

The impact of virtual experiments on students' scientific inquiry skills: Evidence from a three-level meta-analysis

Hongwei Wu¹

Northwest Normal University

Zhengdong Zhang²

nxezzzd@163.com

Northwest Normal University

Systematic Review

Keywords: Virtual Experiments, Physical Experiments, Scientific Inquiry Skills, Three-level meta-analysis, Simulations

Posted Date: August 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10285699/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Psychology on September 29th, 2026. See the published version at <https://doi.org/10.1186/s40359-026-05568-1>.

The impact of virtual experiments on students' scientific inquiry skills: Evidence from a three-level meta-analysis

Hongwei Wu¹, Zhengdong Zhang²

(¹College of Educational Science, Northwest Normal University, Lanzhou 730070, China; ²College of Mathematics and Statistics, Northwest Normal University, Lanzhou 730070, China)

Corresponding author: Zhengdong Zhang. Email:nxezzzd@163.com

Abstract

Virtual experiments (VE) represent a key practical paradigm for digital transformation in science education. However, no consensus exists regarding their intervention effects on students' scientific inquiry skills compared with physical experiments (PE). Accordingly, this study adhered to the PRISMA 2020 guidelines and systematically retrieved empirical literature published from 2000 to 2025. A total of 21 experimental and quasi-experimental studies (QES) were finally included, comprising 41 effect sizes and 3931 participants. A three-level meta-analysis model was employed to compare the relative intervention effects of VE versus PE on students' scientific inquiry skills. We also examined the moderating effects of educational level, subject type, intervention duration, and inquiry skill dimensions. The results revealed that VE exhibited a significant small-to-moderate positive effect on students' scientific inquiry skills compared with PE (Hedges' $g = 0.454$, $p < 0.05$). Moderating effect analysis indicated that only subject type had a marginally significant moderating effect on the intervention effect ($p = 0.079$). Subgroup analysis showed that the relative advantages of VE were most pronounced in high school (Hedges' $g = 0.690$, $p < 0.05$), physics subjects (Hedges' $g = 0.603$, $p < 0.05$), and medium-duration interventions of 4-8 weeks (Hedges' $g = 0.590$, $p < 0.05$). However, VE showed no significant advantage in the experimental design dimension (Hedges' $g = 0.056$, $p > 0.05$). Based on these findings, this study explains the differentiated advantages of VE in reducing cognitive load and PE in enhancing embodied experience from the perspective of cognitive-embodied dynamic balance. It also elucidates the inherent educational mechanism of their complementarity. This study clarifies the respective suitable teaching scenarios for VE and PE. It provides empirical evidence for resolving existing research controversies and promoting the digital implementation of science experiment teaching and the development of virtual-physical integrated experiment teaching.

Keywords

Virtual Experiments; Physical Experiments; Scientific Inquiry Skills; Three-level meta-analysis; Simulations

1 Introduction

Scientific inquiry skills form an integrated cognitive and practical competency system. Students gradually develop these skills in scientific experiment settings. They can be quantified via standardized assessments and reflected through observable behaviors. This system covers five core dimensions: asking scientific questions, formulating testable hypotheses, designing controlled experiments, analyzing and interpreting data, drawing evidence-based conclusions (Chen et al., 2025; Darman et al., 2024; Seeratan et al., 2020). Scientific inquiry skills constitute a vital component of scientific literacy. They underpin logical reasoning, mechanism interpretation and creative problem solving. They are also core learning objectives of science curriculum reform across global basic education (Anam et al., 2025; Hofstein & Lunetta, 2004; Husnaini & Chen, 2019). Science experiment teaching serves as a primary pathway to foster these skills. The digital transformation of science education has continued to advance worldwide. Physical experiments (PE), virtual experiments (VE) and virtual-physical integrated experimental approaches have evolved into complementary teaching paradigms. They are widely adopted in science experiment instruction (Chen et al., 2026; Wörner et al., 2022). The integration of generative AI and VE has further broadened its application scope.

Since VE were widely adopted in basic education classrooms in the 1990s, empirical heterogeneity has existed regarding their effects on scientific inquiry skills relative to PE. Scholars have not reached a consistent conclusion to date (Mulwijk & Lazonder, 2023; Y. Wang et al., 2025). This long-standing academic dispute brings practical difficulties to frontline teachers. Teachers struggle to identify optimal scenarios for the two experimental modes. They cannot design collaborative teaching activities accordingly. This hinders the full educational potential of virtual-physical integrated experiments (Papalazarou et al., 2024). It is therefore essential to distinguish the differentiated effects of VE and PE. Identifying their respective suitable scenarios is equally important. This work not only improves the theoretical framework of digital science education, but also provides empirical support for evidence-based experimental teaching design.

Cognitive Load Theory (CLT) states that human cognitive resources are limited. Learning outcomes depend on the match between effective cognitive load and total cognitive capacity (Sweller, 1994). VE remove redundant operational steps and optimize resource allocation. This enables VE to facilitate cognitive-oriented inquiry skills (Jia & Zeng, 2025; Ouwehand et al., 2025; Pedaste et al., 2015). These skills include asking scientific questions, formulating testable hypotheses, analyzing and interpreting data, and drawing evidence-based conclusions. Embodied Cognition Theory (ECT) defines the link between physical interaction, situational experience and cognitive construction. Real-time interaction between individuals and environments is fundamental to cognition (Wilson, 2002). This explains why PE cultivate practical skills through hands-on operations, especially for experimental design (Pedaste et al., 2015). PE present complementary educational functions. Relying on authentic environments and immediate feedback, PE effectively train practical skills such as instrument assembly, equipment calibration and fault diagnosis (Puntambekar et al., 2021). Their multisensory embodied experience is irreplaceable for long-term skill retention and improvement (Chien et al., 2015; Zacharia & De Jong, 2014). Nevertheless, PE have inherent limitations. Rigid standard procedures consume limited cognitive resources. They distract students from high-order tasks involving variable manipulation and principle deduction (Rau, 2020). In contrast, VE apply process visualization, simulation and digital scene reconstruction. They simplify repetitive operations and streamline cognitive processing. VE show unique strengths in cultivating cognitive-oriented inquiry skills (Wen et al., 2024; Wu et al., 2025). The two theories jointly interpret the complementary nature of VE and PE (Zou et al., 2025). Their viewpoints align with the dual attributes of scientific inquiry skills: cognitive processing and hands-on operation.

Despite the theoretical consensus on complementarity, empirical studies yield conflicting results. Three competing perspectives have emerged to compare the instructional effects of VE and PE. Each perspective is backed by empirical evidence yet confined to specific conditions. The first is the equivalence perspective. Under highly standardized tasks and teacher-led structured guidance, VE and PE produce comparable outcomes. This applies to settings without physical operation requirements. Numerous cross-disciplinary studies and replication works support this view. Controlled experiments by Kapici and colleagues (2019, 2022), the classic study by Triona & Klahr (2003), and a replication by Anam et al. (2025) all confirm equivalent performance in standardized inquiry tasks. Similar findings are reported in chemistry

(Hu-Au & Okita, 2021) and biology (Dodds, 2013; Oser & Fraser, 2015). Pyatt & Sims (2012) further extended this conclusion to pre-service science teacher education. Even so, the equivalence perspective has narrow applicability. Most relevant studies adopt teacher-dominated structured designs. Such settings reduce students' hands-on practice and weaken the embodied advantages of PE. Assessments also focus heavily on cognitive indicators. Thus, the conclusion cannot be generalized to open-ended inquiry classrooms.

The second is the PE superiority perspective. Rooted in ECT, it argues that physical operation is necessary for internalizing practical skills. PE excel at practical tasks such as experimental design and on-site troubleshooting. Controlled experiments in physics verify the practical advantages of PE under unified teaching arrangements (Chen et al., 2014; Husnaini & Chen, 2019). Aktaş & Karamustafaoğlu (2025) confirmed the irreplaceable role of PE for advanced practical skills among pre-service teachers. Chemistry research further demonstrates their complementarity. PE perform better in practical design, while VE show advantages in cognitive interpretation of experimental mechanisms (Chen et al., 2024). Two major flaws exist in this line of research. Most interventions last fewer than four weeks, lacking long-term data on skill retention. In addition, these studies ignore moderating variables such as educational level and intervention duration. They fail to define applicable boundaries, leading to limited external validity.

The third is the VE superiority perspective. Based on CLT, VE reduce unnecessary cognitive burden via digital simulation (Klahr et al., 2007). Learners can therefore concentrate on high-order logical reasoning. VE gain distinct advantages in developing cognitive inquiry skills. Recent studies cover diverse educational levels and disciplines. Physics research across different educational levels proves the benefits of VE for abstract experimental content (Farrokhnia & Esmailpour, 2010; Kapici, 2025; Ogbuanya & Onele, 2018; Wang et al., 2015; Yang & Heh, 2007). Consistent results are found in chemistry (Tatli & Ayas, 2013), integrated science (Ifthinan & Atun, 2019) and biology (Huppert et al., 2002; Toth et al., 2014). VE break hardware restrictions and promote students' higher-order thinking. However, this perspective overlooks the foundational value of embodied experience. It cannot explain the weaknesses of VE in refined practical training.

Meta-analysis is a powerful method to synthesize scattered empirical evidence. It integrates effect sizes from independent studies for quantitative analysis (Glass, 1976). This approach overcomes limitations of single studies and generates generalizable

conclusions (Lipsey & Wilson, 2001). A few existing meta-analyses have compared the effectiveness of VE and PE (Cao & Hsu, 2023; Castro, 2025; X.-M. Wang et al., 2025). Nevertheless, they share three prominent shortcomings. First, most analyses stay at a macro level. They evaluate overall inquiry ability without decomposing specific skill dimensions. Second, mechanism analyses remain superficial. Most only conduct tests with single moderators and cannot clarify optimal application scenarios. Third, traditional statistical models have methodological defects. Educational datasets commonly feature a nested structure of studies and effect sizes. Multiple effect sizes are often reported within one single study. Conventional two-level models fail to address dependence among non-independent effect sizes. This problem causes standard error underestimation by 10% to 30% and elevated Type I error rates (Cheung, 2014; Y. Wang et al., 2025; Yang et al., 2024). Three-level meta-analysis divides total variance into three components: level 1 (sampling error), level 2 (within-study variance) and level 3 (between-study variance) (Assink & Wibbelink, 2016). It effectively corrects data dependence bias, improves the accuracy of effect size estimation and strengthens result robustness.

Apart from methodological issues, heterogeneity also arises from variations in research design. Differences in educational level, subject, intervention duration and inquiry skill dimensions contribute to inconsistent findings (Holzmeister et al., 2023). Younger learners have immature embodied cognition. Cognitive development theory proposed by Piaget reveals obvious developmental differences across educational levels (Piaget, 1976). Physics, chemistry and biology differ greatly in abstractness, which leads to varied cognitive load. More abstract subjects and cognition-oriented tasks amplify the advantages of VE in cognitive optimization. Intervention duration determines cumulative embodied experience and skill internalization. Short interventions (< 4 weeks) hardly build stable embodied cognition. Long interventions (>8 weeks) significantly improve the retention of practical skills. Furthermore, the dual nature of scientific inquiry skills accounts for conflicting empirical results. Practical skills rely heavily on physical manipulation and real feedback, which favor PE. Cognitive skills emphasize abstract reasoning and model construction, which fit the lightweight digital environment of VE (Chen & Klahr, 1999).

To address the above research gaps, the present study adopts a three-level meta-analysis. The research framework is illustrated in Fig. 1. We set strict inclusion criteria for literature selection. We include peer-reviewed articles published between 2000 and 2025 with strict parallel-group designs of VE and PE. Studies on virtual-physical

integrated teaching or single-mode experiment evaluation are excluded. This study takes educational level, subject type, intervention duration and inquiry skill dimensions as potential moderating variables. It explores the heterogeneity of VE and PE effects as well as underlying moderating mechanisms. Three core research questions are proposed:

(1) What are the relative effects of VE compared with PE on students' scientific inquiry skills and their respective dimensions?

(2) Can educational level, subject type, intervention duration, and inquiry skill dimensions statistically explain the effectiveness heterogeneity between VE and PE?

(3) Based on the meta-analytic evidence, what are the respective significant effect scenarios of VE and PE in cultivating scientific inquiry skills?

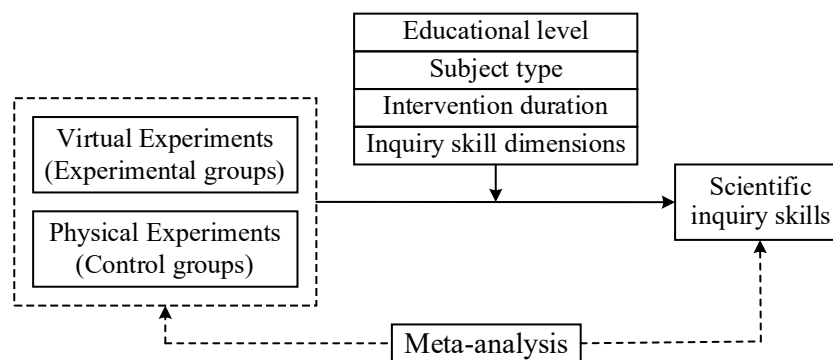


Fig. 1 Meta-analysis Research Framework

2 Methods

This study strictly adhered to the PRISMA 2020 Guidelines and PICOS Framework (Page et al., 2021). A three-level meta-analysis model was used to systematically examine the intervention effects of VE versus PE on students' scientific inquiry skills. This method corrects the non-independent effect size bias that traditional two-level models cannot address. It improves the accuracy of effect size estimation and the robustness of conclusions. The entire research process was conducted by two independent researchers trained in meta-analysis standards. Their training included literature screening, data coding, and effect size calculation. Disagreements were resolved through arbitration by a third-party meta-analysis expert. This ensured research reproducibility and conclusion reliability.

2.1 Definition of Core Concepts

To clarify research boundaries and avoid classification ambiguity, this study provided operational definitions for core concepts:

(1) Virtual experiments (VE): Digital teaching tools that fully construct virtual

experimental environments and objects through computer technology. Learners conduct inquiry activities via digital interaction devices without any real physical instruments or materials. This study focused on non-immersive VE. These are 2D screen-based systems without spatial immersion. Learners interact with them through peripherals such as mice and keyboards. Examples include PhET simulations and the IVPL Internet Virtual Physics Laboratory.

(2) Physical experiments (PE): Traditional experimental teaching activities conducted entirely with real physical instruments, materials, and natural environments. Learners complete the inquiry process through direct manipulation of physical objects.

(3) Combination of physical and virtual experiments (CPVE): Experimental forms that integrate both physical and virtual elements. All studies adopting CPVE intervention were strictly excluded to guarantee the purity of intervention effect comparison between VE and PE.

2.2 Literature Inclusion and Exclusion Criteria

Based on the operational definitions of core concepts, this study strictly adhered to the PICOS framework (Chandler et al., 2019). Inclusion and exclusion criteria were developed from four dimensions: study population, intervention measures, outcome indicators, and study design. This ensured the homogeneity of included studies and controlled for confounding variables.

(1) Study population: Students across all educational levels (primary, junior high, senior high, and university). Research scenarios were limited to natural science experiment teaching, including physics, chemistry, biology, and integrated science.

(2) Intervention and control measures: The experimental group received natural science experiment teaching using VE. The control group received identical teaching using PE. Intervention duration, course content, and task load were exactly the same for both groups. Included studies must have reported baseline equivalence between groups ($p > 0.05$).

(3) Outcome indicators: Studies must have included quantifiable evaluation indicators of scientific inquiry skills. These should be validated scales with established reliability and validity, or quantitative scores for at least one inquiry skill dimension.

(4) Study design: Only empirical comparative studies were included (randomized controlled trials and quasi-experimental study). Studies published up to December 31, 2025, including peer-reviewed advance online publications, were included. Preprints without peer review were excluded.

Exclusion criteria: Studies comparing VE only with traditional lecture-based

teaching without a PE control group; studies using CPVE as the intervention; duplicate publications of the same dataset; studies lacking sufficient statistical information to calculate effect sizes; and grey literature (including but not limited to conference abstracts, government reports, and unpublished manuscripts). Grey literature was excluded because it lacks rigorous peer review, has variable methodological quality, and is difficult to retrieve comprehensively. These factors may introduce systematic bias.

2.3 Literature Sources and Search Strategy

The literature search was conducted from March 1 to April 1, 2026. Five major international databases were searched: Springer Link, Web of Science, EBSCO, Scopus, and Google Scholar. The search strategy was built around core variables: (virtual experiment OR virtual lab OR computer simulation) AND (hands-on experiment OR physical lab) AND (scientific inquiry skills OR experimental ability). Synonyms were supplemented using the ERIC Thesaurus to ensure comprehensive retrieval. Search restrictions included: (1) publication date: January 1, 2000, to December 31, 2025; (2) language: English; (3) research field: interdisciplinary education and natural sciences; (4) study type: empirical research. After the initial search, backward snowballing was performed on the reference lists of all included studies. Forward snowballing was also conducted in the Web of Science database to identify all citing articles. A total of 2810 potentially relevant articles were identified through this search strategy.

2.4 Literature Screening

This study strictly adhered to the standardized PRISMA 2020 Guidelines framework (Moher et al., 2010). A four-stage screening process was adopted: automatic deduplication → title and abstract screening → full-text screening → final inclusion. In the first stage, automatic and manual deduplication yielded 2302 valid records. The second stage involved manual screening of titles and abstracts. The third stage involved full-text screening of remaining articles. The fourth stage confirmed final inclusion. A total of 2214 articles were excluded during the initial screening. After full-text review, 21 valid studies were finally included ($k = 21$). These studies provided 41 effect sizes ($n = 41$) with a total sample size of 3931 participants. Cohen's Kappa test was used to evaluate inter-rater reliability for screening. The Kappa coefficient was 0.896 ($p < 0.001$), indicating reliable screening results. The detailed search steps, screening criteria, and decision process are presented in Fig. 2.

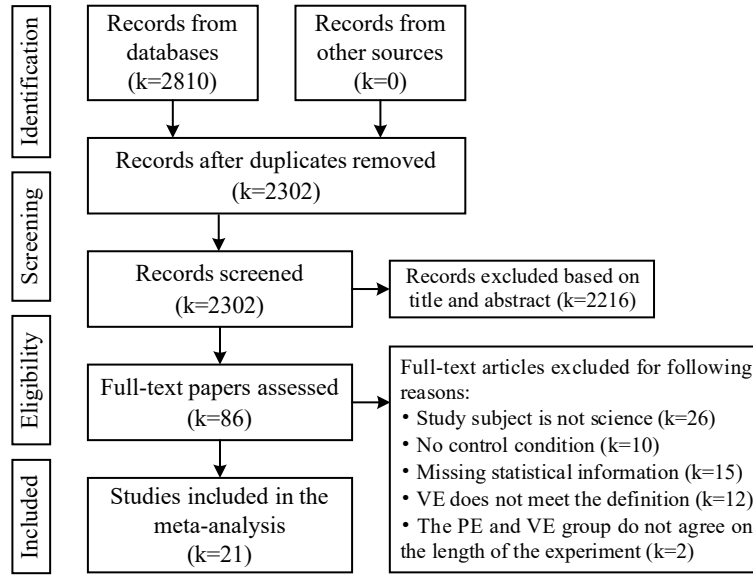


Fig. 2 PRISMA Flow Diagram of Study Selection

2.5 Literature Coding and Quality Control

This study followed the meta-analysis coding standards proposed by Card (2015). A unified coding manual was developed in advance. Two researchers received standardized training before conducting independent coding. Coding content included core information such as authors, publication year, educational level, subject type, intervention duration, inquiry skill dimension, sample size, and descriptive statistics for both experimental and control groups. Educational level was categorized into four levels: primary, junior high, senior high, and university (Ozdemir et al., 2018). Subject type was divided into four categories: physics, chemistry, biology, and integrated science. Intervention duration was classified into three levels: <4 weeks, 4-8 weeks, and >8 weeks (Nie et al., 2025). Inquiry skill dimensions were mapped to five categories based on the scientific inquiry framework of Pedaste et al. (2015): asking scientific questions, formulating testable hypotheses, designing controlled experiments, analyzing and interpreting data, and drawing evidence-based conclusions. This ensured the comparability and standardization of coding results.

To control for the impact of study quality on pooled effect sizes, this study used the Cochrane Risk of Bias Tool (for randomized controlled trials) and the Newcastle-Ottawa Scale (NOS) (for QES) to evaluate methodological quality. The use of dual tools is a consensus practice in educational meta-analysis (Higgins et al., 2003; Wells et al., 2000). The NOS has a maximum score of 9 points. Studies scoring <5 points were defined as low-quality. The Cochrane Risk of Bias Tool includes core domains: random sequence generation, allocation concealment, blinding of participants and

personnel, blinding of outcome assessment, and incomplete outcome data. Studies with ≥ 3 core domains judged as high risk were defined as high-risk bias studies. Quality assessment results were used in sensitivity analysis. The robustness of conclusions was verified by excluding low-quality and high-risk bias studies. After coding, Cohen's Kappa test was used to evaluate inter-rater reliability. The Kappa coefficient was 0.891 ($p < 0.001$), indicating stable and credible coding results.

2.6 Effect Size Calculation

This study used Hedges' g as the pooled effect size indicator. This index corrects for small sample bias and is more suitable than Cohen's d for meta-analyses with large sample size differences (Hedges, 1981). The direction of effect size was defined as follows: positive values indicate higher scientific inquiry skills scores in the VE group than in the PE group; negative values indicate higher scores in the PE group. If original studies reported means, standard deviations, and sample sizes, effect sizes were calculated using standard formulas. If only t -values, F -values, or p -values were reported, formulas recommended by Lipsey & Wilson (2001) were used for conversion (Lipsey & Wilson, 2001). If original studies reported ANCOVA-adjusted means and standard errors, these adjusted data were prioritized for effect size calculation to control for baseline differences.

For multiple effect sizes from the same study, those from independent samples (e.g., different grades or classes) were calculated separately. Those from the same non-independent sample (e.g., different inquiry skill dimensions) were all included in the analysis. The three-level meta-analysis model automatically corrected their correlations without prior pooling. Hedges' g is calculated by adjusting the standardized mean difference (d) with a correction factor (J) (Cohen, 1988; Hedges & Olkin, 2014). The specific formulas are as follows:

$$g = d * J \quad (1)$$

$$d = \frac{M_1 - M_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 1}}} \quad (2)$$

$$J = 1 - 3/[4(n_1 + n_2 - 2) - 1] \quad (3)$$

where M_1 , n_1 , and S_1 represent the mean, sample size, and standard deviation of the experimental group, respectively; M_2 , n_2 , and S_2 represent the mean, sample size, and standard deviation of the control group, respectively; and J is the small sample correction factor.

2.7 Three-level meta-analysis model

Among the 21 included studies, 11 reported multiple effect sizes (total 41). Traditional two-level meta-analysis cannot handle this nested data structure. This leads to underestimated standard errors and increased Type I error rates (Lai & Kwok, 2016). Therefore, this study adopted a three-level meta-analysis model. Total variance was decomposed into three independent levels: level 1 (sampling error of individual effect sizes), level 2 (within-study variance among different effect sizes), and level 3 (between-study variance among different studies).

The model used restricted maximum likelihood (REML) to estimate variance components. A compound symmetry (CS) covariance structure was used to handle correlations among effect sizes within the same study (Cheung, 2014). An unstructured covariance structure was not adopted because it requires estimating more parameters. This can lead to unstable estimation when the number of within-study effect sizes is small (average 1.95 per study in this research). The CS structure assumes equal correlations among all effect sizes within the same study. This matches the data characteristics of this study and is the most commonly used and robust covariance structure for nested effect sizes in educational meta-analysis.

The goodness of fit of the three-level model and the traditional two-level model was compared using the Akaike information criterion (AIC), Bayesian information criterion (BIC), and log-likelihood ratio test (LRT). The three-level model was selected if it had significantly lower AIC and BIC values and a significant LRT ($p < 0.05$). Otherwise, the traditional two-level model was used. Three-level I^2 decomposition was used to assess the degree of heterogeneity. $I^2 > 75\%$ indicates high heterogeneity, $I^2 = 50\%-75\%$ indicates moderate heterogeneity, and $I^2 < 50\%$ indicates low heterogeneity (Higgins et al., 2003). All statistical analyses were performed in R 4.6.0 using the metafor package version 3.0-2 (Viechtbauer, 2010).

2.8 Moderating Effect Analysis

The presence of heterogeneity in meta-analysis indicates that intervention effects are not constant but influenced by multiple moderating variables. To explore the sources of heterogeneity in the effects of VE on scientific inquiry skills, this study selected four core moderating variables: (1) educational level: primary, junior high, senior high, university; (2) subject type: physics, chemistry, biology, integrated science; (3) intervention duration: <4 weeks, 4-8 weeks, >8 weeks; (4) inquiry skill dimension: asking scientific questions, formulating testable hypotheses, designing controlled

experiments, analyzing and interpreting data, drawing evidence-based conclusions, overall scientific inquiry skills.

For categorical moderating variables, subgroup analysis was conducted using a three-level mixed-effects model to calculate pooled effect sizes for each subgroup. The significance of moderating variables was determined using the LRT to compare the fit of nested models, combined with Q tests. The significance level for moderating effects was set at $p < 0.05$. If the overall heterogeneity test was not significant ($p > 0.05$), a three-level fixed-effects model (assuming zero between-study variance at level 3) was used to pool effect sizes. If significant heterogeneity existed, a three-level mixed-effects model was used for moderating effect analysis to systematically explore sources of heterogeneity.

Design characteristics of VE (such as interactivity, feedback type, and immersion) were not included as moderating variables. This was primarily due to insufficient reporting of these characteristics in included studies, which prevented systematic coding. Future research can focus on the impact of VE design characteristics on inquiry skills to provide evidence for optimizing VE design.

2.9 Publication Bias and Sensitivity Analysis

To assess the robustness of research conclusions, this study comprehensively evaluated publication bias using three-level Egger's regression, three-level trim-and-fill method, Orwin's fail-safe N , and three-level funnel plot (Fernández-Castilla et al., 2021). Orwin's fail-safe N calculates the number of unpublished null-effect studies needed to reduce the pooled effect size to a clinically meaningful threshold. This study set the educational meaningful threshold at Hedges' $g = 0.2$. This threshold not only conforms to the classic standard of Cohen (2013) but is also widely adopted by authoritative educational studies such as Hattie (2008). It represents the smallest meaningful effect of educational interventions (Hattie, 2008). If the fail-safe N is greater than " $5k+10$ " (where k is the number of included studies), the impact of publication bias on research conclusions is considered negligible (Rosenthal, 1979).

Sensitivity analysis was systematically conducted using three complementary indicators: Studentized deleted residual (SDR), Cook's distance, and DFBETAS statistics. SDR identifies outliers by sequentially deleting each effect size and refitting the model to calculate the standardized residual. The judgment criterion is $|\text{SDR}| > 1.96$ (Viechtbauer, 2010). In addition, this study conducted leave-one-out sensitivity analysis (sequentially deleting each study and refitting the model). Low-quality and

high-risk bias studies were also excluded and re-analyzed to further verify the robustness of conclusions.

3 Results

This study strictly followed the standard reporting process for three-level meta-analysis. It sequentially presents the basic characteristics of included studies, publication bias tests, and sensitivity analysis. This comprehensively answers the intervention effects of VE on students' scientific inquiry skills and their sources of heterogeneity.

3.1 Basic Characteristics of Included Studies

The literature search covered the period from January 1, 2000, to December 31, 2025. A total of 21 studies were finally included ($k = 21$), yielding 41 effect sizes ($n = 41$) with a total sample size of 3931 participants. The included studies covered all educational levels from primary school to university. They involved four disciplinary fields: physics, chemistry, biology, and integrated science. The distribution of participants by educational level was: primary school (3 studies, 7 effect sizes), junior high school (6 studies, 12 effect sizes), senior high school (7 studies, 16 effect sizes), and university (5 studies, 6 effect sizes). The distribution of effect sizes by inquiry skill dimension was: asking scientific questions (2), formulating testable hypotheses (3), designing controlled experiments (4), analyzing and interpreting data (4), drawing evidence-based conclusions (3), and overall scientific inquiry skills (25). The basic characteristics of included studies are presented in Table 1.

Table 1 Basic Characteristics of Included Studies

Author	Year	Educational level	Discipline type	Intervention duration	Inquiry skill dimensions	Sample size	
						PE	VE
Anam et al.	2025	primary school	Physics	<4weeks	Formulating Testable Hypotheses	25	25
Anam et al.	2025	primary school	Physics	<4weeks	Analyzing and Interpreting Data	25	25
Anam et al.	2025	primary school	Physics	<4weeks	Designing Controlled Experiments	25	25
Anam et al.	2025	primary school	Physics	<4weeks	Drawing Evidence-Based Conclusions	25	25
Chen et al.	2024	junior high school	Chemistry	4-8weeks	Asking Scientific Questions	42	43
Chen et al.	2024	junior high school	Chemistry	4-8weeks	Formulating Testable Hypotheses	42	43
Chen et al.	2024	junior high school	Chemistry	4-8weeks	Designing Controlled Experiments	42	43
Chen et al.	2024	junior high school	Chemistry	4-8weeks	Drawing Evidence-Based Conclusions	42	43
Chen et al.	2024	junior high school	Chemistry	4-8weeks	Analyzing and Interpreting Data	42	43
Chen et al.	2020	senior high school	integrated science	>8weeks	overall scientific inquiry skills	83	79
Chen et al.	2020	senior high school	integrated science	>8weeks	overall scientific inquiry skills	83	79
Chen et al.	2020	senior high school	integrated science	>8weeks	overall scientific inquiry skills	83	79
Dodds	2013	senior high school	Biology	<4weeks	overall scientific inquiry skills	33	33
Efstathiou et al.	2018	primary school	Physics	<4weeks	overall scientific inquiry skills	12	14

Author	Year	Educational level	Discipline type	Intervention duration	Inquiry skill dimensions	Sample size	
						PE	VE
Farrokhnia & Esmailpour	2010	university	Physics	>8weeks	overall scientific inquiry skills	30	35
Hu-Au, E.	2021	university	Chemistry	<4weeks	overall scientific inquiry skills	20	20
Huppert et al.	2002	senior high school	Biology	4-8weeks	overall scientific inquiry skills	89	82
Huppert et al.	2002	senior high school	Biology	4-8weeks	overall scientific inquiry skills	89	82
Huppert et al.	2002	senior high school	Biology	4-8weeks	overall scientific inquiry skills	89	82
Husnaini & Chen	2019	senior high school	Physics	<4weeks	Analyzing and Interpreting Data	34	34
Husnaini & Chen	2019	senior high school	Physics	<4weeks	Designing Controlled Experiments	34	34
Ifthinan & Atun	2019	senior high school	integrated science	4-8weeks	Asking Scientific Questions	31	31
Ifthinan & Atun	2019	senior high school	integrated science	4-8weeks	Formulating Testable Hypotheses	31	31
Ifthinan & Atun	2019	senior high school	integrated science	4-8weeks	Designing Controlled Experiments	31	31
Ifthinan & Atun	2019	senior high school	integrated science	4-8weeks	Analyzing and Interpreting Data	31	31
Ifthinan & Atun	2019	senior high school	integrated science	4-8weeks	Drawing Evidence-Based Conclusions	31	31
Juwariyah & Latifah	2017	junior high school	Physics	4-8weeks	overall scientific inquiry skills	32	32
Kapici	2025	junior high school	Physics	<4weeks	overall scientific inquiry skills	77	79
Kapici & Akcay	2020	university	integrated science	4-8weeks	overall scientific inquiry skills	21	21
Kapici et al.	2022	junior high school	Physics	4-8weeks	overall scientific inquiry skills	28	28
Kapici et al.	2022	junior high school	Physics	4-8weeks	overall scientific inquiry skills	33	27
Kapici et al.	2019	junior high school	Physics	4-8weeks	overall scientific inquiry skills	33	34
Ogbuanya & Onele	2018	university	Physics	4-8weeks	overall scientific inquiry skills	68	74
Oser, R., & Fraser, B. J.	2015	senior high school	Biology	>8weeks	overall scientific inquiry skills	153	166
Pyatt et al.	2012	junior high school	integrated science	>8weeks	overall scientific inquiry skills	98	86
Pyatt et al.	2012	junior high school	integrated science	>8weeks	overall scientific inquiry skills	86	98
Toth et al.	2014	university	Biology	4-8weeks	overall scientific inquiry skills	16	16
Toth et al.	2014	university	Biology	4-8weeks	overall scientific inquiry skills	12	12
Triona & Klahr	2003	primary school	Physics	<4weeks	overall scientific inquiry skills	46	46
Triona & Klahr	2003	primary school	Physics	<4weeks	overall scientific inquiry skills	46	46
Yang & Heh	2007	senior high school	Physics	4-8weeks	overall scientific inquiry skills	75	75

The methodological quality of included studies was generally good. Among 14 randomized controlled trials, 12 (85.7%) had low risk of bias. The average Newcastle-Ottawa Scale (NOS) score for 7 quasi-experimental study was 6.8 out of 9, all meeting medium or higher quality standards.

3.2 Publication Bias Test

This study used a three-level random-effects model for comprehensive publication bias assessment (Borenstein et al., 2021). The three-level Egger's regression test showed no statistically significant regression coefficient for standard error ($\beta = -0.650, p > 0.05$). This confirmed no significant publication bias was detected. The three-level trim-and-fill method based on the random-effects model identified no missing studies to be imputed, further validating this conclusion. The Orwin's fail-safe N was calculated as

2585, far exceeding the critical value of 115 ($5k + 10$, $k = 21$). This indicates the impact of publication bias on research conclusions is negligible. The three-level funnel plot is presented in Fig. 3.

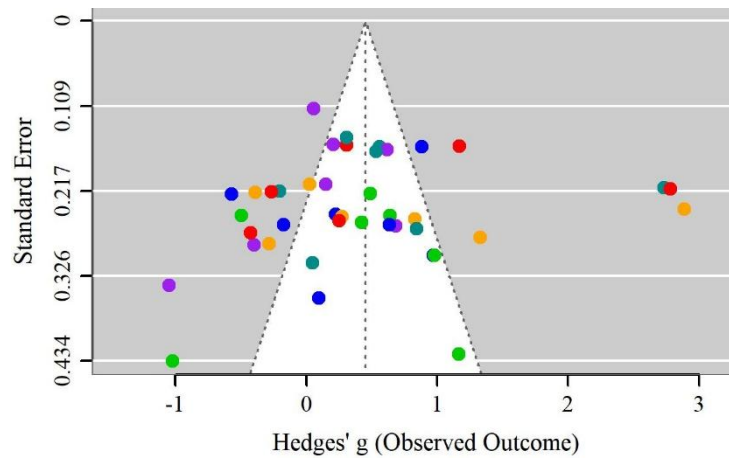


Fig. 3 Three-Level Funnel Plot for Publication Bias

3.3 Sensitivity Analysis

This study conducted systematic sensitivity analysis using three complementary indicators: studentized deleted residual (SDR), Cook's distance, and DFBETAS statistics. The outlier judgment criterion was $|SDR| > 1.96$ (Viechtbauer, 2010). The results identified three statistically deviant effect sizes (IDs: 17, 18, 33) (Fig. 4). However, all effect sizes had Cook's distance < 0.45 (Fig. 5) and DFBETAS values within the $[-1, 1]$ interval (Fig. 6). Combining the results of all three indicators, the three SDR outliers did not have a substantial impact on the overall effect size estimation.

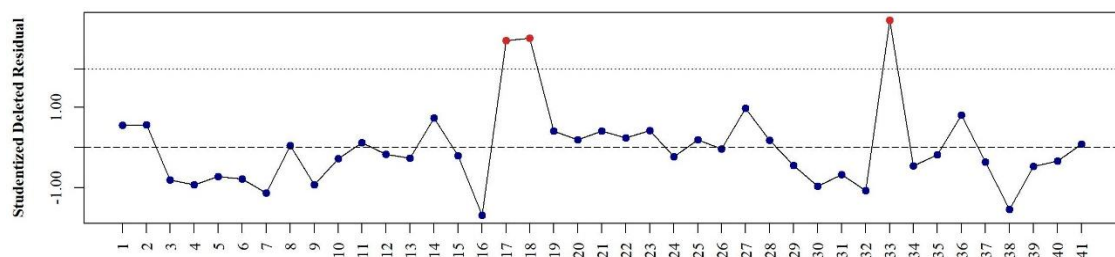


Fig. 4 Studentized Deleted Residual Plot

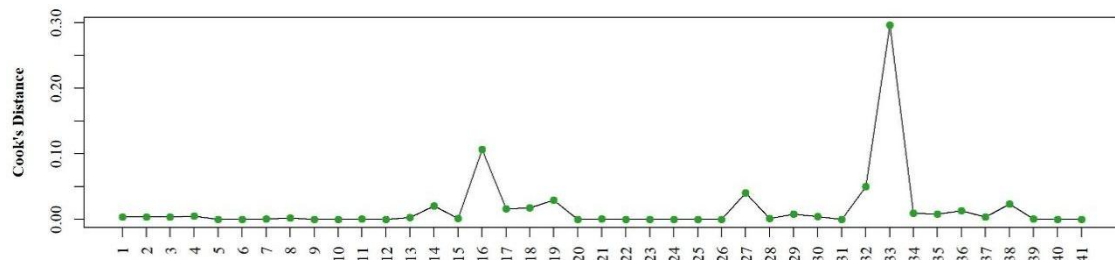


Fig. 5 Cook's Distance Plot

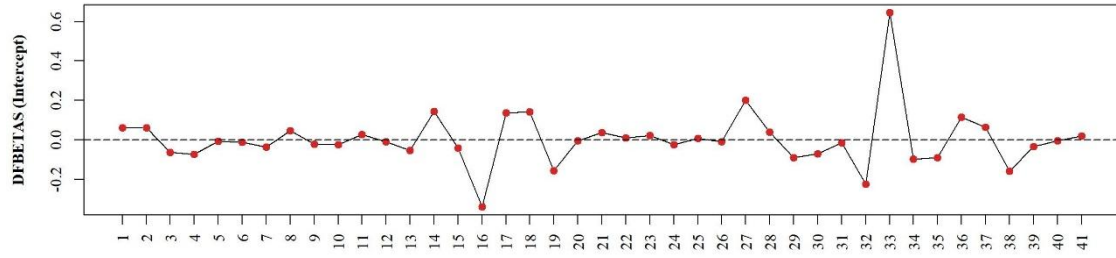


Fig. 6 DFBETAS Plot

The leave-one-out sensitivity analysis at the study level showed that after excluding any single study, the pooled effect size fluctuated between 0.353 and 0.514. This was highly consistent with the main model effect size (0.454), and all results remained statistically significant ($p < 0.05$). This indicates the research conclusions have high robustness. After excluding the three SDR outliers, the pooled effect size increased slightly from Hedges' $g = 0.454$ to Hedges' $g = 0.530$. It remained statistically significant ($p < 0.01$), and the 95% CI highly overlapped with the original result (Table 2). The effect size change was only 16.7%, far below the 50% empirical threshold for conclusion instability.

Table 2 Comparison of Overall Effect Sizes Between Original Data and Outlier-Excluded Data

Analysis sample	k	n	Hedges' g	SE	95% CI	t	df	p
Complete data	21	41	0.454	0.185	[0.081, 0.827]	2.458	40	0.018
After excluding outliers	19	38	0.530	0.171	[0.184, 0.876]	3.108	37	0.004

3.4 Overall Effect Size and Heterogeneity Analysis

To verify the necessity of the three-level model, this study constructed two nested two-level simplified models for comparison. All models used the Kenward-Roger method to estimate degrees of freedom, which is more robust for small samples (Cheung, 2014). The log-likelihood ratio test (LRT) results showed that the full three-level model was significantly superior to the two simplified models (all $p < 0.01$). It also had lower AIC and BIC values (three-level model: AIC = 99.507, BIC = 104.574; Model 1: AIC = 175.150, BIC = 178.528; Model 2: AIC = 106.981, BIC = 110.359). This indicates the three-level model is the best-fitting model for this study's data.

This study also used the traditional two-level model for robustness comparison. The results showed a pooled effect size of Hedges' $g = 0.446$ ($p < 0.020$), highly consistent with the three-level model result. The point estimate difference was only 1.8%. Despite the minimal point estimate difference, the traditional two-level model underestimated the standard error by 41%, which would lead to serious statistical inference bias. Therefore, this study adopted the three-level random-effects model as the main analysis model.

The three-level random-effects model results based on all 41 effect sizes showed that VE had a significant small-to-moderate positive effect on students' scientific inquiry skills compared with PE (Hedges' $g = 0.454$, 95% CI [0.081, 0.827], $p < 0.05$). The effect size scatter plot showed that 73.2% (30/41) of effect sizes were positive (Fig. 7). The forest plot visually presented the distribution of effect sizes and confidence intervals for each independent study (Fig. 8). According to Cohen's (1988) effect size criteria (Cohen, 1988), this result approaches a medium effect. Combined with the average effect size range for educational intervention meta-analyses (0.3-0.5) (Card, 2015), this indicates the improvement effect of VE on scientific inquiry skills is above the medium level in the educational field. Furthermore, because the three-level model corrected for nested effect size bias, the confidence interval of the effect size in this study is narrower and the estimation is more accurate. This three-level meta-analysis, which strictly separated pure VE and pure PE effects, provides clearer and more unbiased evidence for the debate on "whether VE can replace PE to improve scientific inquiry skills".

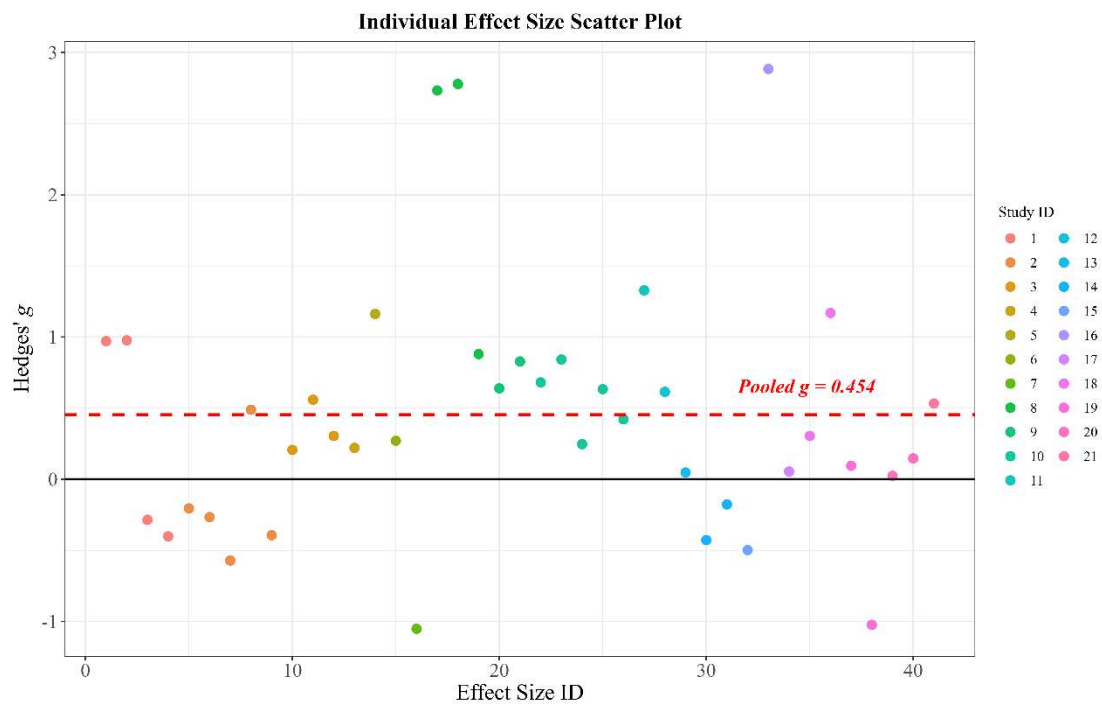


Fig. 7 Effect Size Scatter Plot

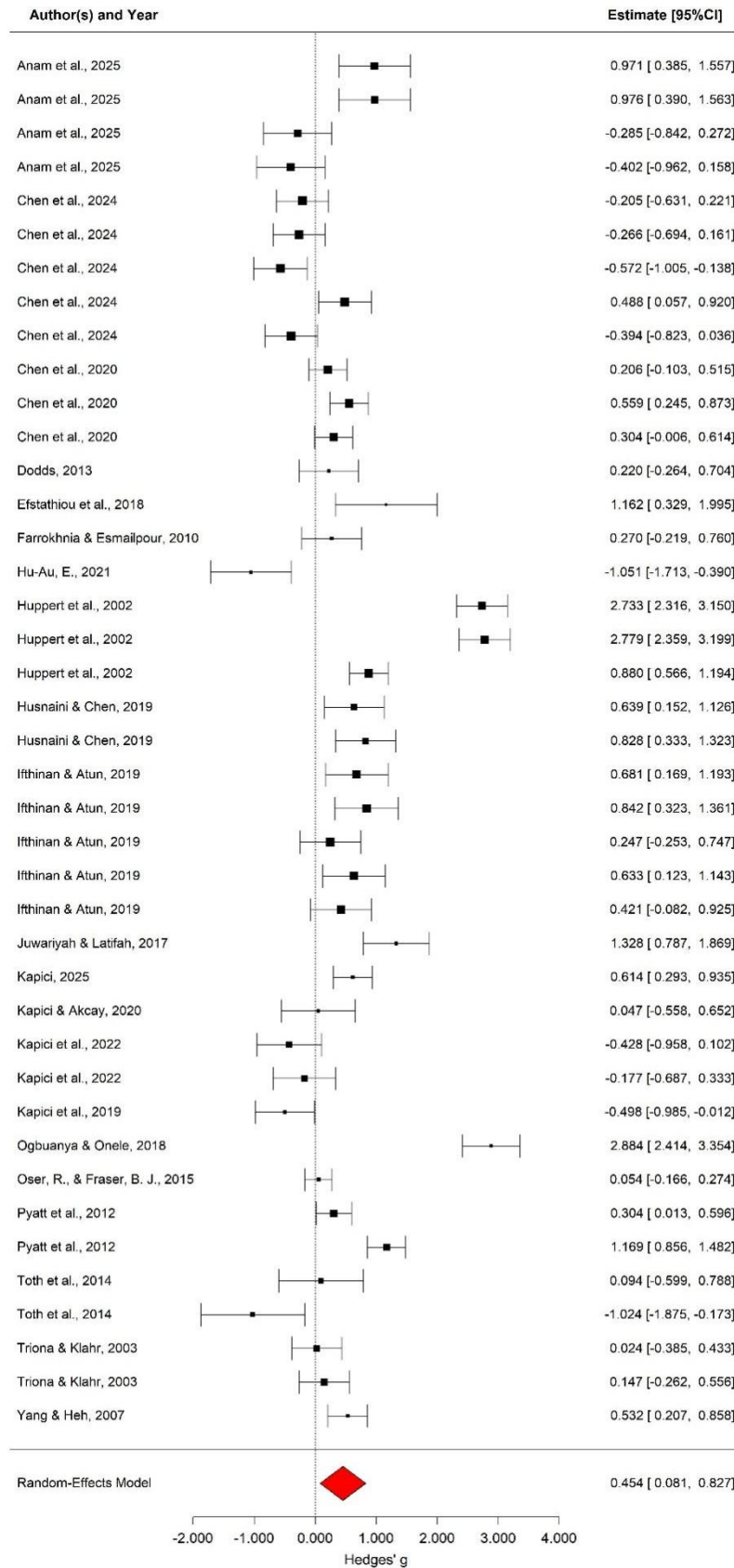


Fig. 8 Forest Plot of Included Studies

The Q test results indicated significant overall heterogeneity among the effect sizes of included studies ($Q(40) = 551.786, p < 0.001$). The heterogeneity decomposition

results showed that 31.32% of total heterogeneity came from multiple effect sizes within the same study (Level 2: $\sigma^2 = 0.252$, $p < 0.001$, $I^2 = 31.32\%$). 62.88% came from between-study variation (Level 3: $\sigma^2 = 0.506$, $p < 0.01$, $I^2 = 62.88\%$). The remaining 5.80% was sampling error. The intraclass correlation coefficient (ICC) of the three-level model was 0.647, which is a high level in meta-analysis. This indicates extremely strong nested clustering effects in the data. Based on the ICC = 0.647 calculation, using the traditional two-level model would underestimate the standard error by 41%, an extremely high degree of bias. This provides statistical justification for choosing the three-level model in this study. This result suggests that the effects of VE on scientific inquiry skills have significant heterogeneity, which mainly comes from study-level characteristic differences such as intervention duration, VE design, and disciplinary field.

This study conducted statistical power analysis for moderating effects based on the three-level meta-analysis model. The results showed that the statistical power to detect a moderate moderating effect ($R^2 = 0.10$) was 1.00, far exceeding the academic standard of 0.80. This indicates the study has sufficient statistical power to detect moderate-strength moderating effects. The statistical power to detect a small moderating effect ($R^2 = 0.05$) was 0.59, which did not reach the acceptable standard. These results indicate that non-significant moderating effects truly represent no moderating effect for moderate effects. For weak moderating effects, the non-significant results are mainly limited by insufficient statistical power, and potential weak moderating effects cannot be ruled out.

3.5 Effect Size Analysis of Scientific Inquiry Skills Subdimensions

To explore whether the effects of VE differ across different dimensions of inquiry skills, this study conducted univariate three-level meta-analysis for dimensions with ≥ 3 effect sizes. The “asking scientific questions” dimension had only 2 effect sizes with insufficient statistical power and was not included in the analysis. The effect sizes and heterogeneity test results for each dimension are presented in Table 3.

Table 3 Univariate Three-Level Random-Effects Meta-Analysis Results for Each Dimension of scientific inquiry skills

Dependent variable	<i>k</i>	<i>n</i>	Hedges' <i>g</i>	SE	95% CI	<i>p</i>	τ^2	Statistical power
Overall scientific inquiry skills	17	25	0.488	0.235	[0.003, 0.972]	0.049	0.972	0.069
Formulating testable hypotheses	3	3	0.497	0.400	[-1.226, 2.220]	0.341	0.413	0.107
Designing controlled experiments	4	4	0.052	0.312	[-0.941, 1.045]	0.879	0.325	0.030
Analyzing and interpreting data	4	4	0.445	0.303	[-0.519, 1.410]	0.238	0.301	0.110
Drawing evidence-based conclusions	3	3	0.190	0.277	[-1.001, 1.381]	0.563	0.165	0.059

The results in Table 3 show that in the subsample containing only dimensional data, the significant positive effect of VE on the overall dimension of inquiry skills was reconfirmed (Hedges' $g = 0.488$, $p < 0.05$). However, none of the subdimension effects reached statistical significance ($p > 0.05$). Among them, the “formulating testable hypotheses” dimension had the largest effect size (Hedges' $g = 0.497$), and the “designing controlled experiments” dimension had the smallest effect size (Hedges' $g = 0.052$). However, neither of their 95% CIs included zero.

Statistical power analysis was conducted using the R metapower package ($\alpha = 0.05$, two-tailed test, using measured heterogeneity τ^2 for each dimension). The results showed that the statistical power for each dimension was: formulating testable hypotheses (0.107), analyzing and interpreting data (0.110), designing controlled experiments (0.030), and drawing evidence-based conclusions (0.059). All were far below the acceptable standard of 0.8 (Cohen, 2013). This suggests that the non-significant dimensional effects are likely due to low statistical power caused by insufficient sample size. The possibility of true effects cannot be excluded (Borenstein et al., 2021).

3.6 Moderating Effect Analysis

Given the significant heterogeneity in the overall effect ($Q(40) = 551.786$, $p < 0.001$), this study used a three-level mixed-effects model for moderating effect analysis. It explored the moderating roles of educational level, subject type, and intervention duration. The multilevel R^2 index was used to quantify the explanatory rate of each moderating variable for heterogeneity (Cheung, 2015) (Table 4).

Table 4 Moderating Effect Analysis of VE vs PE

Moderator variable	k	n	Hedges' g	95%CI	F	df_1/df_2	p	Level 2/3 variance
Educational level	21	41			1.590	4/37	0.197	0.257/0.578
Primary school	3	7	0.446	[-0.584,1.477]			0.386	
Junior high school	6	12	0.252	[-0.479,0.983]			0.490	
Senior high school	7	16	0.690	[0.026,1.353]			0.042	
University	5	6	0.344	[-0.509,1.196]			0.419	
Discipline type	21	41			2.279	4/37	0.079	0.247/0.522
Science	4	11	0.451	[-0.375,1.277]			0.275	
Physics	11	17	0.603	[0.071,1.136]			0.028	
Chemistry	2	6	-0.532	[-1.731,0.668]			0.375	
Biology	4	7	0.575	[-0.290,1.440]			0.186	
Intervention duration	21	41			1.999	3/38	0.131	0.250/0.574
< 4weeks	7	12	0.300	[-0.388,0.988]			0.383	
4-8weeks	10	22	0.590	[0.027,1.153]			0.041	

> 8weeks	4	7	0.370	[-0.515,1.255]			0.402	
Inquiry skill dimensions	21	41			1.244	6/35	0.308	0.288/0.529
Overall scientific inquiry skills	17	25	0.488	[0.047,0.929]			0.031	
Asking scientific questions	2	2	0.394	[-0.770,1.557]			0.497	
Formulating testable hypotheses	3	3	0.620	[-0.770,1.557]			0.233	
Designing controlled experiments	4	4	0.056	[-0.896,1.009]			0.905	
Analyzing and interpreting data	4	4	0.450	[-0.504,1.403]			0.345	
Drawing evidence-based conclusions	3	3	0.310	[-0.726,1.345]			0.547	

The variance decomposition results of all moderating models were consistent with the overall model. heterogeneity mainly came from between-study variation (accounting for > 65%). The moderating effect test results showed that the moderating effects of educational level ($F(4,37) = 1.590, p > 0.05$), intervention duration ($F(3,38) = 1.999, p > 0.05$), and inquiry skill dimension ($F(6,35) = 1.244, p > 0.05$) did not reach significant levels. The moderating effect of subject type approached significance ($p = 0.079$) but still did not meet the 0.05 significance threshold. The multilevel R^2 results showed that all moderating variables explained 0% of between-study heterogeneity (Level 3). Only subject type explained 2.2% of within-study heterogeneity (Level 2), and intervention duration explained 1.0%. All other moderating variables had 0% explanatory rates. This indicates that the moderating variables included in this study can barely explain the heterogeneity of effect sizes. Most heterogeneity comes from unincluded moderating variables such as VE design characteristics, teacher teaching methods, and student individual differences.

It is important to note that non-significant moderating effects indicate no statistically significant differences between subgroups. No conclusion can be drawn that one subgroup effect is superior to another. The subgroup results reported below only indicate that the effect size of that subgroup is significantly different from zero. They only represent potential effect directions and require verification with larger sample studies. Subgroup analysis results showed that significant subgroup effects were only detected in senior high school (Hedges' $g = 0.690, p < 0.05$), physics discipline ($g = 0.603, p < 0.05$), and medium intervention duration of 4-8 weeks (Hedges' $g = 0.590, p < 0.05$). The chemistry subgroup effect size was negative and non-significant (Hedges' $g = -0.532, p > 0.05$), with a statistical power of only 0.21. It cannot reliably detect true effects, so no conclusion can be drawn that VE is less effective in chemistry. The effect size of the experimental design dimension was close to zero (Hedges' $g = 0.056, p > 0.05$). This result is consistent with the general observation in the field that current VE are mostly confirmatory experiments lacking open design space.

In summary, this study found no significant moderating effects of educational level,

intervention duration, or inquiry skill dimension on VE effects. Only subject type suggested a potential moderating trend. These results indicate that the intervention effects of VE have certain universality and are not significantly affected by the above factors. This has important guiding significance for educational policy making and teaching practice. Since the statistical power of this study to detect small moderating effects ($R^2 = 0.05$) was only 0.59, the possibility of weak moderating effects cannot be ruled out. Further verification with large sample studies is needed.

4 Discussion and Conclusions

4.1 Discussion

4.1.1 Overall Effect of VE on Students' Scientific Inquiry Skills

This study found that VE had a significant small-to-moderate positive effect on students' scientific inquiry skills compared with PE (Hedges' $g = 0.454$, $p < 0.05$). This was under the premise of controlling teaching content, intervention duration, and total inquiry task load. According to Cohen's (1988) effect size evaluation criteria, this value approaches a medium effect level. Combined with the average effect size range for educational intervention meta-analyses (0.3-0.5) reported by Card (2015), this result indicates that the improvement effect of VE is at the upper limit of the average educational intervention level. It has clear practical application value.

The advantage of using a three-level meta-analysis is its ability to handle multiple non-independent effect sizes from the same study. The ICC of the three-level model was 0.647, which is a high level in educational meta-analysis. This indicates extremely strong clustering correlation in the data. Model comparison results showed that the point estimate difference between the two models was only 0.008 (1.8%) (three-level model: Hedges' $g = 0.454$; two-level model: Hedges' $g = 0.446$). This suggests that the nested data structure has no substantial impact on the estimation of effect size magnitude. However, based on the ICC = 0.647 calculation, using the traditional two-level model would underestimate the standard error by 41%, leading to serious statistical inference bias. This finding clarifies that nested data structures mainly affect statistical inference accuracy rather than the true effect size itself. It further resolves the field's understanding of how nested data influence effects. This study also used the traditional two-level model for robustness comparison. The results showed a pooled effect size of Hedges' $g = 0.446$ ($p < 0.05$), consistent with the three-level model result and both reaching significance. This further verified the robustness of the conclusion.

Comparison with existing meta-analytic studies shows that the results of this study are highly consistent with international mainstream conclusions, but with clear research

boundaries and methodological differences. Compared with Song et al.'s (2025) chemistry meta-analysis using traditional lecture-based teaching as the control group (Hedges' $g = 0.385$), this study used PE as the parallel control. The difference in control group intervention form is the core reason for the slight increase in effect size. Compared with the two-level meta-analyses by Cao & Hsu (2023) and Castro (2025), this study corrected for within-study effect size dependence bias through three-level modeling, optimizing parameter estimation accuracy. Compared with Wang et al.'s (2025) limited sample integration, this study expanded to include recent empirical literature from 2024-2025, covering all educational levels and multiple natural science disciplines. The external validity of the conclusions has been further improved.

Although the results of this study are consistent with international mainstream empirical laws, existing single theoretical perspectives cannot fully explain the mechanism of effectiveness differentiation between virtual and physical experiments in different teaching scenarios. Neither Cognitive Load Theory (CLT) nor Embodied Cognition Theory (ECT) alone can completely explain the differentiated effectiveness of VE and PE. Therefore, this study integrates the core viewpoints of CLT and ECT. It systematically explains the relative effect differences between virtual and physical experiments from the perspective of cognitive-embodied dynamic balance. This explanatory perspective complements the “dual-pathway educational model” proposed by Zou et al. (2025), but focuses more on the dual characteristics of scientific inquiry skills: “cognitive processing + real-world operation”. The dual-pathway educational model emphasizes two independent learning paths of cognition and embodiment. In contrast, the cognitive-embodied dynamic balance perspective highlights the dynamic balance relationship between the two paths. It argues that the relative effect of VE is the net value of cognitive gains minus embodied losses. This perspective can clearly explain the differentiated characteristics of VE effects in different contexts.

From the deductive logic of this perspective, when inquiry tasks focus on high-order cognitive processing such as abstract logical deduction and model construction, the benefits of VE in simplifying redundant operations and optimizing cognitive resource allocation can cover the losses from lack of real embodied interaction. This ultimately results in relative gains. The significant positive effect in the physics discipline subgroup (Hedges' $g = 0.603$, $p < 0.05$) is an empirical manifestation of this mechanism. Physics has the highest degree of abstract principles among natural science disciplines. Inquiry activities have low dependence on concrete equipment operations, adapting well to the digital visualization advantages of VE. Conversely, if inquiry tasks

rely on refined physical operation, on-site troubleshooting, and multi-sensory real-time feedback, the embodied educational advantages of PE will offset the cognitive gains of VE. The near-zero effect size in the experimental design dimension (Hedges' $g = 0.052$, $p > 0.05$) confirms this pattern. The experimental design process requires judging the feasibility of schemes based on the physical characteristics of equipment. Current mainstream commercial VE mostly use preset operation procedures, lacking open independent design space. The shortcomings of missing embodied experience are significantly amplified. This is consistent with Zacharia & De Jong's (2014) conclusion that PE have more advantages in cultivating students' experimental design capabilities.

The four moderating variables (educational level, subject type, intervention duration, and inquiry skill dimension) essentially achieve contextual regulation of VE intervention effects by changing the relative weights of cognitive gains and embodied losses. The three significant subgroup effects observed in this study (senior high school, physics, and 4-8 week intervention) can all be mechanistically explained based on the cognitive-embodied dynamic balance perspective. This study also clarifies the applicable boundary of this theoretical perspective: it only applies to parallel comparison studies of non-immersive VE and pure PE. It does not apply to research scenarios of virtual-physical integrated teaching or VE compared with traditional lectures. The explanatory power for immersive VE still requires subsequent special empirical verification.

4.1.2 Moderating Effect Analysis of VE on Scientific Inquiry Skills

Based on the overall effect heterogeneity test results ($Q = 551.786$, $p < 0.001$), analysis was conducted around the four moderating variables. Non-significant overall moderating effects and significant local subgroup effects are not statistically contradictory (Borenstein et al., 2021). The former indicates that the variable cannot systematically explain the overall cross-study heterogeneity. The latter only indicates that VE have positive gains relative to PE within a specific sample. All subgroup results are exploratory findings and cannot be generalized to global scenarios.

The overall moderating test results for the four variables showed that the moderating effects of educational level, intervention duration, and inquiry skill dimension were not significant ($p > 0.05$). The moderating effect of subject type approached the significance threshold ($p = 0.079$). This study had a statistical power of 1.00 to detect moderate moderating effects ($R^2 = 0.10$), exceeding the 0.8 threshold. This confirms that there are no independent moderating effects of moderate or higher intensity among the four variables. For small weak moderating effects ($R^2 = 0.05$), the

statistical power was only 0.59, not reaching the 0.8 threshold. The possibility of weak moderating effects cannot be completely ruled out statistically (Liu, 2015).

Based on the above statistical power analysis, the deviation between this study's theoretical expectations and empirical results can be attributed to the following objective factors: (1) Insufficient statistical power for weak effects, making it difficult to capture small moderating patterns with the existing sample size (Cumming, 2013). (2) Inconsistent operational definitions of moderating variables in original literature (e.g., some literature counted intervention duration as single class hours, others as total project duration). Measurement heterogeneity diluted the explanatory power of variables (Lipsey & Wilson, 2001). (3) Limited by original literature reporting, this study did not include moderating variables such as inquiry structure, VE design characteristics, and teacher technological literacy. Omitted variable interference masked the true effects of target variables (Hunter & Schmidt, 2004). (4) Original theoretical hypotheses were mostly deduced based on independent effects of single variables, ignoring interactive coupling characteristics between variables. True intervention effects are driven by multiple variables synergistically, greatly compressing the independent moderating amplitude of single variables. These limitations point the way for future research. Future studies can expand literature to include key moderating variables and deepen research on variable relationships.

In the educational level dimension, non-significant overall moderation indicates that the educational effectiveness of VE has statistical universality across all educational levels from primary school to university. This conclusion still needs to be consolidated by subsequent expanded cross-regional large samples. A significant positive effect was only observed in the senior high school subgroup (Hedges' $g = 0.690$, $p < 0.05$). From the perspective of cognitive development laws, senior high school students' abstract logical thinking has entered a mature stage (Piaget, 1976). They can fully utilize VE visualization simulations to break through the constraints of concrete equipment. Additionally, senior high school science experiment topics have a higher proportion of abstract content, making cognitive gains dominant in the cognitive gain-embodied loss trade-off system. Meanwhile, the within-study heterogeneity (Level 2 variance) of this subgroup was the lowest among all educational levels, with higher sample homogeneity, making it easier to detect true effects.

In the intervention duration dimension, non-significant overall moderation proves that the relative gains of VE do not show a linear increasing trend with intervention duration. Only the medium-duration intervention subgroup of 4-8 weeks was

statistically significant (Hedges' $g = 0.590, p < 0.05$). This duration matches the regular unit teaching arrangement rhythm in primary and secondary schools (Kali, 2021). It can accumulate sufficient inquiry experience while avoiding learner burnout caused by overly long cycles, thus optimizing the cognitive gain-embodied loss ratio. The within-study heterogeneity of this subgroup was also the lowest among all duration groups, with better conclusion robustness than other groups.

In the subject type dimension, the overall moderating effect was close to significance, indicating potential differentiation trends in VE effectiveness across different disciplines. This is worthy of further exploration with expanded samples in the future. Only the physics subgroup had a significantly positive effect (Hedges' $g = 0.603, p < 0.05$). The within-study heterogeneity of this group was the lowest among all disciplines, with the highest conclusion credibility. The chemistry subgroup had a negative effect size (Hedges' $g = -0.532, p > 0.05$), but this subgroup only included 2 studies (6 effect sizes). The statistical power for moderate effect sizes was less than 0.3, meaning that even if a true effect exists, there is less than a 30% chance of detecting it. Therefore, no definitive conclusion can be drawn that "VE are inferior to PE in chemistry". The related negative trend only serves as an exploratory inference to be verified.

In the inquiry skill dimension, non-significant overall moderation indicates that except for the experimental design dimension, the gain amplitudes of VE for the other four cognitive-oriented inquiry skills tend to be balanced. The overall inquiry skills summary dimension had the largest sample size and highest statistical power, making it easier to detect true effects (Hedges' $g = 0.488, p < 0.05$). The effect size of the experimental design dimension was close to zero (Hedges' $g = 0.052$), indicating the current industrial status that mainstream commercial VE products generally lack open independent design modules and have fixed experimental operation procedures. Based on this, an exploratory inference can be formed: PE have relative advantages in cultivating practice-oriented experimental design capabilities. This inference still needs to be repeatedly verified by large sample studies using unified standardized assessment tools.

4.1.3 Theoretical Contributions

This study forms marginal academic innovations from three dimensions: method optimization, theoretical integration, and empirical supplementation. It provides new evidence and perspectives for the field of VE empowering scientific inquiry.

(1) Methodologically, this study verified the adaptive advantages of the three-level

meta-analysis model for highly nested non-independent effect size data in this field through likelihood ratio tests and AIC/BIC indicators. Based on this study's data ($ICC = 0.647$, high nesting degree), model comparison showed that the two-level meta-analysis model underestimated the standard error by an average of 41%, but the effect size estimation difference was only 0.008. This confirms that nested data mainly affect statistical inference accuracy rather than true effect size magnitude, clarifying the core statistical bias mechanism in educational technology meta-analysis. This study standardized the reporting of hierarchical variance and ICC, providing a standardized analysis and reporting example for highly nested data in this field. It offers a reference for model selection and bias control in similar studies.

(2) Theoretically, this study integrates the core viewpoints of CLT and ECT. It explains the internal mechanism of effectiveness differentiation between VE and PE from the cognitive-embodied balance perspective: VE improve abstract concept processing efficiency by reducing extraneous cognitive load, while PE promote practical skill internalization by strengthening multi-channel embodied experience. This integrated perspective compensates for the insufficient explanatory power of single theories. It provides a unified logic for analyzing effectiveness differences in digital experimental teaching, enriching the theoretical explanations for virtual-physical experiment effectiveness comparisons.

(3) Empirically, based on long-term data from 2000-2025, this study systematically sorted out three competing academic hypotheses: substitution, complementarity, and equivalence. It constructed a conditional advantage integrated explanatory framework, providing integrated evidence for gradually resolving the virtual-physical experiment effectiveness controversy. Exploratory subgroup analysis revealed effectiveness trends in three core scenarios: equivalent effectiveness in highly structured teacher-led scenarios; PE advantages in refined practical tasks; and significant VE gains in senior high school physics abstract logic teaching. This finding challenges the binary thinking of opposing advantages and disadvantages between the two. It promotes the field to shift towards differentiated adaptation and collaborative application of virtual and physical approaches within defined boundaries. It empirically supports the positioning of VE as a complementary carrier to PE, providing robust quantitative support for the digital transformation of science education.

4.1.4 Practical Implications

Based on the core findings and exploratory subgroup results, this study proposes reference optimization strategies for digital science experiment teaching from four

aspects: teaching mode, scheme design, implementation optimization, and product iteration. These provide empirical evidence for the implementation of science experiment teaching.

(1) Construct a virtual-physical collaborative teaching mode based on scenario adaptation trends. Combined with the positive effect characteristics of the senior high school physics subgroup, VE should be prioritized in teaching scenarios where PE are difficult to effectively implement. These include abstract concept explanation, high-risk experimental operations, and scenarios limited by objective conditions. Rely on their advantages of visualization, repeatability, and zero risk to break through PE teaching bottlenecks. In teaching links focusing on embodied practice such as instrument assembly, fault debugging, and phenomenon observation, retain the important teaching status of PE. Through scenario complementarity and advantage coupling of the two types of experiments, improve the comprehensive educational effectiveness of science experiment teaching.

(2) Design layered adaptive teaching schemes based on disciplinary heterogeneity trends. Based on the marginally significant moderating effect of subject type, implement targeted disciplinary optimization strategies: Biology and integrated science teaching can rely on the visualization characteristics of VE to concretely present inquiry content that is difficult to physically operate, such as microevolution, celestial motion, and long-period life activities. Chemistry teaching can adopt a closed-loop teaching mode of “virtual preview-physical operation-virtual expansion”. This balances students’ embodied practical training and cognitive processing efficiency, promoting the coordinated development of practical abilities and scientific inquiry thinking.

(3) Avoid fragmented teaching applications based on the positive results of medium-duration interventions. Based on the exploratory significant effect of 4-8 week continuous interventions, it is recommended to deeply integrate VE into unit overall teaching design. Adapt to the normal classroom teaching rhythm and avoid over-reliance on single, fragmented temporary digital teaching. Through systematic and continuous VE interventions, cultivate students’ scientific inquiry skills in the long term.

(4) Iteratively upgrade VE products based on the identified shortcomings from empirical findings. In response to the empirical finding of no significant effect in the experimental design dimension, it is recommended to carry out targeted iterations of existing VE products. Add open inquiry modules such as independent variable manipulation and personalized experimental scheme design. Support experimental error tracing and multi-level immediate feedback mechanisms. Fill the current

functional shortcomings of VE in independent inquiry and open design, better empowering the cultivation of students' high-order scientific inquiry skills.

4.2 Conclusions

This study adopted a three-level meta-analysis model that handles effect size non-independence. It integrated 21 experimental and quasi-experimental study from 2000-2025 (41 effect sizes, 3931 students) to compare the relative effects of VE and PE on students' scientific inquiry skills. It analyzed sources of heterogeneity through four moderating variables and identified the respective advantageous teaching scenarios of the two types of experiments. This provides large-sample quantitative evidence for the virtual-physical experiment effectiveness controversy. The main conclusions are as follows:

(1) VE exhibited a significant small-to-moderate positive effect on students' scientific inquiry skills compared with PE (Hedges' $g = 0.454$, $p < 0.05$). Limited by subsample size, this study did not detect significant between-group differences in scientific inquiry skills dimensions between the two types of experiments. Among them, the effect size of the experimental design dimension was close to zero (Hedges' $g = 0.052$), preliminarily indicating that PE may have potential advantages in this dimension. This trend requires further verification by large sample studies.

(2) Educational level, intervention duration, and inquiry skill dimension had no significant moderating effects ($p > 0.05$). Only subject type showed a marginally significant moderating trend ($p = 0.079$). Statistical power analysis showed that this study can completely rule out the existence of moderate-intensity moderating effects. However, the statistical power for small-intensity moderating effects was only 0.59, and the potential existence of weak moderating effects cannot be ruled out.

(3) VE intervention effects have significant scenario specificity. They have significant advantages under the conditions of senior high school educational level, physics discipline, and medium intervention duration of 4-8 weeks. In contrast, PE have potential value in cultivating experimental design skills. Since the overall moderating effect did not reach statistical significance, this study cannot precisely define the optimal application boundaries of the two types of experiments for empowering scientific inquiry skills cultivation.

In summary, through rigorous three-level meta-analysis, this study systematically evaluated the relative effectiveness of VE and PE in cultivating scientific inquiry skills. It clarified the overall effectiveness and scenario specificity of VE, providing robust quantitative evidence for the virtual-physical experiment teaching effectiveness

controversy. The nested effect size analysis process established in this study can provide methodological reference for similar meta-analyses. The proposed “cognitive-embodied dynamic balance” perspective lays a foundation for exploring the educational mechanisms of digital teaching tools and constructing a theoretical system for virtual-physical integrated teaching.

5 Limitations and Future Directions

This study adopted a three-level meta-analysis model to systematically evaluate the intervention effects of VE on students’ scientific inquiry skills and their sources of heterogeneity. It mainly has three limitations. At the sample level, only English literature was included, with samples dominated by European and American students, limiting the cross-cultural generalizability of conclusions. Subgroup sample sizes for chemistry and university educational level were small, resulting in insufficient statistical power for small-intensity moderating effects and limited subgroup inference validity. At the research scope level, virtual-physical integrated experiment literature was not included. Only binary comparisons between VE and PE were conducted, not covering mainstream virtual-physical integrated teaching paradigms. Most included studies were short-term interventions, lacking long-term follow-up data, making it impossible to evaluate the long-term sustainability of effects. No distinction was made between immersive/non-immersive VE types, preventing analysis of their heterogeneity impacts. At the methodological level, although sensitivity analysis confirmed that quasi-experimental samples did not interfere with core conclusions, inherent selection bias and unmeasured confounding variables in quasi-experiments may affect results. Core conclusions should be interpreted cautiously. Only the frequentist three-level meta-analysis framework was adopted, without cross-validation with Bayesian statistics, resulting in incomplete robustness testing. Only published literature was included, with insufficient grey literature, leading to potential publication bias.

In response to the above limitations, future research can advance from three dimensions: sample expansion, boundary extension, and method improvement. This will enrich the empirical and theoretical system of VE teaching. At the sample level, multi-lingual literature can be integrated to include more studies from different cultural backgrounds. Fill the sample gaps in weak subgroups such as chemistry and university educational level. Improve subgroup statistical power, explore effect differences in different cultural and educational scenarios, and enhance the cross-cultural

generalizability of conclusions. At the scope level, break through the binary research framework of VE and PE. Include virtual-physical integrated experimental paradigms to systematically compare the applicable scenarios and effectiveness of the three teaching modes. Introduce longitudinal follow-up designs to explore the long-term evolution laws of effects. Compare the effectiveness of immersive and non-immersive VE to enrich their heterogeneity research. At the methodological level, promote the establishment of unified operational definitions and measurement standards for scientific inquiry skills to reduce the impact of measurement tool heterogeneity. Adopt multivariate three-level meta-analysis to analyze complex interactive effects between moderating variables. Combine Bayesian methods for cross-validation to improve conclusion robustness. Expand the retrieval scope of grey literature to reduce the impact of publication bias.

Declarations

Ethical Approval (Not applicable).

Informed Consent (Not applicable).

Statement Regarding Research Involving Human Participants and/or Animals (Not applicable).

Consent to Participate (Not applicable).

Consent to Publish The author agrees to the publication of the paper in the journal.

Funding The research received no funding.

Author's Contribution Hongwei Wu: Formal analysis, Data curation, Software, Methodology, Writing-original draft, Writing-review & editing. Zhengdong Zhang: Formal analysis, Data curation, Methodology, Validation, Writing-original draft, Writing-review & editing.

Competing Interests Authors declare that they have no conflict of interest.

Availability of data and materials Data are available on request from the authors.

References

(* = references included in the meta-analysis)

Aktaş, İ., & Karamustafaoğlu, O. (2025). The effect of guided inquiry-based virtual and physical laboratories on science learning outcomes. *Research in Science & Technological Education*, 43(4), 1145-1166. <https://doi.org/10.1080/02635143.2024.2423071>

* Anam, R. S., Gumilar, S., & Handayani, M. (2025). The effects of teaching with real, virtual, and real-virtual experimentation modes on conceptual knowledge and science process skills among sixth-grade primary school students: a case study on concepts of electricity. *Education*, 135(3), 393-407. <https://doi.org/10.1080/03004279.2023.2192224>

Assink, M., & Wibbelink, C. J. (2016). Fitting three-level meta-analytic models in R: A

ep-by-step tutorial. *The quantitative methods for psychology*, 12(3), 154-174. <https://doi.org/10.20982/tqmp.12.3.p154>

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Cao, X., & Hsu, Y. (2023). Systematic review and meta-analysis of the impact of virtual experiments on students' learning effectiveness. *Interactive Learning Environments*, 31(10), 7433-7454. <https://doi.org/10.1080/10494820.2022.2072898>

Card, N. A. (2015). *Applied meta-analysis for social science research*. Guilford Publications.

Castro, R. (2025). The effects of chemistry virtual laboratories in academic achievement of secondary level learners: A meta-analysis. *Integrated Science Education Journal*, 6(1), 24-37. <https://doi.org/10.37251/isej.v6i1.1379>

Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. (2019). Cochrane handbook for systematic reviews of interventions. *Hoboken: Wiley*, 4(1002), 753-754. <https://doi.org/10.2105/AJPH.2020.305609>

Chen, G., Li, Y., Luo, R., Oubibi, M., Zhou, Y., & Gao, Y. (2026). The Effectiveness of Educational Virtual Environments in K-12 Science Education: A Meta-Analysis. *Journal of science education and technology*, 1-21. <https://doi.org/10.1007/s10956-025-10274-5>

* Chen, J. C., Huang, Y., Lin, K. Y., Chang, Y. S., Lin, H. C., Lin, C. Y., & Hsiao, H. S. (2020). Developing a hands-on activity using virtual reality to help students learn by doing. *Journal of Computer Assisted Learning*, 36(1), 46-60. <https://doi.org/10.1111/jcal.12389>

Chen, S., CHANG, W. H., LAI, C. H., & TSAI, C. Y. (2014). A comparison of students' approaches to inquiry, conceptual learning, and attitudes in simulation-based and microcomputer-based laboratories. *Science Education*, 98(5), 905-935. <https://doi.org/10.1002/sce.21126>

* Chen, S., Xue, S., Yang, D., Zhu, L., & Ye, M. (2024). Exploring differences in student learning and inquiry skills between hands-on and virtual chemistry laboratories. *Journal of Chemical Education*, 101(10), 4102-4113. <https://doi.org/10.1021/acs.jchemed.4c00557>

Chen, Y.-C., Sui, C.-J., & Chang, C.-Y. (2025). Profiling Scientific Inquiry Skills: Exploring Relationships with Scientific Competency and Background Factors in High School Students. *Journal of Baltic Science Education*, 24(5), 824-842. <https://doi.org/10.33225/jbse/25.24.824>

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child development*, 70(5), 1098-1120.

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological methods*, 19(2), 211. <https://doi.org/10.1037/a0032968>

Cheung, M. W.-L. (2015). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, 5, 1521. <https://doi.org/10.3389/fpsyg.2014.01521>

Chien, K.-P., Tsai, C.-Y., Chen, H.-L., Chang, W.-H., & Chen, S. (2015). Learning differences and eye fixation patterns in virtual and physical science laboratories. *Computers & Education*, 82, 191-201. <https://doi.org/10.1016/j.compedu.2014.11.023>

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

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, a*

nd meta-analysis. Routledge.

Darman, D. R., Suhandi, A., Kaniawati, I., Samsudin, A., & Wibowo, F. C. (2024). Development and validation of Scientific Inquiry Literacy Instrument (SILI) using rasch measurement model. *Education Sciences*, 14(3), 322. <https://doi.org/10.3390/educsci14030322>

* Dodds, H. E. (2013). *Can virtual science foster real skills? A study of inquiry skills in a virtual world*. Capella University.

* Efstathiou, C., Hovardas, T., Xenofontos, N. A., Zacharia, Z. C., deJong, T., Anjewierden, A., & van Riesen, S. A. (2018). Providing guidance in virtual lab experimentation: the case of an experiment design tool. *Educational technology research and development*, 66(3), 767-791. <https://doi.org/10.1007/s11423-018-9576-z>

* Farrokhnia, M. R., & Esmailpour, A. (2010). A study on the impact of real, virtual and comprehensive experimenting on students' conceptual understanding of DC electric circuits and their skills in undergraduate electricity laboratory. *Procedia-Social and Behavioral Sciences*, 2(2), 5474-5482. <https://doi.org/10.1016/j.sbspro.2010.03.893>

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van de n Noortgate, W. (2021). Detecting selection bias in meta-analyses with multiple outcomes: A simulation study. *The Journal of Experimental Education*, 89(1), 125-144. <https://doi.org/10.1080/00220973.2019.1582470>

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher*, 5(10), 3-8. <https://doi.org/10.3102/0013189x005010003>

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *journal of Educational Statistics*, 6(2), 107-128.

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. Academic press.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ-BRITISH MEDICAL JOURNAL*, 327(7414), 557-560. <https://doi.org/10.1136/bmj.327.7414.557>

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education*, 88(1), 28-54. <https://doi.org/10.1002/sce.10106>

Holzmeister, F., Johannesson, M., Böhm, R., Dreber, A., Huber, J., & Kirchler, M. (2023). *Heterogeneity in effect size estimates: Empirical evidence and practical implications*.

* Hu-Au, E., & Okita, S. (2021). Exploring differences in student learning and behavior between real-life and virtual reality chemistry laboratories. *Journal of science education and technology*, 30(6), 862-876. <https://doi.org/10.1007/s10956-021-09925-0>

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

* Huppert, J., Lomask, S. M., & Lazarowitz, R. (2002). Computer simulations in the high school: Students' cognitive stages, science process skills and academic achievement in microbiology. *International Journal of Science Education*, 24(8), 803-821. <https://doi.org/10.1080/09500690110049150>

* Husnaini, S. J., & Chen, S. (2019). Effects of guided inquiry virtual and physical laboratories on conceptual understanding, inquiry performance, scientific inquiry self-efficacy, and enjoyment. *Physical Review Physics Education Research*, 15(1), 010119. <https://doi.org/10.1103/PhysRevPhysEducRes.15.010119>

* Ifthinan, D. N. M., & Atun, S. (2019). The impact of inquiry-based virtual laboratory on students' inquiring abilities. *MIER Journal of Educational Studies Trends and Practices*, 50-61. <https://doi.org/10.1366/000370278774330856>

Jia, X., & Zeng, M. PISA Science Literacy Assessment Frameworks, Characteristics and Implications (2006-2025).

Jia, X., & Zeng, M. (2025). PISA Science Literacy Assessment Frameworks, Characteristics and Implications (2006-2025). 150-157.

* Juwariyah, S., & Latifah, E. (2017). Guided inquiry method employing virtual laboratory to improve scientific working skills. *Jurnal Pendidikan Sains Universitas Negeri Malang*, 5(1), 17-25. <https://doi.org/10.17977/jps.v5i1.9016>

* Kapici, H., & Akcay, H. (2020). Enhancing pre-service science teachers' inquiry skills in hands-on and virtual laboratory environments. *Themes in e-Learning*, 13.

Kali, Y. (2021). Guiding frameworks for the design of inquiry learning environments. In *International handbook of inquiry and learning* (pp. 39-59). Routledge.

* Kapici, H. O. (2025). The Impact of Digital Inquiry vs. Classroom-Based Science Instruction on Learners' Outcomes. *Journal of science education and technology*, 1-20. <https://doi.org/10.1007/s10956-025-10246-9>

* Kapici, H. O., Akcay, H., & Cakir, H. (2022). Investigating the effects of different levels of guidance in inquiry-based hands-on and virtual science laboratories. *International Journal of Science Education*, 44(2), 324-345. <https://doi.org/10.1080/09500693.2022.2028926>

* Kapici, H. O., Akcay, H., & de Jong, T. (2019). Using hands-on and virtual laboratories alone or together—which works better for acquiring knowledge and skills? *Journal of science education and technology*, 28(3), 231-250. <https://doi.org/10.1007/s10956-018-9762-0>

Klahr, D., Triona, L. M., & Williams, C. (2007). Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *Journal of research in science teaching*, 44(1), 183-203. <https://doi.org/10.1002/tea.20152>

Lai, M. H., & Kwok, O.-m. (2016). Estimating standardized effect sizes for two-and three-level partially nested data. *Multivariate Behavioral Research*, 51(6), 740-756. <https://doi.org/10.1080/00273171.2016.1231606>

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE publications, Inc.

Liu, J. (2015). *Statistical power in meta-analysis* University of South Carolina].

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2010). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International journal of surgery*, 8(5), 336-341. <https://doi.org/10.1016/j.ijsu.2010.02.007>

Muiliwijk, S. E., & Lazonder, A. W. (2023). Learning from physical and virtual investigation: A meta-analysis of conceptual knowledge acquisition. *Frontiers in Education*,

Nie, X., Tian, Y., Liu, M., Wu, D., & Guo, Y. (2025). The impact of generative artificial intelligence on students' higher order thinking: Evidence from a three-level meta-analysis. *Education and Information Technologies*, 30(17), 25359-25390. <https://doi.org/10.1007/s10639-025-13735-x>

* Ogbuanya, T. C., & Onele, N. O. (2018). Investigating the effectiveness of desktop virtual reality for teaching and learning of electrical/electronics technology in universities. *Computers in the Schools*, 35(3), 226-248. <https://doi.org/10.1080/07380569.2018.1492283>

* Oser, R., & Fraser, B. J. (2015). Effectiveness of virtual laboratories in terms of learning environment, attitudes and achievement among high-school genetics students. *Curriculum and Teaching*, 30(2), 65-80. <https://doi.org/10.7459/ct/30.2.05>

Ouwehand, K., Lespiau, F., Tricot, A., & Paas, F. (2025). Cognitive load theory: emerging trends and innovations. *Education Sciences*, 15(4), 458. <https://doi.org/10.3390/educsci15040458>

Ozdemir, M., Sahin, C., Arcagok, S., & Demir, M. K. (2018). The effect of augmented reality applications in the learning process: A meta-analysis study. *Eurasian Journal of Educational Research*, 18(74), 165-186. <https://doi.org/10.14689/ejer.2018.74.9>

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . Brennan, S. E. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Papalazarou, N., Lefkos, I., & Fachantidis, N. (2024). The effect of physical and virtual inquiry-based experiments on students' attitudes and learning. *Journal of science education and technology*, 33(3), 349-364. <https://doi.org/10.1007/s10956-023-10088-3>

Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A., Kamp, E. T., . . . Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, 14, 47-61. <https://doi.org/10.1016/j.edurev.2015.02.003>

Piaget, J. (1976). Piaget's theory. In *Piaget and his school: A reader in developmental psychology* (pp. 11-23). Springer.

Puntambekar, S., Gnesdilow, D., Dornfeld Tissenbaum, C., Narayanan, N. H., & Rebello, N. S. (2021). Supporting middle school students' science talk: A comparison of physical and virtual labs. *Journal of research in science teaching*, 58(3), 392-419. <https://doi.org/10.1002/tea.21664>

* Pyatt, K., & Sims, R. (2012). Virtual and physical experimentation in inquiry-based science labs: Attitudes, performance and access. *Journal of science education and technology*, 21(1), 133-147. <https://doi.org/10.1007/s10956-011-9291-6>

Rau, M. A. (2020). Comparing multiple theories about learning with physical and virtual representations: Conflicting or complementary effects? *Educational Psychology Review*, 32(2), 297-325. <https://doi.org/10.1007/s10648-020-09517-1>

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3), 638. <https://doi.org/10.1037/0033-2909.86.3.638>

Seeratan, K. L., McElhaney, K. W., Mislevy, J., McGhee Jr, R., Conger, D., & Long, M. C. (2020). Measuring students' ability to engage in scientific inquiry: A new instrument to assess data analysis, explanation, and argumentation. *Educational Assessment*, 25(2), 112-135. <https://doi.org/10.1080/10627197.2020.1756253>

Song, Q., Li, M., Zhang, Z., & Zhou, Q. (2025). Learning effectiveness of virtual technologies in chemistry using a three-level meta-analysis. *Education and Information Technologies*, 30(11), 14737-14759. <https://doi.org/10.1007/s10639-025-13321-1>

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295-312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)

Tatli, Z., & Ayas, A. (2013). Effect of a virtual chemistry laboratory on students' achievement. *Journal of Educational Technology & Society*, 16(1), 159-170. <https://doi.org/10.2307/jedutechsoci.16.1.159>

* Toth, E. E., Ludvico, L. R., & Morrow, B. L. (2014). Blended inquiry with hands-on a

nd virtual laboratories: The role of perceptual features during knowledge construction. *Interactive Learning Environments*, 22(5), 614-630. <https://doi.org/10.1080/10494820.2012.693102>

* Triona, L. M., & Klahr, D. (2003). Point and click or grab and heft: Comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition and Instruction*, 21(2), 149-173. https://doi.org/10.1207/S1532690XCI2102_02

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software*, 36, 1-48.

Wang, J., Guo, D., & Jou, M. (2015). A study on the effects of model-based inquiry pedagogy on students' inquiry skills in a virtual physics lab. *Computers in Human Behavior*, 49, 658-669. <https://doi.org/10.1016/j.chb.2015.01.043>

Wang, X.-M., Yu, X.-H., Yu, D.-d., Hwang, G.-J., & Lan, M. (2025). Does combining real and virtual experiments improve learning achievement in physics? Evidence from a meta-analysis (2001-2021). *Educational Research Review*, 46, 100661. <https://doi.org/10.1016/j.edurev.2024.100661>

Wang, Y., Jiang, X., Hu, J., Hou, C., Li, B., & Wang, Q. (2025). A meta-analysis of the impact of virtual experiments on students' development. *Interactive Learning Environments*, 1-22. <https://doi.org/10.1080/10494820.2025.2536574>

Wells, G. A., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M., & Tugwell, P. (2000). The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses.

Wen, P., Lu, F., & Mohamad Ali, A. Z. (2024). Using attentional guidance methods in virtual reality laboratories reduces students' cognitive load and improves their academic performance. *Virtual Reality*, 28(2), 110. <https://doi.org/10.1007/s10055-024-01012-0>

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4), 625-636. <https://doi.org/10.3758/BF03196322>

Wörner, S., Kuhn, J., & Scheiter, K. (2022). The best of two worlds: A systematic review on combining real and virtual experiments in science education. *Review of Educational Research*, 92(6), 911-952. <https://doi.org/10.3102/00346543221079417>

Wu, T. T., Sarwono, E., & Huang, Y. M. (2025). Incorporating Computational Thinking In to Virtual Laboratories to Enhance Learning Motivation, Engagement, and Higher-Order Thinking Skills. *Journal of Computer Assisted Learning*, 41(2), e70017. <https://doi.org/10.1111/jcal.70017>

Yang, C., Zhang, J., Hu, Y., Yang, X., Chen, M., Shan, M., & Li, L. (2024). The impact of virtual reality on practical skills for students in science and engineering education: A meta-analysis. *International Journal of STEM Education*, 11(1), 1-15. <https://doi.org/10.1186/s40594-024-00487-2>

* Yang, K.-Y., & Heh, J.-S. (2007). The impact of internet virtual physics laboratory instruction on the achievement in physics, science process skills and computer attitudes of 10th-grade students. *Journal of science education and technology*, 16(5), 451-461. <https://doi.org/10.1007/s10956-007-9062-6>

Zacharia, Z. C., & De Jong, T. (2014). The effects on students' conceptual understanding of electric circuits of introducing virtual manipulatives within a physical manipulatives-oriented curriculum. *Cognition and Instruction*, 32(2), 101-158. <https://doi.org/10.1080/07370008.2014.887>

Zou, L., Zhang, Z., Mavilidi, M., Chen, Y., Herold, F., Ouwehand, K., & Paas, F. (2025). The synergy of embodied cognition and cognitive load theory for optimized learning. *Nature Human Behaviour*, 9(5), 877-885. <https://doi.org/10.1038/s41562-025-02152-2>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PRISMA2020ReportingChecklist.docx](#)