate the effect of P4C. Thus, we hypothesize the following:

H3.1. *The length of each P4C session will significantly moderate the cognitive effects of P4C.*

H3.2. *The frequency of P4C sessions will significantly moderate the cognitive effects of P4C.*

H3.3. *The overall duration of the P4C intervention will significantly moderate the cognitive effects of P4C.*

2.3.3. Grade Level

Whether children at different developmental stages could engage in and benefit equally from P4C is an important theoretical and empirical debate. Piaget (1933), Kitchener (1990), and White (1992) have argued that children below the formal operational stage lack abstract thinking skills required for philosophical inquiry. While contemporary developmental psychologists, such as Alison Gopnik, have argued that young children possess more advanced cognitive capacities than Piaget originally proposed, children are endowed with abilities to form theories, test hypotheses, and reason about causality and counterfactuals from a very early age (Gopnik 2009; Gopnik and Wellman 2012), they are “philosophical babies” (Gopnik 2009) and “scientists in the crib” (Gopnik et al. 1999). P4C scholars have also provided theoretical and practical evidence that even young children can do philosophy (Matthews 1980; Lipman 1990; Murris 1999). Empirical studies generally support this opinion. For example, Demirtaş et al. (2018) found that P4C enhanced preschoolers' ability to generate questions and give elaborated responses; Gardner (1998) observed moral reasoning gains in a K-5 longitudinal study. Meta-analytic evidence indicates consistent significant effects across grades 2–5 and 6–10 (Yan et al. 2018). Nevertheless, in order to rigorously address theoretical concerns about children's cognitive readiness raised in the literature, it remains important to examine whether grade level systematically moderates P4C's cognitive effects. To this end, we hypothesize:

H4. *Grade Level will significantly moderate the cognitive effects of P4C.*

2.3.4. Cultural Context

P4C originated in the USA and is now practiced in over 60 countries (Gregory 2021). Despite its global spread, P4C's cultural adaptability remains an open question. Grounded in western philosophical traditions and emphasizing inquiry-based, child-centered pedagogy, the program may not fully align with all educational systems in different contexts. Yan et al. (2018) found significantly ($Q = 5.16, p < 0.05$) larger effects in non-Western ($d = 0.69$) than in Western contexts ($d = 0.39$). Cultural context is commonly examined as a moderator in meta-analyses, with the Western vs. non-Western distinction being one of the common categorizations in cross-cultural research (Triandis and Gelfand 1998; Watkins

2014; Liu and Baumeister 2016). Accordingly, culture warrants investigation as a moderator, so we hypothesize:

H5. *Cultural context (Western vs. non-Western) will significantly moderate the cognitive effects of P4C.*

2.3.5. Research Design

Research design has been identified as a potential moderator. García-Moriyón et al. (2005) categorized studies into three types: independent groups pretest–posttest, independent groups post-test, and single group pretest–posttest, and found that research design had a significant moderating effect. In contrast, Yan et al. (2018) classified the included studies as randomized controlled trials (RCTs) or quasi-experimental and reported no significant differences. Given these mixed findings, the moderating role of research design remains worthy of further examination. Thus, we propose:

H6. *The type of research design will significantly moderate the cognitive effects of P4C.*

2.3.6. Publication Year

Publication year, as a temporal variable, has been widely considered a potential moderator in meta-analyses (Nakagawa et al. 2022). Time-lag bias—where studies with larger or statistically significant effects are published more quickly than smaller or non-significant ones—may lead to a decline in reported effect sizes over time (i.e., a decline effect) (Nakagawa et al. 2022; Koricheva and Kulinskaya 2019). García-Moriyón et al. (2005) found that more recent P4C studies tended to report smaller effect sizes, and explain that this might be due to the adoption of more rigorous research designs. Given these trends, this study investigates publication year as a moderator and proposes:

H7. *Publication year will significantly moderate the cognitive effects of P4C.*

2.3.7. Sample Size

Sample size is a critical methodological moderator in meta-analytic research. Smaller studies are frequently associated with larger effect sizes, a phenomenon termed the small-study effect. This may stem from publication bias (i.e., the preferential publication of statistically significant results) and greater random error in small samples (Luo et al. 2020). Small-scale trials may also benefit from more favorable implementation conditions such as more intensive teacher training and higher delivery consistency, potentially leading to larger observed effects (Yan et al. 2018). Examining sample size as a moderator is therefore essential for accurately interpreting heterogeneity and enhancing the robustness of the findings. Accordingly, we propose:

H8. *Sample size will significantly moderate the cognitive effects of P4C.*

3. Method

This study followed the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) guidelines (Page et al. 2021) and was preregistered on the Open Science Framework (OSF) (registration number: 10.17605/OSF.IO/EUAPX). Coded data and scripts are openly accessible on the OSF platform.

3.1. Search Procedures

The final comprehensive search was conducted on 13 March 2025 using Google Scholar, Web of Science (WOS), Scopus, and ProQuest. The searching keywords were (“philosophy

for children”) OR (“P4C”) OR (“philosophy with children”) OR (“P4wc”) OR (“philosophy in schools”) OR (“community of inquiry” AND “child*”) OR (“community of enquiry” AND “child*”) OR (“community of philosophical inquiry”) OR (“community of philosophical enquiry”) OR (“philosophical inquiry” AND “child*”) OR (“philosophical enquiry” AND “child*”). To minimize publication bias, both peer-reviewed journal articles and grey literature (conference papers and dissertations) were considered (Hopewell et al. 2005). All retrieved records were managed using Zotero. Duplicate records and non-English publications were removed prior to screening.

3.2. Eligibility Criteria and Study Selection

Potential studies were evaluated against the eligibility criteria to determine their suitability for the meta-analysis. The systematic review methodology involved establishing inclusion criteria based on the research topic, internationally recognized PICOS principles are typically adopted (Moher et al. 2009). Following the PICOS (Population, Intervention, Comparison, Outcomes, and Study design) principles, the inclusion criteria were: (1) the participants were children under 18 years old, consistent with *the United Nations Convention on the Rights of the Child* (United Nations General Assembly 1989); (2) the intervention group received the P4C curriculum or any course employing the pedagogy of P4C; (3) the control group received regular instruction without P4C components; (4) the eligible studies reported measurable outcomes regarding P4C’s impact on children’s cognitive performance, with sufficient statistics—such as sample sizes (N), means (M), standard deviations (SD), *t*-values or *p*-values—for effect size calculation; (5) studies adopted experimental or quasi-experimental designs.

Two researchers independently screened the literature through a four-stage process: (1) titles, abstracts, and keywords were reviewed to exclude records clearly ineligible based on PICOS criteria; (2) screening results were cross-verified, resolving discrepancies through discussion; (3) full texts were retrieved, and articles without accessible full texts were excluded; (4) eligible full texts underwent comprehensive review against inclusion criteria, followed by a second verification round. This dual-reviewer approach with consensus-building ensured rigor and minimized bias.

The PRISMA flowchart of the literature screening is shown in Figure 1. A total of 2579 articles were identified. After removing 710 duplicate records, 428 non-English publications, and 105 documents that were not journal articles, conference papers, or dissertations, 1336 records remained. Titles and abstracts screening excluded 1236 records; nine studies were removed due to unavailable full texts. After full-text review based on inclusion and exclusion criteria, 58 articles were further excluded. Ultimately, 33 studies were retained, providing 53 effect sizes from 4568 participants, spanning the period from 1979 to 2023.

3.3. Data Extraction, Feature Coding and Quality Assessment

To ensure rigorous and reliable data extraction, two researchers independently coded the included studies. When a single study reported multiple effect sizes, each was coded separately. Extracted information included: (1) basic study information, such as first author, publication year and relevant statistical data (e.g., N, M, SD, *t*-values or *p*-values); (2) potential moderator variables, including outcome indicators (for studies reporting multiple cognitive outcomes, all relevant measures were systematically recorded), instructional dosage (session length, frequency, and total duration), grade level, country, and research design. Missing study characteristics were coded as “Not Reported” when the relevant information was unavailable in the original publications.

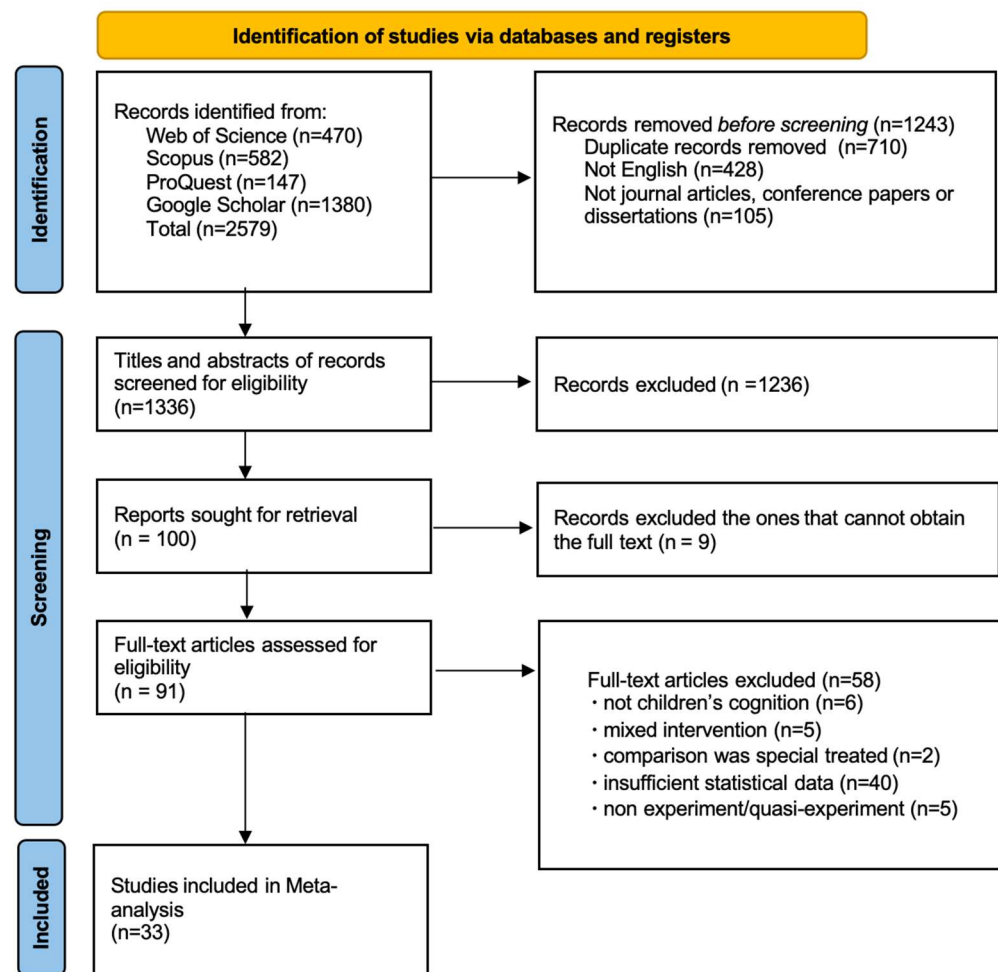


Figure 1. PRISMA Flow Diagram of Literature Screening.

Coding criteria were based on both theoretical and practical considerations. Following Piaget (1933) and subsequent debates (Kitchener 1990; White 1992) regarding children's capacity for philosophical inquiry prior to the formal operational stage (typically younger than 11–12 years, or below Grade 6), and consistent with Yan et al. (2018), grade levels were coded as $\leq$ Grade 5 vs. $\geq$ Grade 6.

Cultural context was coded as Western vs. Non-Western at the country level. Following prior meta-analyses and cross-cultural studies (Yan et al. 2018; Triandis and Gelfand 1998; Watkins 2014; Liu and Baumeister 2016), we classified studies as Western (USA, UK, Australia, Iceland) or non-Western (China, South Korea, Malaysia, Iran, Turkey). Western contexts are generally higher in individualism and lower in power distance, whereas non-Western contexts tend to be more collectivist and hierarchical (Hofstede 1984; Hofstede et al. 2010; Li 2012). We acknowledge that this distinction is a heuristic simplification, but it provides a pragmatic framework widely used in cross-cultural research.

Session length followed the widely adopted guideline that a standard class period lasts 40–50 min (National Health Commission of the People's Republic of China 2012), and categorized as $\leq$ 50 min vs. $>$ 50 min. Considering a typical school semester spans approximately 20 weeks, the total duration of intervention was categorized into two groups: $\leq$ 20 weeks vs. $>$ 20 weeks. Research design was coded as randomized controlled trial (RCTs), non-randomized controlled trials (Non-RCT), or one-group pretest–posttest design.

Outcome domains were classified according to the definitions provided by the original authors and the characteristics of the instruments employed. For example, reasoning skills were frequently assessed using the New Jersey Test of Reasoning Skills (NJTRS) (Lam 2012;

Marashi 2008; Pålsson 1995); critical thinking was often measured with standardized instruments such as the Cornell Critical Thinking Test (CCTT) or the California Critical Thinking Skills Test (CCTST) (Erfani and Rezaei 2016; Acar and Arslan 2023); creativity was evaluated using the Torrance Test of Creativity, as well as the “What Could You Use It For?” and “What Could It Be?” tasks (Pourtaghi et al. 2014; Cinquino 1981). General cognitive abilities were assessed with the Cognitive Abilities Test (CAT) (Fair et al. 2015a, 2015b; Topping and Trickey 2007a, 2007b), while reading comprehension was measured using a variety of instruments, including the General English Proficiency Test (Tian and Liao 2016), the Test of Reading Comprehension (TORCH) (Youssef et al. 2016), TOEFL (Othman and Hashim 2006), the reading comprehension subtest of the Comprehensive Test of Basic Skills (CTBS) (Yeazell 1982), and researcher-designed Reading Comprehension Tests (Tok and Mazl 2015).

The researchers subsequently cross-validated their coding, resolving discrepancies through discussion. Independent double coding by two researchers ensured consistent and reliable data extraction. The resulting inter-coder reliability coefficient ($Kappa = 0.93$) indicated an excellent level of agreement.

The methodological quality of the included studies was evaluated using the *Quality Assessment Tool for Observational Cohort and Cross-Sectional Studies* provided by the National Institutes of Health (NIH 2021). This tool consists of 14 items, each with five possible responses: Yes, No, Cannot Determine (CD), Not Reported (NR), and Not Applicable (NA). Scoring was performed by assigning 1 point for “Yes” and 0 points for all other responses. Studies were then classified as “good” (total score >7), “fair” (5–7), or “poor” (<5) (Zhu et al. 2024). Two authors assessed quality independently, and the inter-rater agreement was $Kappa = 0.78$.

3.4. Data Analysis and Synthesis

3.4.1. Model Selection

Meta-analysis can be conducted with fixed-effects or random-effects models. The fixed-effects model assumes that all studies estimate a single true effect size, with variation across studies attributed solely to sampling error. However, given that the included studies differed in participant populations, cultural contexts, intervention formats, and outcome measures, it was more appropriate to adopt a random-effects model, which allows the true effect size to vary across studies (Borenstein et al. 2009). Additionally, further heterogeneity tests will be conducted to validate model choice during the subsequent analysis of the meta-analytic data.

Moreover, several primary studies included in this meta-analysis reported multiple effect sizes from the same sample, creating statistical dependence. Traditional meta-analytic methods assume independence among effect sizes and typically extract only one effect size per study (Assink and Wibbelink 2016). This approach ignores the inherent dependence, which may lead to an overestimation of the overall effect size (Lipsey and Wilson 2001). In contrast, a three-level meta-analytic model accounts for the dependency of effect sizes within the same study by partitioning the variance into three levels. Level 1 represents sampling error arising from the selection of participants within individual studies; Level 2 reflects variance among multiple effect sizes within the same study, indicating within-study heterogeneity; Level 3 captures variance among effect sizes from different studies, indicating between-study heterogeneity (Cheung 2014). Compared with traditional meta-analytic techniques, the three-level model addresses the issue of dependent effect sizes while retaining more information and increasing statistical power (Assink and Wibbelink 2016). For these reasons, the present study employed a three-level random effects model to examine the main effects, heterogeneity, moderator effects, and publication bias.

3.4.2. Effect Size Calculation

Given that some included studies had relatively small sample sizes, Hedges' g with 95% confidence intervals (CIs) was chosen as the measure of effect size. Compared with Cohen's d , g includes a correction factor that reduces bias in small samples (Hedges 1981), while they yield nearly identical results in large-sample studies, making it more accurate under such conditions (Zheng et al. 2011). The interpretation of effect sizes followed Cohen's conventional benchmarks: effect sizes less than 0.2 were considered negligible, those between 0.2 and 0.5 indicated a small effect, those between 0.5 and 0.8 indicated a medium-to-large effect, and values above 0.8 indicated a large effect (Cohen 2013).

3.4.3. Heterogeneity Analysis and Moderator Analysis

The present study used one-tailed log-likelihood ratio tests to examine the variance at Level 2 and Level 3, determining whether these components are significant. If significant, moderator analyses were conducted to explore potential sources of heterogeneity (Gao et al. 2024). Moderator variables were incorporated as covariates in the three-level meta-analytic model to estimate the magnitude of their moderating effects (Gao et al. 2024). The moderator variables considered included (1) categorical moderators: cognitive ability type, dosage variables (duration, frequency, and weeks), grade level, cultural context, stimulus type, and research design and (2) continuous moderators: publication year and sample size.

3.4.4. Publication Bias Assessment

Publication bias refers to the phenomenon that studies reporting statistically significant findings are more likely to be published (Franco et al. 2014). In this study, publication bias is assessed qualitatively and quantitatively using funnel plots, Rosenthal's Fail-Safe N test, and the Trim-and-Fill method. For qualitative assessment, a symmetric funnel plot indicates low risk of publication bias (Sterne and Harbord 2004). Fail-Safe N is calculated using Rosenthal's method, with values exceeding $5 \times k + 10$ (k = the number of effect sizes) suggesting no obvious publication bias, whereas smaller values indicate potential bias (Rosenthal 1979). Next, the Trim-and-Fill analysis for multilevel models is used to estimate the missing studies that could make the funnel plot symmetrical (Duval and Tweedie 2000; Fernández-Castilla et al. 2021). If $R_0^+ > 3$ and $L_0^+ > 2$ in the Trim-and-Fill analysis, publication bias existed (Fernández-Castilla et al. 2021).

3.4.5. Statistical Procedures

All analyses are performed in R 4.2.0 using the metafor package (Viechtbauer 2010). R code is based on the procedures designed by Assink and Wibbelink (2016) and Rodgers and Pustejovsky (2021). Model parameters estimated using restricted maximum likelihood (REML) (Viechtbauer 2010), and statistical significance is determined at a two-tailed $p < 0.05$ level.

4. Results

4.1. Study Characteristics

The three-level meta-analysis of P4C's cognitive effects included 33 studies with 53 effect sizes, encompassing 4568 participants from 1979 to 2023. Among the included studies, two were derived from doctoral dissertations, while the remainder were peer-reviewed journal articles. The basic characteristics of the included studies are presented in Table 1. A more detailed coding table listing the specific measurement tools used in each study has been made available on the OSF preregistration platform (<https://doi.org/10.17605/OSF.IO/EUAPX>, accessed on 1 October 2025). The quality assessment results indicated that the included studies were rated as either "good" ($n = 31$) or "fair" ($n = 2$).

Table 1. Basic information of the included studies.

ID	Year	Study (First Author)	Sample Size	Grade	Country	Dosage	Design	Outcome Indicators
						Duration/Frequency/Weeks		
1	2023	Acar (Acar and Arslan 2023)	23	S	Turkey	40 min/2 times/10 weeks	NR	CT, SK
2	2023	Şişman (Şişman et al. 2023)	34	S	Turkey	80 min/1 time/8 weeks	NR	MC
3	2022	Akbayir (Akbayir and Tedikçi 2022)	38	S	Turkey	120 min/1 time/8 weeks	NR	MA
4	2022	Işıklar (Işıklar and Öztürk 2022)	40	L	Turkey	40 min/2 times/10 weeks	NR	CT, PS
5	2021	Wu (Wu 2021)	173	S	China	40 min/2 times/4 weeks	R	CT
6	2020	Zulkifli (Zulkifli and Hashim 2020)	61	S	Malaysia	—/—/11 sessions	NR	CT
7	2020	Khanmohammadi (Khanmohammadi et al. 2020)	44	S	Iran	—/1 time/12 weeks	NR	LO, PM
8	2020	Mehnehj (Mehnehj et al. 2020)	50	S	Iran	60 min/2 time/6 weeks	NR	RS
9	2019	Ventista (Ventista 2019)	738	L	UK	30 min/1 time/40 weeks	NR	CT, C
10	2017	Abbasi (Abbasi et al. 2017)	50	L	Iran	—/—/12 weeks	NR	CT
11	2016	Tian (Tian and Liao 2016)	62	S	China	100 min/1 time/10 weeks	NR	RC
12	2016	Youssef (Youssef et al. 2016)	246	S	Australia	—/—/24 weeks	NR	RC
13	2016	Erfani (Erfani and Rezaei 2016)	40	S	Iran	120 min/1 time/12 weeks	NR	CT
14	2015	Fair (a) (Fair et al. 2015b)	275	S	USA	60 min/1 time/22–26, 4–10 weeks	R	CA
15	2015	Fair (b) (Fair et al. 2015a)	183	S	USA	60 min/1 time/22–26 weeks	R	CA
16	2015	Tok (Tok and Mazl 2015)	74	L	Turkey	120 min/2 time/10 weeks	NR	RC, LC
17	2014	Pourtaghi (Pourtaghi et al. 2014)	32	S	Iran	75 min/12 weeks	NR	C
18	2012	Lam (Lam 2012)	28	S	China	90 min/2 times/16 weeks	R	RS
19	2009	Marashi (Marashi 2008)	60	S	Iran	70 min/11 sessions	NR	RS
20	2007	Topping (a) (Topping and Trickey 2007a)	115	L	UK	60 min/1 time/58 weeks	NR	CA
21	2007	Topping (b) (Topping and Trickey 2007b)	177	L	UK	60 min/1 time/58 weeks	NR	CA
22	2006	Othman (Othman and Hashim 2006)	45	S	Malaysia	—/—/16 weeks	NR	RS, RC
23	2000	Jo (Jo 2000)	54	L	South Korea	30 min/4 times/24 weeks	NR	CM
24	1998	Sprod (Sprod 1998)	54	S	UK	70 min/1 time/40 weeks	NR	RS
25	1995	Pálsson (Pálsson 1995)	126	L	Iceland	80 min/1 time/24 weeks	NR	RS
26	1993	Chamberlain (Chamberlain 1993)	160	L	USA	60 min/5 times/12 weeks	NR	RS, HC
27	1989	Slade (Slade 1989)	50	S	Australia	120 min/12 sessions	NR	RS
28	1986	Jenkis (Jenkins 1986)	60	S	UK	45 min/about 20 weeks	NR	RS
29	1985	Martin (Martin and Weinstein 1985)	964	L, S	USA	—/—/about 32 weeks	PP	CT
30	1984	Iorio (Iorio et al. 1984)	336	L	USA	45 min/2 times/40 weeks	PP	CT
31	1982	Yeazell (Yeazell 1982)	100	L	USA	45 min/1 time/32 weeks	NR	RC
32	1981	Cinquino (Cinquino 1981)	47	L	USA	120 min/1 time/28 weeks	PP	RS, C
33	1979	Cummings (Cummings 1979)	29	L	USA	40 min/2 times/7.5weeks	NR	LT

Notes: —, not reported. Grade: L, Lower grade ($\leq$ Grade 5); S, Senior grade ($\geq$ Grade 6). Stimulus: I, IAPC-developed stimulus; N-I, Non-IAPC developed stimulus. Research Design: R, Randomized controlled trial; NR, Non-randomized controlled trial; PP, One-group pretest–posttest design. Outcome Indicators: C, Creativity; CA, Cognitive Abilities; CT, Critical Thinking; CM, Constructing Meaning; HC, Higher Cognitive Processes; L, Logic; LC, Listening Comprehension; LO, Learning Outcome; MA, Math Achievement; MC, Meta-Cognition; PM, Philosophical Mentality; PS, Problem Solving skills; RC, Reading Comprehension; RS, Reasoning Skills; SK, Speaking Skills.

ID	Study Authors-Year	Sample Size	Grade	Country	Dosage	Design	Outcome Indicators
					Duration/Frequency/Weeks		
1	(Acar and Arslan 2023)	23	S	Turkey	40 min/2 times/10 weeks	NR	CT, SK
2	(Şişman et al. 2023)	34	S	Turkey	80 min/1 time/8 weeks	NR	MC
3	(Akbayir and Tedikçi 2022)	38	S	Turkey	120 min/1 time/8 weeks	NR	MA
4	(Işıklar and Öztürk 2022)	40	L	Turkey	40 min/2 times/10 weeks	NR	CT, PS
5	(Wu 2021)	173	S	China	40 min/2 times/4 weeks	R	CT
6	(Zulkifli and Hashim 2020)	61	S	Malaysia	—/—/11 sessions	NR	CT
7	(Khanmohammadi et al. 2020)	44	S	Iran	—/1 time/12 weeks	NR	LO, PM
8	(Mehnehj et al. 2020)	50	S	Iran	60 min/2 time/6 weeks	NR	RS
9	(Ventista 2019)	738	L	UK	30 min/1 time/40 weeks	NR	CT, C
10	(Abbasi et al. 2017)	50	L	Iran	—/—/12 weeks	NR	CT
11	(Tian and Liao 2016)	62	S	China	100 min/1 time/10 weeks	NR	RC
12	(Youssef et al. 2016)	246	S	Australia	—/—/24 weeks	NR	RC
13	(Erfani and Rezaei 2016)	40	S	Iran	120 min/1 time/12 weeks	NR	CT
14	(Fair et al. 2015b)	275	S	USA	60 min/1 time/22–26, 4–10 weeks	R	CA
15	(Fair et al. 2015a)	183	S	USA	60 min/1 time/22–26 weeks	R	CA
16	(Tok and Mazl 2015)	74	L	Turkey	120 min/2 time/10 weeks	NR	RC, LC
17	(Pourtaghi et al. 2014)	32	S	Iran	75 min/12 weeks	NR	C
18	(Lam 2012)	28	S	China	90 min/2 times/16 weeks	R	RS
19	(Marashi 2008)	60	S	Iran	70 min/11 sessions	NR	RS
20	(Topping and Trickey 2007a)	115	L	UK	60 min/1 time/58 weeks	NR	CA
21	(Topping and Trickey 2007b)	177	L	UK	60 min/1 time/58 weeks	NR	CA
22	(Othman and Hashim 2006)	45	S	Malaysia	—/—/16 weeks	NR	RS, RC
23	(Jo 2000)	54	L	South Korea	30 min/4 times/24 weeks	NR	CM
24	(Sprod 1998)	54	S	UK	70 min/1 time/40 weeks	NR	RS
25	(Pálsson 1995)	126	L	Iceland	80 min/1 time/24 weeks	NR	RS
26	(Chamberlain 1993)	160	L	USA	60 min/5 times/12 weeks	NR	RS, HC
27	(Slade 1989)	50	S	Australia	120 min/12 sessions	NR	RS

28	(Jenkins 1986)	60	S	UK	45 min/about 20 weeks	NR	RS
29	(Martin and Weinstein 1985)	964	L, S	USA	—/—/about 32 weeks	PP	CT
30	(Iorio et al. 1984)	336	L	USA	45 min/2 times/40 weeks	PP	CT
31	(Yeazell 1982)	100	L	USA	45 min/1 time/32 weeks	NR	RC
32	(Cinquino 1981)	47	L	USA	120 min/1 time/28 weeks	PP	RS, C
33	(Cummings 1979)	29	L	USA	40 min/2 times/7.5weeks	NR	LT

4.2. Assessment of Publication Bias

Figure 2 shows the funnel plot of the 53 effect sizes, with standard error (SE) of the g on the vertical axis and the g values on the horizontal axis. Each black circle represents an individual effect size, plotted according to its estimated magnitude (g) and corresponding standard error. The plot indicates that the majority of effect sizes are evenly and symmetrically distributed around the mean, suggesting a low risk of publication bias and supporting the robustness of the findings.

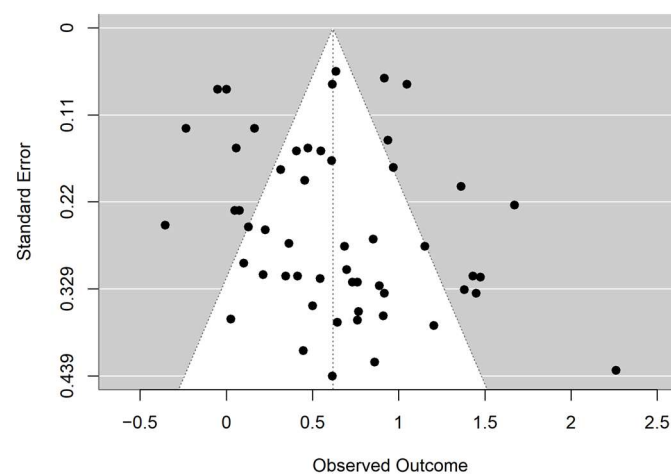


Figure 2. Funnel Plot.

Rosenthal's Fail-Safe N indicated that 6790 missing or unpublished studies with null results would be required to cancel the overall effect size to non-significance. This number far exceeds $5 \times k + 10$ (in this case, $k = 53$, the result would be 275), further suggesting a low risk of publication bias and supporting the robustness of the findings.

Moreover, the Trim-and-Fill method yielded $R_0^+ = 1$ (<2) and $L_0^+ = 0.44$ (<3), indicating that the impact of publication bias was minimal and could be considered negligible.

4.3. Heterogeneity Test Results and Overall Effect

To examine the significance of the within-study variance (Level 2) and the between-study variance (Level 3), we conducted two fit tests comparing the fit of the three-level model (which included both within study and between-study variances) to the fit of two two-level models (one including only within-study variance, and the other including only between-study variance), respectively. Results are shown in Table 2. The proportion of total variance attributable to sampling variance (Level 1) was 11.47%, within-study variance (Level 2) accounted for 41.40%, and between-study variance (Level 3) for 47.13%. One-tailed log-likelihood ratio tests indicated that both Level 2 variance ($p < 0.001$) and Level 3 variance ($p < 0.05$) were significant, supporting the presence of substantial heterogeneity, confirming the appropriateness of employing a three-level random-effects model (Assink and Wibbelink 2016; Hunter and Schmidt 1990).

Findings revealed a moderate-to-large overall effect size ($g = 0.59$, 95% CI = [0.43, 0.75]), and the forest plot (Figure 3) illustrates the individual and pooled effect sizes.

Table 2. Results of Heterogeneity Test and Overall Effects.

Model	<i>k</i>	# <i>es</i>	<i>N</i>	<i>t</i>	Hedges' <i>g</i>	95%CI	%Var. at Level 1	%Var. at Level 2	%Var. at Level 3
REM	33	53	4568	7.20 ***	0.59	[0.43, 0.75]	11.47	41.40	47.13

Notes: REM, Random Effect Model; *k*, number of studies; #*es*, number of effect sizes; *N*, sample size; *t*, *t* test value for the difference between mean effect size and 0; Hedges' *g*, effect sizes; CI, confidence interval; % Var, percentage of variance that is distributed at one of the three levels of the meta-analytic model. Level 1, sample variance; Level 2, variance between effect sizes from the same study; Level 3, variance between studies; ***, $p < 0.001$. The same below.

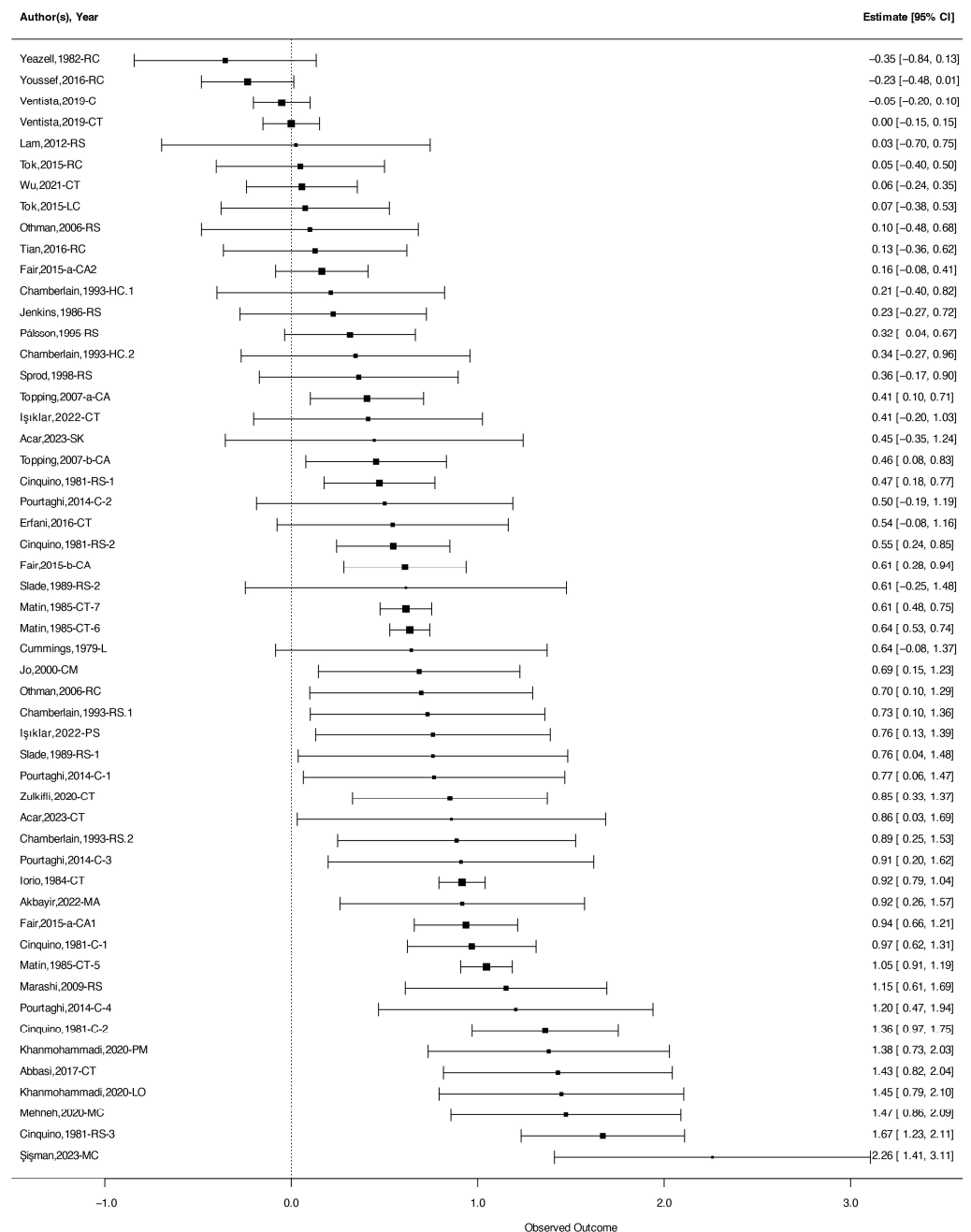


Figure 3. Forest plot based on a random-effect model displaying effect sizes with 95% confidence intervals. Notes: The studies included in the forest plot are as follows: Yeazell, 1982-RC (Yeazell 1982); Youssef, 2016-RC (Youssef et al. 2016); Ventista, 2019-C (Ventista 2019); Ventista, 2019-CT (Ventista 2019);

Lam,2012-RS (Lam 2012); Tok,2015-RC (Tok and Mazl 2015); Othman,2006-RS (Othman and Hashim 2006); Tian,2016-RC (Tian and Liao 2016); Fair,2015-a-CA2 (Fair et al. 2015b); Chamberlain,1993-HC (Chamberlain 1993); Jenkins,1986-RS (Jenkins 1986); Pålsson,1995-RS (Pålsson 1995); Chamberlain,1993-HC (Chamberlain 1993); Sprod,1998-RS (Sprod 1998); Topping,2007-a-CA (Topping and Trickey 2007a); Işıklar,2022-CT (Işıklar and Öztürk 2022); Acar,2023-SK (Acar and Arslan 2023); Topping,2007-b-CA (Topping and Trickey 2007b); Cinquino,1981-RS-1 (Cinquino 1981); Pourtaghi,2014-C-2 (Pourtaghi et al. 2014); Erfani,2016-CT (Erfani and Rezaei 2016); Cinquino,1981-RS-2 (Cinquino 1981); Fair,2015-b-CA (Fair et al. 2015a); Slade,1989-RS-2 (Slade 1989); Matin,1985-CT-7 (Martin and Weinstein 1985); Matin,1985-CT-6 (Martin and Weinstein 1985); Cummings,1979-L (Cummings 1979); Jo,2000-CM (Jo 2000); Othman,2006-RC (Othman and Hashim 2006); Chamberlain,1993-RS (Chamberlain 1993); Işıklar,2022-PS (Işıklar and Öztürk 2022); Slade,1989-RS-1 (Slade 1989); Pourtaghi,2014-C-1 (Pourtaghi et al. 2014); Zulkifli,2020-CT (Zulkifli and Hashim 2020); Acar,2023-CT (Acar and Arslan 2023); Chamberlain,1993-RS (Chamberlain 1993); Pourtaghi,2014-C-3 (Pourtaghi et al. 2014); Iorio,1984-CT (Iorio et al. 1984); Akbayir,2022-MA (Akbayir and Tedikçi 2022); Fair,2015-a-CA1 (Fair et al. 2015b); Cinquino,1981-C-1 (Cinquino 1981); Matin,1985-CT-5 (Martin and Weinstein 1985); Marashi,2009-RS (Marashi 2008); Pourtaghi,2014-C-4 (Pourtaghi et al. 2014); Cinquino,1981-C-2 (Cinquino 1981); Khanmohammadi,2020-PM (Khanmohammadi et al. 2020); Abbasi,2017-CT (Abbasi et al. 2017); Khanmohammadi,2020-LO (Khanmohammadi et al. 2020); Mehneh,2020-MC (Mehneh et al. 2020); Cinquino,1981-RS-3 (Cinquino 1981); Şişman,2023-MC (Şişman et al. 2023).

4.4. Moderator Analyses

Subgroup analyses were conducted to examine potential moderators, including cognitive ability type (creativity, cognitive abilities, critical thinking, reading comprehension, reasoning skills, and others), session length (≤ 50 min vs. > 50 min), session frequency (once per week vs. twice per week), total duration (≤ 20 weeks, > 20 weeks), grade level ($\leq$ Grade 5 vs. $\geq$ Grade 6), cultural context (Western vs. non-Western), and research design (RCT, Non-RCT controlled trial, and One-group pretest–posttest). Meta-regression tested publication year and sample size. Results are shown in Table 3.

Table 3. Results of Moderator Analyses.

		<i>k</i>	<i>#es</i>	<i>g</i>	95% CIs	<i>t</i> ₀	β_1	95% CIs	<i>t</i> ₁	<i>F</i> (<i>df</i> ₁ , <i>df</i> ₂)
Cognitive Domains	Creativity	3	7	0.73	[0.33, 1.13]	3.66 ***				<i>F</i> (5, 47) = 1.91 <i>p</i> = 0.11
	Cognitive Abilities	4	5	0.51	[0.10, 0.92]	2.52 *	−0.21	[−0.56, 0.41]	−0.31	
	Critical Thinking	9	11	0.66	[0.37, 0.95]	4.54 ***	−0.07	[−0.53, 0.38]	−0.32	
	Reading Comprehension	5	5	0.08	[−0.33, 0.50]	0.40	−0.65	[−1.23, −0.07]	−2.26 *	
	Reasoning Skills	9	13	0.56	[0.27, 0.85]	3.94 ***	−0.17	[−0.63, 0.29]	−0.75	
	Others	10	12	0.82	[0.52, 1.13]	5.40 ***	0.09	[−0.41, 0.60]	−0.37	
Dosages	Session Length	≤ 50 min	9	12	0.36	[0.05, 0.67]	2.30 *			<i>F</i> (2, 50) = 1.57
		> 50 min	18	31	0.65	[0.45, 0.87]	5.97 ***	0.29	[−0.09, 0.67]	
		Not reported	6	10	0.74	[0.37, 1.10]	4.06 ***	0.38	[−0.10, 0.86]	
	Weekly Frequency	1 time	14	21	0.58	[0.31, 0.84]	4.34 ***			<i>F</i> (3, 49) = 0.10
		2 times	8	11	0.53	[0.17, 0.89]	2.92 *	−0.05	[−0.50, 0.40]	
		others	2	5	0.59	[−0.08, 1.27]	1.26	0.02	[−0.70, 0.74]	
		Not reported	9	16	0.66	[0.33, 0.99]	4.04 ***	0.08	[−0.34, 0.51]	
	Intervention Duration	≤ 10 weeks	18	28	0.63	[0.40, 0.86]	5.42 ***			<i>F</i> (2, 50) = 0.90
		> 20 weeks	13	21	0.49	[0.25, 0.74]	4.12 ***	−0.14	[−0.46, 0.19]	
		Not reported	3	4	0.89	[0.30, 1.47]	3.03 **	0.26	[−0.38, 0.89]	
Grade	$\leq$ Grade 5	13	19	0.49	[0.25, 0.74]	4.00 ***				<i>F</i> (2, 50) = 1.02
	$\geq$ Grade 6	20	29	0.62	[0.41, 0.83]	5.96 ***	0.13	[−0.19, 0.44]	0.80	
	Mixed	1	5	0.97	[0.31, 1.64]	2.93 **	0.48	[−0.23, 1.19]	1.35	
Culture Context	Western	17	28	0.46	[0.23, 0.68]	4.10 ***				<i>F</i> (1, 51) = 2.96 <i>p</i> = 0.09
	non-Western	16	25	0.74	[0.50, 0.98]	6.17 ***	0.28	[−0.05, 0.61]	1.72	
Type of Study Design	RCT	4	5	0.36	[−0.11, 0.82]	1.55				<i>F</i> (2, 50) = 1.44
	Non-RCT	26	39	0.57	[0.39, 0.76]	6.17 ***	0.22	[−0.28, 0.72]	0.88	
	One-group pre-post	3	9	0.88	[0.45, 1.32]	4.06 ***	0.53	[−0.11, 1.16]	1.67	
Year		33	53				0.002	[−0.01, 0.01]	0.38	<i>F</i> (1, 51) = 0.15
Sample Size		33	53				−0.001	[−0.002, −0.000]	−2.02 *	<i>F</i> (1, 51) = 4.07 *

Notes: *t*₀, *t* test value for the difference between mean effect size and 0. β_1 , estimated regression coefficient; *t*₁, *t* test value for the difference between mean effect size and reference category; *F*(*df*₁, *df*₂), the result of the omnibus test; *, *p* < 0.05; **, *p* < 0.01; ***, *p* < 0.001.

Moderator analyses indicated that cognitive ability type [$F(5, 47) = 1.91, p = 0.11$], session length [$F(2, 50) = 1.57, p = 0.22$], weekly frequency [$F(3, 49) = 0.10, p = 0.96$], total duration [$F(2, 50) = 0.90, p = 0.41$], grade level [$F(2, 50) = 1.02, p = 0.37$], research design [$F(2, 50) = 1.44, p = 0.247$], and publication year [$F(1, 51) = 0.15, p = 0.704$] were all non-significant moderators.

One moderator was marginally significant: cultural context [$F(1, 51) = 2.96, p = 0.09$]. Effect sizes were larger in non-Western contexts ($g = 0.74$, 95% CI [0.50, 0.98]) than in Western contexts ($g = 0.46$, 95% CI [0.23, 0.68]).

Sample size emerged as a significant moderator [$F(1, 51) = 4.07, p = 0.049$], indicating that the cognitive effects of P4C decreased slightly as sample size increased ($\beta = -0.001, p = 0.049$).

5. Discussion

5.1. Overall Effects of P4C on Children's Cognitive Abilities

To our knowledge, this study represents the first application of a three-level meta-analytic approach to systematically evaluate the impact of P4C on children's cognitive abilities. The results indicate a moderate-to-strong and significant positive effect ($g = 0.59$, 95% CI = [0.43, 0.75], $p < 0.001$), confirming *H1*. This finding aligns robustly with prior empirical evidence (e.g., [Acar and Arslan 2023](#); [Jo 2000](#); [Pourtaghi et al. 2014](#); [Zulkifli and Hashim 2020](#)).

P4C's significant cognitive benefits can be effectively explained by its two foundational principles—introducing philosophy to children and the pedagogy of the Community of Inquiry (COI), both pioneered by Lipman and Sharp.

Bringing philosophy to children does not mean transmitting philosophical knowledge. Rather, it entails respecting and cultivating children's innate philosophical curiosity, leveraging the intrinsic consonance between philosophy and sound thinking ([Lipman 2003](#)), and through philosophically rich stimulus, children are encouraged to raise and persistently explore questions of genuine interest, thereby stimulating sustained cognitive engagement.

The COI framework is deeply rooted in a rich tapestry of theoretical traditions, including Socratic dialogue, John Dewey's epistemology of inquiry and political philosophy, Charles Sanders Peirce's theory of inquiry, Justus Buchler's theory of judgment, and the social constructivist perspectives of George Herbert Mead and Lev Vygotsky ([Lipman 2003](#); [Yarmel and Gregory 2025](#)). Within a COI, children are encouraged to express their viewpoints on issues of common concern in a climate of respect and attentive listening. Crucially, each perspective is subject to collective critical scrutiny and rational evaluation. This ongoing dialectic provokes a wealth of cognitive activities, including conceptual definition, reasoning, exemplification and counter-exemplification, analogy-making, hypothesis identification, inference, the exploration of alternative possibilities, and so on. Through such dynamic social interaction, participants not only co-construct more reasonable and meaningful judgments, but also enhance their individual cognitive capacities and inquiry skills through communal thinking and inquiry. The latter is realized through what [Vygotsky \(1978\)](#) termed "internalization", whereby higher mental functions are transformed from socially mediated interactions (interpsychical) into individualized cognitive functions (intrapsychical) ([Lipman 2003](#); [Kennedy 2013](#); [Liu 2020](#); [Gregory 2021](#); [Yarmel and Gregory 2025](#)).

Notably, the effect size obtained in this study ($g = 0.59$) was slightly larger than [Yan et al. \(2018\)](#)'s earlier meta-analysis ($d = 0.43$, 95% CI = [0.33, 0.53], $p < 0.001$). This divergence may be attributable to methodological advancements in the present study, including broader search strategies, access to more comprehensive databases, and an extended time span. These enhancements likely resulted in the inclusion of a larger, more diverse, and potentially more representative sample of primary studies. Crucially,

the persistence of substantial effect underscores P4C's enduring effectiveness over time. Despite evolving educational landscapes and methodological variations in research designs, contemporary studies continue to affirm the cognitive benefits of P4C.

5.2. Moderating Factors Influencing the Cognitive Effects of P4C

5.2.1. Moderating Effect of Cognitive Domains

We found that the type of cognitive domain assessed did not significantly moderate the overall effect, failing to support *H2*. However, notable variations emerged across domains: P4C had significant positive effects on creativity, critical thinking, reasoning, and general cognitive abilities, but its effect on reading comprehension was small and non-significant.

Findings highlight P4C's robust and significant positive effect on creativity ($g = 0.72$), a dimension unexplored in prior meta-analyses. This finding confirms theoretical expectations (Lipman 2003; Fisher 2007) and empirical studies (Lam 2021; Camhy and Iberer 1988; Naderi et al. 2012; Ghaedi et al. 2015). P4C cultivates a safe and open environment that values curiosity, questioning, and diverse perspectives over fixed answers. Through prompts like "Who has a new idea?" and "Could it be...?", and by supporting children to express their thoughts through multiple modes including language, imagery, and embodied forms, P4C enables children to "think outside the box" (Lam 2021), thereby fostering their ability to make connections, generate novel ideas, explore multiple possibilities, and propose alternative solutions, ultimately promoting the multidimensional development of creativity (Fisher 2008; Akan and Çüçen 2023; Stokell et al. 2017).

The effect of P4C on critical thinking was moderate to large and significant ($g = 0.66$), in line with theoretical claims (Lipman 2003; Fisher 2008) and empirical evidence (Erfani and Rezaei 2016; Işıklar and Öztürk 2022). P4C also exerted a moderate to large effect on reasoning skills ($g = 0.56$). This aligns with Lipman (2003)'s theoretical expectations and earlier empirical findings (Lipman 1976; Karras 1979; Gasparatou and Kampeza 2012). This effect size is notably lower than Yan et al. (2018)'s earlier estimate ($d = 1.02$), which drew from two studies, whereas the present analysis is based on 13 effect sizes yielding from 9 studies. Notably, our result closely matches García-Moriyón et al. (2005)'s meta-analysis spanning 1976–2002 ($d = 0.58$), suggesting the reasoning benefit of P4C is robust. In P4C, facilitators and peers engage children by asking for questions and reasons, requesting examples or counterexamples, probing sources of information, examining assumptions and implications, identifying fallacies, and appealing to standards. These cognitive moves help children develop and internalize reasoning and critical thinking skills such as questioning, accommodating dissent, and self-correction, while also nurturing the intellectual virtues essential to critical thinking, including curiosity, fairness, open-mindedness, respect for others, and rationality (Lipman 2003; Fisher 2008; Shorer and Quinn 2023).

P4C also yielded a moderate, significant improvement in general cognitive ability ($g = 0.51$), mirroring the results of a ten-year longitudinal study by Colom et al. (2014), which reported an effect size of 0.44 for overall cognitive ability.

Specifically, P4C had a small and not significant effect on reading comprehension. Two possible explanations may account for this finding. First, P4C does not directly target reading skills, and improvements in reasoning may not immediately transfer to reading comprehension. As Ventista (2019) has suggested, "long-term implementation of P4C may positively influence academic performance". Further large-scale longitudinal studies are needed to establish its effects on academic achievement. Second, the sensitivity and appropriateness of the measurement tools used may also contribute to the variability in results. For instance, Othman and Hashim (2006), who employed TOEFL—a test emphasizing reasoning—reported a moderate-to-large effect size ($g = 0.6$), whereas Tok and Mazl (2015), who used a self-designed test focusing more on literal comprehension, found a

much smaller effect ($g = 0.05$). Although causality remains to be confirmed, this difference suggests that the design of measurement tools may influence the evaluation of outcomes. Future research would therefore benefit from adopting validated, sensitive, appropriate instruments that align closely with the targeted cognitive constructs, enabling a more accurate estimation of P4C's cognitive effects.

Our findings confirm the positive effects of P4C, through the implementation of a community of inquiry, on children's reasoning abilities, critical thinking, and creativity. While existing research has shown significant positive impacts of P4C on metacognitive skills (Şişman et al. 2023; Mehnehj et al. 2020) and problem-solving (Işıklar and Öztürk 2022), the limited number of studies included in this analysis prevented subgroup analysis in these areas. Future research should comprehensively incorporate more relevant analyses to accumulate sufficient studies for subgroup analyses in these areas.

5.2.2. Moderating Effect of Dosage Variables

Among dosage-related variables, none showed statistically significant moderating effects, providing no support for *H3.1*, *H3.2*, or *H3.3*.

Nevertheless, descriptive patterns suggest that session length may still matter: sessions lasting ≤ 50 min yielded a moderate effect ($g = 0.36$, $p < 0.05$), whereas sessions extending beyond 50 min showed much larger effects ($g = 0.65$, $p < 0.001$). A typical P4C session typically includes text reading, question generation, and agenda-setting before discussion. Limited session duration may not provide sufficient time for participation, deeper thinking, and thorough dialogue, thereby reducing the effectiveness. While longer sessions appear more beneficial, practical constraints such as curriculum scheduling and student fatigue must be considered.

Subgroup analyses indicate no significant difference by frequency. Although intervention duration was not a statistically significant moderator, the pattern of effect sizes across subgroups is noteworthy. Shorter-term interventions (≤ 20 weeks) produced slightly higher effects ($g = 0.61$) than longer-term interventions ($g = 0.49$). One possible explanation is the Hawthorne effect—where the introduction of a novel and enthusiastically implemented program may temporarily enhance student engagement and performance, generating short-term gains that are difficult to sustain (Topping and Trickey 2015). These early effects, while encouraging, may not fully reflect the enduring impact of the intervention. Thus, even relatively brief P4C programs can produce meaningful benefits, but sustained implementation may help consolidate and stabilize these effects.

Due to missing data in the dosage variables, the generalizability of our findings may be limited, and care should be taken when generalizing the results.

5.2.3. Moderating Effect of Grade Level

Findings of this study show that P4C exerts moderate-to-large positive effects on children's cognitive abilities across all pre-college educational stages, with no significant moderating role of grade level ($p = 0.37$), thereby not providing support for *H4*. This aligns with Yan et al. (2018)'s finding, who also reported no evidence of grade-level moderation. It challenges a major objection to P4C—namely, the claim by Piaget (1933) and his followers that young children lack the higher-order thinking skills required for philosophical inquiry (Kitchener 1990; White 1992). Contrary to this assumption, our findings demonstrate that young children can engage in philosophical thinking and achieve significant cognitive gains through P4C. Our findings are congruent with contemporary developmental theories that view young children as capable learners and “philosophical babies”, offering evidence from a different perspective that even at an early age, children can engage in reasoning, reflection,

and metacognitive activities. Moreover, the results demonstrate that young children can engage in philosophical thinking and achieve significant cognitive gains through P4C.

The consistent cognitive benefits of P4C across grade levels can be attributed to its pedagogy. Rather than transmitting simplified philosophical content, P4C creates an intellectual safe environment in which children are encouraged to explore philosophical questions arising from their lived experiences (Lipman et al. 1980; Matthews 1980), using everyday language rather than technical jargon. Such inquiry is accessible to learners of all ages, as it draws on their prior knowledge and experience. With teacher facilitation, students engage in cognitive moves such as evaluating reasons, offering examples, questioning assumptions, and forming reasoned judgments (Gregory 2008; Kennedy 2004). These dialogues promote development within each learner's zone of proximal development, enabling even young children can engage in P4C and derive cognitive benefits.

5.2.4. Moderating Effect of Cultural Context

Cultural context exhibited a marginally significant moderating effect ($p = 0.09$), with interventions implemented in non-Western contexts ($g = 0.74$) demonstrating larger effects than those in Western contexts ($g = 0.46$), supporting *H5*.

Although P4C originated in western philosophy, the findings highlight its cultural adaptability. As an educational program grounded in the community of inquiry, P4C maintains a clear theoretical core while allowing flexible adaptation. Including using culturally relevant stimulus—such as Confucian texts and idioms in China (Cao and Huang 2024; Lin and Huang 2025), or fairy tales in Denmark (Jespersen 2017); and context-sensitive pedagogical modifications, such as Jackson's p4cHI model in Hawai'i, which aligns with local Aloha values (Jackson 2001, 2012). These localized practices enable P4C to resonate more deeply with context and demonstrate its compatibility with diverse educational settings. While implementation challenges exist—such as teacher-centered traditions and exam-driven systems in China (Ku 2008; Wu 2021; Lam 2019), "P4C has been proven to both necessitate and induce broader educational reforms (e.g., toward student empowerment, inquiry-based pedagogy)" (Gregory 2013), this suggests that P4C not only adapts to but also gradually transforms educational environments, supporting its applicability and long-term value across diverse cultures. For instance, Lam (2019) and Wu (2021) observed shifts toward more dialogical and child-centered practices, with positive outcomes captured through measurement and observation.

Consistent with Yan et al. (2018), the present study also found larger effects in non-Western than in Western contexts. We speculate that this difference may be partly driven by sample size effects. In this meta-analysis, non-Western studies ($n = 17$) involved an average of only 53 participants, whereas western studies ($n = 16$) averaged 229. Small-scale projects often benefit from more favorable implementation condition, such as greater researcher involvement, closer monitoring of fidelity, while large-scale studies are more susceptible to heterogeneity in delivery, including variation in teacher expertise, dosage, and classroom conditions (Yan et al. 2018). Baseline pedagogical differences may also amplify P4C's impact. In many non-Western contexts, where education is shaped by collectivist values and higher power distance, instruction often emphasizes teacher authority, structured curriculum, and exam preparation, with limited space for critical thinking and autonomy (Lam 2012; Wu 2021; Imani et al. 2016; Kole 2025), the P4C pedagogy—child-centered, inquiry-driven, and encouraging critical reflection and student participation—may offer a more novel and engaging learning experience, potentially yielding greater gains in these contexts than in Western ones (Imani et al. 2016; Kiany and Movahedian 2010). Taken together, small-sample effects and baseline differences help explain this pattern, though the marginal significance ($p = 0.09$) calls for further larger cross-cultural evidence.

5.2.5. Moderating Effect of Research Design

Findings of this study indicate that the type of research design did not significantly moderate the cognitive effects of P4C, which did not support *H6*. However, there was a gradient pattern: RCTs yielded the smallest effect size ($g = 0.36$), one-group pretest–posttest designs produced the largest ($g = 0.88$), and Non-RCTs fell in between ($g = 0.57$). This gradient aligns with expectations based on internal validity: RCTs minimize bias through random allocation, whereas one-group designs lack control conditions and are more susceptible to confounding factors, potentially inflating effect sizes. This finding somewhat supports [García-Moriyón et al. \(2005\)](#)’s conclusion that more rigorous designs tend to yield smaller, more conservative estimates. Given the relatively small number of RCTs included in this analysis ($k = 4$, $\#es = 5$), the statistical power to detect moderating effects remains limited. Future research should prioritize well-controlled trials to better clarify the role of research design in shaping the estimated cognitive effects of P4C.

5.2.6. Moderating Effect of Publication Year

Publication year did not emerge as a significant moderator ($\beta = 0.002$, $p > 0.05$) in this meta-regression model. This finding does not support *H7* and offers insight into the evolving landscape of P4C research. It may indicate a degree of temporal robustness in the cognitive benefits of P4C, suggesting that its effectiveness has persisted across decades.

5.2.7. Moderating Effect of Sample Size

Meta-regression results revealed a significant negative relationship between sample size and effect size ($\beta = -0.001$, $p < 0.05$), indicating that smaller studies tended to report larger effects, supporting *H8*. This finding is consistent with the well-documented phenomenon of the small-study effects ([Luo et al. 2020](#)). In the context of educational interventions, large-scale studies often involve more implementation heterogeneity, including variation in teacher expertise, dosage, classroom conditions, and delivery consistency. For instance, [Fair et al. \(2015b\)](#) noted considerable discrepancies in session counts between grade levels in a large-scale P4C study, potentially contributing to uneven outcomes. These results highlight the importance of interpreting P4C effect sizes with attention to study scale, and the need for future research to incorporate detailed implementation monitoring to improve interpretability and robustness.

6. Conclusions, Limitation and Implications

6.1. Conclusions

This three-level meta-analysis of 33 studies (53 effect sizes) confirmed that P4C has a moderate-to-large, significant effect on children’s cognitive development, particularly in higher-order thinking skills such as creativity, reasoning, and critical thinking.

Subgroup and meta-regression analyses reveal that the cognitive effects of P4C are consistently positive across grade levels, research design, and publication years, indicating its cross-age, cross-design, and temporal stability. Dosage-related variables (session length, frequency, and total duration) were not significant moderators of cognitive outcomes, although descriptively longer sessions yielded larger effects than shorter ones. However, cultural context demonstrated marginally significant moderating trends, suggesting potential contextual and methodological influences that warrant further investigation. Sample size was found to be a significant moderator: smaller studies tended to report larger effect sizes.

6.2. Limitations

Despite its broad coverage, this meta-analysis has several limitations that future research should address.

First, although the included studies span over five decades and come from multiple regions, the geographic representation of the included studies was uneven, with concentration in Anglophone and select Asian countries. The exclusion of non-English publications may have omitted culturally distinct practices, limiting global generalizability. Future studies should incorporate multilingual databases and broaden regional coverage.

Second, although cultural context was examined as a moderator, our Western vs. non-Western coding is a heuristic simplification that overlooks within-group diversity. Country-level analyses were not feasible because several countries were represented by very few studies, limiting the validity of such comparisons. Future research should employ country-specific analyses to better capture how cultural values shape P4C's effectiveness.

Third, although variations in measurement tools are a likely source of heterogeneity, their scattered and inconsistent use across studies made it unfeasible to treat them as a moderator in this meta-analysis. Future meta-analyses should systematically examine the role of measurement tools to account for variability in effect sizes, while primary studies should adopt validated and appropriately sensitive instruments that more accurately capture the targeted cognitive constructs that P4C is designed to foster, thereby enabling more precise estimates of P4C's effects.

Fourth, several studies involved small sample sizes, which may have reduced statistical power and external validity. Large-scale studies are needed to provide more robust evidence.

Additionally, some moderator categories (e.g., RCTs, creativity outcomes) had few effect sizes, limiting the reliability of subgroup comparisons; missing data on key variables (e.g., intervention dosage) across studies introduced uncertainty, biasing analyses and reducing effect size precision. Future studies should broaden the coverage of underrepresented domains, and ensure comprehensive reporting of relevant data to enhance validity.

Finally, although key moderators were tested, other influential factors—such as teacher training quality and instructional fidelity—could not be tested due to limited reporting in primary studies. Future research should explore these factors through more detailed moderator analyses and mixed-methods approaches to uncover how these elements shape P4C's effectiveness.

6.3. Implications

The present three-level meta-analysis may contribute to the field in several ways. First, by synthesizing 33 studies conducted over more than five decades, it provides the most comprehensive quantitative evidence to date on the cognitive effects of P4C. Second, to the best of our knowledge, this is the first study in the P4C field to employ a three-level meta-analytic approach instead of a traditional meta-analysis, improving the credibility of the findings. Third, this study also investigates moderating factors that have not been systematically examined in previous reviews, such as session length, weekly frequency, and creativity-related outcomes, offering new insights into the conditions that shape the cognitive impact of P4C. Taken together, this study not only provides robust evidence for the effectiveness of P4C but also offers important implications for policy-making, educational practice, and future research.

For educational policymakers and curriculum developers, the findings provide compelling evidence to consider P4C as a viable option for integrating into contemporary education systems. Its significant and consistent cognitive benefits—across grade levels, cultural contexts, and historical periods—highlight its broad adaptability and enduring relevance. Given the rising demand for higher-order thinking in today's complex world—marked by

information overload, rapid advances in artificial intelligence, growing value pluralism, and rising societal uncertainty, P4C can be adopted as a stand-alone subject or embedded within existing subjects to foster key cognitive competences.

For P4C practitioners, the results underscore the importance of thoughtful program design. Longer sessions (over 50 min) were linked to stronger cognitive gains, suggesting that adequate time is essential for effective inquiry. While P4C works across diverse settings, culturally responsive implementation—such as using locally relevant materials and age-appropriate facilitation—can further enhance its impact.

For researchers, the findings underscore the need for more rigorous and transparent empirical studies on P4C. Future research should use larger, more diverse samples to improve statistical power and generalizability. To enhance interpretability and implementation relevance, researchers are encouraged to report key contextual variables—such as teacher qualifications, training quality, instructional fidelity, and classroom interaction patterns. Incorporating process measures and mixed-method approaches can help unpack how and under what conditions P4C is most effective.

By fostering children's engagement in a community of philosophical inquiry, P4C has once again been shown to enhance key cognitive competences such as critical thinking, reasoning, creativity and general cognitive abilities. As UNESCO (2007) affirms, P4C helps prepare individuals to “shoulder responsibilities in the face of the great challenges of the contemporary world”. Therefore, in the face of information overload, rapid advances in artificial intelligence, growing value pluralism, and rising societal uncertainty, P4C offers valuable insights for educational policy and curriculum reform, contributing to the development of reflective, responsible, and future-ready citizens.

Author Contributions: Conceptualization, C.W. (Caiyun Wei) and L.C. (Lele Chen); methodology, C.W. and L.C.; software, C.W.; validation, C.W. and L.C.; formal analysis, C.W.; investigation, C.W. and L.C.; resources, C.W. and L.C.; data curation, C.W. and L.C.; writing—original draft preparation, C.W.; writing—review and editing, C.W. and L.C.; visualization, C.W.; supervision, L.C.; project administration, L.C.; funding acquisition, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the National Social Science Fund of China, grant number BAA220161.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Coded data and scripts are available in the publicly accessible Open Science Framework (OSF) repository: <https://doi.org/10.17605/OSF.IO/EUAPX> (accessed on 1 October 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Abbasi, Mohammad, Zabih Pirani, Mohammad Reza Sarmadi, and Davoud Taghvaei. 2017. The effectiveness of philosophy education program for children on critical thinking and its components in female students. *Iranian Journal of Educational Sociology* 1: 96–103.
- Ab Wahab, Mohd Kaziman, Hafizah Zulkifli, and Khadijah Abdul Razak. 2022. Impact of Philosophy for Children and Its Challenges: A Systematic Review. *Children* 9: 1671. [CrossRef] [PubMed]
- Acar, Feride, and Recep Şahin Arslan. 2023. Enhancing Critical Thinking and English-Speaking Skills of Gifted Students through Philosophy for Children Approach. *International Journal of Assessment Tools in Education* 10: 345–75. [CrossRef]
- Akan, Ramazan, and A. Kadir Çüçen. 2023. The Interrelation between Philosophy for Children (P4C) and Creative Thinking. *Balkan Journal of Philosophy* 1: 35–40. [CrossRef]
- Akbayir, Kamil, and Sevda Kaya Tedikçi. 2022. The Effect of Philosophy Education on Mathematics Success. *Education Quarterly Reviews* 5: 192–207. [CrossRef]

- Assink, Mark, and Carlijn J. Wibbelink. 2016. Fitting Three-Level Meta-Analytic Models in R: A Step-by-Step Tutorial. *The Quantitative Methods for Psychology* 12: 154–74. [CrossRef]
- Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to Meta-Analysis*. Hoboken: John Wiley & Sons, Ltd.
- Boyras, Celal, and Yasin Ayday. 2024. Investigation of the Effect of Philosophy for Children Approach on Listening Comprehension of Fourth-Grade Primary School Students. *European Journal of Education Studies* 11: 1–15. [CrossRef]
- Cam, Philip. 2017. Philosophy for children, values education and the inquiring society. In *Philosophy in Schools*. Edited by Michael Hand and Carrie Winstanley. New York: Routledge, pp. 7–15.
- Camhy, Daniela G., and Gunter Iberer. 1988. Philosophy for Children: A research project for further mental and personality development of primary and secondary school pupils. *Thinking: The Journal of Philosophy for Children* 7: 18–25. [CrossRef]
- Cao, Jianbo, and Rui Huang. 2024. *Philosophy for Children: A Case-Based Course*. Nanjing: Jiangsu Phoenix Science & Technology Press.
- Carroll, John Bissell. 1993. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge: Cambridge University Press.
- Chamberlain, Mary Ann. 1993. Philosophy for Children Program and the Development of Critical Thinking of Gifted Elementary Students. Ph.D. thesis, University of Kentucky, Lexington, KY, USA.
- Cheung, Mike W.-L. 2014. Modeling Dependent Effect Sizes with Three-Level Meta-Analyses: A Structural Equation Modeling Approach. *Psychological Methods* 19: 211–29. [CrossRef]
- Cinquino, Dolly. 1981. An evaluation of a philosophy program with 5th and 6th grade academically talented students. *Thinking: The Journal of Philosophy for Children* 2: 79–83. [CrossRef]
- Cohen, Jacob. 2013. *Statistical Power Analysis for the Behavioral Sciences*. New York: Routledge.
- Colom, Roberto, Félix García Moriyón, Carmen Magro, and Elena Morilla. 2014. The long-term impact of Philosophy for Children: A longitudinal study (preliminary results). *Analytic Teaching and Philosophical Praxis* 35: 50–56. Available online: <https://journal.viterbo.edu/index.php/atpp/article/view/1129> (accessed on 1 October 2025).
- Cooke, Patricia A. 2015. The Impact of Engaging in Philosophy with Middle School Children on the Development of Critical Thinking. Ph.D. thesis, University of Rochester, Rochester, NY, USA.
- Cummings, Nancy Pekin. 1979. Improving the logical skills of fifth graders. *Thinking: The Journal of Philosophy for Children* 1: 90–92. [CrossRef]
- De Cesaris, Alessandro. 2018. New literacy and multidimensional thinking: P4C and the challenges of the digital age. In *Philosophy and Community Practices*. Edited by Roberto Franzini Tobaldeo and Graziano Lingua. New York: Peter Lang Publishing, pp. 145–55.
- Demirtaş, Vesile Yıldız, Filiz Karadağ, and Kurtul Gülenç. 2018. Levels of the Questions Formulated by Preschool Children During the Philosophical Inquiry Process and the Qualities of Their Answers. *International Online Journal of Educational Sciences* 10: 277–94. [CrossRef]
- Duval, Sue, and Richard Tweedie. 2000. Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics* 56: 455–63. [CrossRef]
- Erfani, Nasrolah, and Mandana Rezaei. 2016. The effect of teaching philosophy on the critical thinking of female students. *International Journal of Pharmaceutical Research & Allied Science* 5: 484–91.
- Fair, Frank, David Schapiro, and Nancy Fair. 2015a. Socrates in the Schools: Gains at Three-Year Follow-Up. *Journal of Philosophy in Schools* 2: 5–16. [CrossRef]
- Fair, Frank, Lory E. Haas, Carol Gardosik, Daphne D. Johnson, Debra P. Price, and Olena Leipnik. 2015b. Socrates in the Schools from Scotland to Texas: Replicating a Study on the Effects of a Philosophy for Children Program. *Journal of Philosophy in Schools* 2: 18–37. [CrossRef]
- Fernández-Castilla, Belén, Lies Declercq, Laleh Jamshidi, S. Natasha Beretvas, Patrick Onghena, and Wim Van den Noortgate. 2021. Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education* 1: 125–44. [CrossRef]
- Fisher, Robert. 2007. Dialogic Teaching: Developing Thinking and Metacognition through Philosophical Discussion. *Early Child Development and Care* 177: 615–31. [CrossRef]
- Fisher, Robert. 2008. *Teaching Thinking: Philosophical Enquiry in the Classroom*, 3rd ed. New York: Continuum.
- Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. Publication Bias in the Social Sciences: Unlocking the File Drawer. *Science* 345: 1502–5. [CrossRef]
- Gao, Shuling, Delin Yu, Mark Assink, Ko Ling Chan, Ling Zhang, and Xianxin Meng. 2024. The Association between Child Maltreatment and Pathological Narcissism: A Three-Level Meta-Analytic Review. *Trauma, Violence, & Abuse* 1: 275–90. [CrossRef]
- Gao, Zhenyu. 2018. Constructing a Philosophy for Children Curriculum System Based on Core Competencies. *Shanghai Educational Research* 1: 19–23.

- García-Moriyón, Felix, Irene Rebollo, and Roberto Colom. 2005. Evaluating Philosophy for Children: A meta-analysis. *Thinking: The Journal of Philosophy for Children* 17: 14–22. [CrossRef]
- Gardner, Susan T. 1998. Philosophy for Children Really Works! A Report on a Two Year Empirical Study. *Critical and Creative Thinking* 6: 1–15.
- Gasparatou, Renia, and Maria Kampeza. 2012. Introducing P4C in Kindergarten in Greece. *Analytic Teaching and Philosophical Praxis* 33: 72–82. Available online: <https://journal.viterbo.edu/index.php/attp/article/view/1076> (accessed on 1 October 2025).
- Ghaedi, Yahya, Mona Mahdian, and Fatemeh Khoshnavay Fomani. 2015. Identifying Dimensions of Creative Thinking in Preschool Children During Implementation of Philosophy for Children (P4C) Program: A Directed Content Analysis. *American Journal of Educational Research* 3: 547–51. [CrossRef]
- Gopnik, Alison. 2009. *The Philosophical Baby: What Children's Minds Tell Us About Truth, Love and the Meaning of Life*. New York: Random House.
- Gopnik, Alison, and Henry M. Wellman. 2012. Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin* 6: 1085–108. [CrossRef] [PubMed]
- Gopnik, Alison, Andrew N. Meltzoff, and Patricia K. Kuhl. 1999. *The Scientist in the Crib: Minds, Brains, and How Children Learn*. New York: William Morrow & Co.
- Gordon, Jean, Gábor Halász, Magdalena Krawczyk, Tom Leney, Alain Michel, David Pepper, Elzbieta Putkiewicz, and Jerzy Wiśniewski. 2009. *Key Competences in Europe: Opening Doors for Lifelong Learners Across the School Curriculum and Teacher Education*. CASE Network Reports No. 87. Warsaw: Center for Social and Economic Research. Available online: <https://www.econstor.eu/bitstream/10419/87621/1/613705459.pdf> (accessed on 12 June 2025).
- Gregory, Maughn. 2008. *Philosophy for Children Practitioner Handbook*. Montclair: Institute for the Advancement of Philosophy for Children.
- Gregory, Maughn. 2013. Precollege philosophy education: What can it be? The IAPC model. In *Philosophy in Schools*. Edited by Michael Hand and Carrie Winstanley. London: Routledge, pp. 69–85.
- Gregory, Maughn. 2021. Philosophy for Children and children's philosophical thinking. In *A History of Western Philosophy of Education in the Contemporary Landscape*. Edited by Anna Pagès. London: Bloomsbury, pp. 153–77.
- Gu, Xiaoqing, and Menghua Hu. 2018. The learning effects of e-schoolbag: A meta-analysis of 39 studies at home and abroad. *E-Education Research* 39: 19–25.
- Hedges, Larry V. 1981. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 6: 107–28. [CrossRef]
- Higa, William R. 1980. Philosophy for Children in Hawaii: A quantitative evaluation. *Thinking: The Journal of Philosophy for Children* 2: 21–31. [CrossRef]
- Hofstede, Geert. 1984. *Culture's Consequences: International Differences in Work-Related Values*. Beverly Hills: Sage Publications.
- Hofstede, Geert, Gert Jan Hofstede, and Michael Minkov. 2010. *Cultures and Organizations: Software of the Mind*. New York: McGraw-Hill.
- Hopewell, Sally, Mike Clarke, and Sue Mallett. 2005. Grey Literature and Systematic Reviews. In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Edited by Hannah R. Rothstein, Alexander J. Sutton and Michael Borenstein. Chichester: John Wiley & Sons, Ltd., pp. 49–72.
- Hunter, John E., and Frank L. Schmidt. 1990. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks: Sage Publications.
- Imani, Homa, Ghodsi Khadijah, and Maryam Seif Naraghi. 2016. The Role of Philosophy for Children (P4C) Teaching Approach for Improving the Reading Comprehension Skills of Guidance School Female Students. *Iranian Journal of Educational Sociology* 1: 54–59. Available online: <http://iase-idje.ir/article-1-38-en.html> (accessed on 1 October 2025).
- Iorio, John, Mark Weinstein, and John Martin. 1984. A review of District 24's Philosophy for Children program. *Thinking: The Journal of Philosophy for Children* 5: 28–35. [CrossRef]
- İşıklar, Sema, and Yasemin Abalı Öztürk. 2022. The effect of Philosophy for Children (P4C) curriculum on critical thinking through philosophical inquiry and problem solving skills. *International Journal of Contemporary Educational Research* 9: 130–42. [CrossRef]
- Jackson, Thomas. 2001. The art and craft of 'gently Socratic' inquiry. In *Developing Minds: A Resource Book for Teaching Thinking*, 3rd ed. Edited by Arthur L. Costa. Alexandria: ASCD, pp. 1–13.
- Jackson, Thomas. 2012. Home grown. *Educational Perspectives* 44: 3–7. [CrossRef]
- Jenkins, Joseph. 1986. Philosophy for Children programme at a Gloucestershire comprehensive school in Great Britain. *Thinking: The Journal of Philosophy for Children* 6: 33–37. [CrossRef]
- Jespersen, Per. 2017. P4C in Denmark. In *History, Theory and Practice of Philosophy for Children*. Edited by Saeed Naji and Rosnani Hashim. London: Routledge, pp. 143–49.

- Jo, Seon-Hee. 2000. Literacy: Constructing meaning through philosophical inquiry. *Analytic Teaching* 21: 44–52. Available online: <https://journal.viterbo.edu/index.php/at/article/view/728> (accessed on 1 October 2025).
- Kanani Harandi, Setare, Mohammad Nourian, Dariush Noroozi, and Mahmood Abaei Koopaei. 2021. The effect of Philosophy for Children curriculum on the growth of students' creativity. *Thinking and Children* 12: 203–30.
- Karras, Ray W. 1979. Final Evaluation of the Pilot Program in Philosophical Reasoning in Lexington Elementary Schools 1978–79. *Thinking: The Journal of Philosophy for Children* 1: 26–32. [CrossRef]
- Kennedy, David. 2004. Communal philosophical dialogue and the intersubject. *International Journal of Applied Philosophy* 18: 201–16. [CrossRef]
- Kennedy, David. 2013. Developing Philosophical Facilitation: A Toolbox of Philosophical 'Moves'. In *Philosophy in Schools: An Introduction for Philosophers and Teachers*. Edited by Sara Goering, Nicholas J. Shudak and Thomas E. Wartenburg. New York: Routledge, pp. 110–18.
- Kennedy, Nadia, and David Kennedy. 2011. Community of philosophical inquiry as a discursive structure, and its role in school curriculum design. *Journal of Philosophy of Education* 45: 265–83. [CrossRef]
- Khanmohammadi, Marmar, Mohammad Seifi, and Faezeh Nateghi. 2020. Effectiveness of applying Philosophy for Children (P4C) teaching approach in a media literacy course. *Interdisciplinary Journal of Virtual Learning in Medical Sciences* 11: 112–20.
- Kiany, Gholamreza, and Mostafa Movahedian. 2010. *A comparative Study of English Policies, Assessment Procedures Training Programs in Iran, China & Singapore*. Tehran: Tarbiat Modares University.
- Kiely, Kim. 2014. Cognitive function. In *Encyclopedia of Quality of Life and Well-Being Research*. Edited by Kim M. Michalos. Dordrecht: Springer, pp. 974–78. [CrossRef]
- Kitchener, Richard F. 1990. Do children think philosophically? *Metaphilosophy* 21: 416–31. [CrossRef]
- Kole, John Kirwa Tum. 2025. Applying Hofstede's Cultural Dimensions in Education: Insights, Critiques, and Implications for Diverse Classrooms. *Open Journal of Social Sciences* 13: 94–110. [CrossRef]
- Koricheva, Julia, and Elena Kulinskaya. 2019. Temporal instability of evidence base: A threat to policy making? *Trends in Ecology & Evolution* 34: 895–902. [CrossRef]
- Ku, Yee-lai. 2008. Critical Thinking of Chinese Students: Conceptualization, Assessment and Instruction. Ph.D. thesis, The University of Hong Kong, Hong Kong, China.
- Lam, Chi-Ming. 2012. Continuing Lipman's and Sharp's pioneering work on Philosophy for Children: Using Harry to foster critical thinking in Hong Kong students. *Educational Research and Evaluation* 18: 187–203. [CrossRef]
- Lam, Chi-Ming. 2019. Integrating Philosophy into English Curriculum: The Development of Thinking and Language Competence. *The Journal of Educational Research* 112: 700–9. [CrossRef]
- Lam, Chi-Ming. 2021. Development of Thinking and Language Skills through Philosophy: A Case Study in Hong Kong. *Cambridge Journal of Education* 51: 127–42. [CrossRef]
- Lewis, Lizzy, and Nick Chandley, eds. 2012. *Philosophy for Children through the Secondary Curriculum*. London: Bloomsbury Publishing.
- Li, Jin. 2012. *Cultural foundations of learning: East and West*. Cambridge, UK: Cambridge University Press.
- Lin, Xu, and Rui Huang. 2025. *Philosophical Psychology: Idiom Stories*. Shanghai: Shanghai Education Press.
- Lipman, Matthew. 1976. Philosophy for Children. *Metaphilosophy* 7: 17–39. [CrossRef]
- Lipman, Matthew. 1990. Response to Professor Kitchener. *Metaphilosophy* 4: 432–33. [CrossRef]
- Lipman, Matthew. 2003. *Thinking in Education*. Cambridge, UK: Cambridge University Press.
- Lipman, Matthew, Ann Margaret Sharp, and Frederick S. Oscanyan. 1980. *Philosophy in the Classroom*. Philadelphia: Temple University Press.
- Lipsey, Mark W., and David B. Wilson. 2001. *Practical Meta-Analysis*. Thousand Oaks: Sage Publications.
- Liu, Dong, and Roy F. Baumeister. 2016. Social networking online and personality of self-worth: A meta-analysis. *Journal of Research in Personality* 64: 79–89. [CrossRef]
- Liu, Xueliang. 2020. The Basic Framework of Philosophy for Children and Its Transcendence over Dualism. *Educational Development Research* 22: 21–27.
- Lord, Pippa, Afrah Dirie, Kelly Kettlewell, and Ben Styles. 2021. *Evaluation of Philosophy for Children: An Effectiveness Trial*. London: Educational Endowment Foundation.
- Luo, Chongliang, Tianjing Li, Yong Chen, and Wensheng Guo. 2020. Accounting for small-study effects using a bivariate trim and fill meta-analysis procedure. *medRxiv* 7: 1–16. [CrossRef]
- Luo, Jie. 2013. *Systematic Evaluation and Meta-Analysis: Theory and Practice*. Beijing: Military Medical Science Press.
- Marashi, S. Mansour. 2008. Teaching philosophy to children: A new experience in Iran. *Analytic Teaching* 27: 12–15.
- Martin, John, and Mark Weinstein. 1985. Thinking skills and Philosophy for Children: The Bethlehem Program, 1982–1983. *Analytic Teaching* 5: 28–31. Available online: <https://journal.viterbo.edu/index.php/at/article/view/304> (accessed on 1 October 2025).
- Matthews, Gareth B. 1980. *Philosophy and the Young Child*. Cambridge: Harvard University Press.

- Mehnehj, Falah T., Falah Mehnehj, Turaj, Mohamad Yamini, and Hosein Mahdian. 2020. The effect of teaching philosophy on metacognitive and irrational beliefs of male elementary school students. *Journal of Research and Health* 10: 183–92. [CrossRef]
- Meyer, John. 1989. A quest of the possible? Evaluation of the impact of the Pixie Programme on 8–10 year olds. *Analytic Teaching* 9: 63–75. Available online: <https://journal.viterbo.edu/index.php/at/article/view/470> (accessed on 1 October 2025).
- Michalik, Kerstin. 2023. Philosophising with children—Learning to live with uncertainty as a condition of human life and a challenge for teaching and learning. *Teachers and Teaching* 29: 680–94. [CrossRef]
- Moher, David, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *BMJ* 339: b2535. [CrossRef] [PubMed]
- Murris, Karin. 1999. Philosophy with Preliterate Children. *Thinking: The Journal of Philosophy for Children* 14: 23–33. [CrossRef]
- Naderi, Ezatollah, Nia Amir Tajali, Ali Shariatmadari, and Naraghi Maryam Seif. 2012. The effect of P4C program on creativity of high school boys in Tehran. *Thinking and Children* 3: 91–117.
- Nakagawa, Shinichi, Malgorzata Lagisz, Michael D. Jennions, Julia Koricheva, Daniel W. A. Noble, Timothy H. Parker, Alfredo Sánchez-Tójar, Yefeng Yang, and Rose E. O’Dea. 2022. Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods in Ecology and Evolution* 13: 4–21.
- National Health Commission of the People’s Republic of China. 2012. *Health Requirements of Daily Learning Time for Secondary and Elementary School Students*; GB/T 17223-2012. Beijing: National Health Commission. Available online: <https://openstd.samr.gov.cn/bzgk/gb/newGbInfo?hcno=D6DA88D3260600A078EEF6CE21AC61A6> (accessed on 22 May 2025).
- National Institute of Health. 2021. Study Quality Assessment Tools. Available online: <https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools> (accessed on 12 August 2025).
- OECD. 2010. The Definition and Selection of Key Competencies: Executive Summary. Available online: [https://one.oecd.org/document/EDU/EDPC/ECEC/RD\(2010\)26/en/pdf](https://one.oecd.org/document/EDU/EDPC/ECEC/RD(2010)26/en/pdf) (accessed on 12 June 2025).
- Othman, Moomala, and Rosnani Hashim. 2006. Critical thinking and reading skills: A comparative study of the reader response and the Philosophy for Children approaches. *Thinking: The Journal of Philosophy for Children* 18: 26–34. [CrossRef]
- Oyler, Joe. 2017. Philosophy with Children: The Lipman-Sharp approach to Philosophy for Children. In *Encyclopedia of Educational Philosophy and Theory*. Edited by Michael A. Peters. New York: Springer, pp. 1–7.
- Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, and et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery* 88: 105906. [CrossRef]
- Partnership for 21st Century Skills. 2009. *P21 Framework Definitions*; Washington, DC: Partnership for 21st Century Skills. Available online: <https://files.eric.ed.gov/fulltext/ED519462.pdf> (accessed on 12 June 2025).
- Pálsson, Hreinn. 1995. We think more than before about others and their opinions. *Thinking: The Journal of Philosophy for Children* 12: 24–29. [CrossRef]
- Piaget, Jean. 1933. Children’s Philosophies. In *A Handbook of Child Psychology*. Edited by Carl Murchison. Worcester: Clark University Press, pp. 377–91.
- Pourtaghi, Vahidreza, Afzalsadat Hosseini, and Elahe Hejazi Moghari. 2014. Effectiveness of Implementing Philosophy for Children Program on Students’ Creativity. *Scientific Journal of Pure and Applied Sciences* 7: 1–10.
- Rahdar, Anisheh, Abdulwahab Pourghaz, and Afsaneh Marziyeh. 2018. The impact of teaching Philosophy for Children on critical openness and reflective skepticism in developing critical thinking and self-efficacy. *International Journal of Instruction* 11: 539–56. [CrossRef]
- Reed, Stephen K. 2020. *Cognitive Skills You Need for the 21st Century*. New York: Oxford University Press.
- Rodgers, Melissa A., and James E. Pustejovsky. 2021. Evaluating Meta-Analytic Methods to Detect Selective Reporting in the Presence of Dependent Effect Sizes. *Psychological Methods* 26: 141–60. [CrossRef]
- Rosenthal, Robert. 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin* 86: 638–41. [CrossRef]
- Sabbagh Hasanazadeh, Talat. 2024. The effectiveness of teaching Philosophy for Children on math performance and math self-efficacy and reducing math anxiety of students with math learning disorders. *Thinking and Children* 16: 273–307. [CrossRef]
- Santi, Marina. 2019. Collaborative problem-solving and citizenship education: A philosophical escape in the age of competencies. *Childhood & Philosophy* 15: 1–19.
- Säre, Egle, Piret Luik, and Tiia Tulviste. 2016. Improving pre-schoolers’ reasoning skills using the Philosophy for Children programme. *Trames: A Journal of the Humanities and Social Sciences* 20: 273–84. [CrossRef]
- Shorer, Alison, and Katie Quinn. 2023. *Philosophy for Children Across the Primary Curriculum: Inspirational Themed Planning*. London: Routledge.
- Slade, Christina. 1989. Logic in the classroom. *Thinking: The Journal of Philosophy for Children* 8: 14–20. [CrossRef]
- Sprod, Tim. 1998. “I can change your opinion on that”: Social constructivist whole class discussions and their effect on scientific reasoning. *Research in Science Education* 28: 463–80. [CrossRef]
- Sterne, Jonathan A., and Roger M. Harbord. 2004. Funnel Plots in Meta-Analysis. *The Stata Journal* 4: 127–41.

- Stokell, Kathy, Diane Swift, and Babs Anderson. 2017. P4C in the primary school. In *Philosophy for Children: Theories and Praxis in Teacher Education*. Edited by Babs Anderson. London: Routledge, pp. 66–71.
- Şişman, Zehra Betül, İhsan Ünlü, and Alper Kaşkaya. 2023. The effect of P4C application in the Philosophy for Children education program on the metacognitive awareness of secondary school 5th grade students. *International Journal of Educational Technology and Scientific Research* 3: 406–26.
- Tian, Shiauping, and Pei-Fen Liao. 2016. Philosophy for Children with Learners of English as a Foreign Language. *Journal of Philosophy in Schools* 3: 40–58. [CrossRef]
- Tok, Şükran, and Ayşegül Mazl. 2015. The effect of Stories for Thinking on reading and listening comprehension: A case study in Turkey. *Research in Education* 93: 1–18. [CrossRef]
- Topping, Keith J., and Steve Trickey. 2007a. Collaborative philosophical enquiry for school children: Cognitive effects at 10–12 years. *British Journal of Educational Psychology* 77: 271–88. [CrossRef]
- Topping, Keith J., and Steve Trickey. 2007b. Collaborative philosophical inquiry for schoolchildren: Cognitive gains at 2-year follow-up. *British Journal of Educational Psychology* 77: 787–96. [CrossRef]
- Topping, Keith J., and Steve Trickey. 2015. Philosophy for children: Short and long term effects. In *The Routledge International Handbook of Research on Teaching Thinking*. Edited by Rupert Wegerif, Li Li and James C. Kaufman. London: Routledge, pp. 103–112.
- Triandis, Harry C., and Michele J Gelfand. 1998. Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology* 1: 118–28. [CrossRef]
- Trickey, Steve, and Keith J. Topping. 2004. 'Philosophy for Children': A systematic review. *Research Papers in Education* 19: 365–80. [CrossRef]
- UNESCO. 2007. *Philosophy: A School of Freedom*. Paris: UNESCO.
- United Nations General Assembly. 1989. *Convention on the Rights of the Child*; New York: UN Office of the High Commissioner for Human Rights. Available online: <https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child> (accessed on 20 May 2025).
- Ventista, Ourania Maria. 2018. A literature review of empirical evidence on the effectiveness of Philosophy for Children. In *Family Resemblances: Current Proposals in Philosophy for Children*. Madrid: Anaya Educación, pp. 448–469.
- Ventista, Ourania Maria. 2019. An Evaluation of the "Philosophy for Children" Programme: The Impact on Cognitive and Non-Cognitive Skills. Ph.D. thesis, Durham University, Durham, UK.
- Viechtbauer, Wolfgang. 2010. Conducting Meta-Analyses in R with the Metafor Package. *Journal of Statistical Software* 36: 1–48. [CrossRef]
- Voogt, Joke, and Natalie Pareja Roblin. 2012. A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies* 44: 299–321. [CrossRef]
- Vygotsky, Lev Semenovich. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Cambridge: Harvard University Press.
- Watkins, David. 2014. Correlates of approaches to learning: A cross-cultural meta-analysis. In *Perspectives on Thinking, Learning, and Cognitive Styles*. Edited by Robert J. Sternberg and Li-fang Zhang. New York: Routledge, pp. 165–95.
- White, John. 1992. The roots of philosophy. *Royal Institute of Philosophy Supplements* 33: 73–88. [CrossRef]
- Wu, Caiwei. 2021. Training teachers in China to use the Philosophy for Children approach and its impact on critical thinking skills: A pilot study. *Education Sciences* 11: 206. [CrossRef]
- Yan, Sijin, Lynne Masel Walters, Zhuoying Wang, and Chia-Chiang Wang. 2018. Meta-analysis of the effectiveness of Philosophy for Children programs on students' cognitive outcomes. *Analytic Teaching and Philosophical Praxis* 39: 13–33. Available online: <https://journal.viterbo.edu/index.php/atpp/article/view/1160> (accessed on 1 October 2025).
- Yarmel, Aaron, and Maughn Gregory. 2025. Noise Reduction in the Community of Philosophical Inquiry. *Journal of Philosophy of Education*, qhaf059. [CrossRef]
- Yeazell, Mary I. 1982. Improving reading comprehension through Philosophy for Children. *Reading Psychology: An International Quarterly* 3: 239–46. [CrossRef]
- Youssef, Chadi, Marilyn Campbell, and Donna Tangen. 2016. The effects of participation in a P4C program on Australian elementary school students. *Analytic Teaching and Philosophical Praxis* 37: 1–19.
- Zhang, Hua. 2016. On the Connotations of A Key Competence. *Global Education* 4: 10–24.
- Zheng, Hao-Min, Zhong-Lin Wen, and Yan Wu. 2011. The Appropriate Effect Sizes and Their Calculations in Psychological Research. *Advances in Psychological Science* 12: 1868–78.
- Zhu, Yan-Han, Bin He, and Lei Sun. 2024. The effects of state power on prosocial behavior: A three-level meta-analysis. *Advances in Psychological Science* 11: 1786–99. [CrossRef]

- Zulkifli, Hafizhah, and Rosnani Hashim. 2019. The development of questioning skills through Hikmah (wisdom) pedagogy. *Creative Education* 10: 2593–605. [CrossRef]
- Zulkifli, Hafizhah, and Rosnani Hashim. 2020. Philosophy for Children (P4C) in improving critical thinking in a secondary moral education class. *International Journal of Learning, Teaching and Educational Research* 19: 29–45. Available online: <http://ijlter.org/index.php/ijlter/article/view/1824/pdf> (accessed on 12 June 2025). [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.