Id: 1867423

Participatory Educational Research (PER)
Vol.13(4), pp. 195-213, July 2026
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.26.55.13.4

Effectiveness of Immersive Technologies in Science Education: A Meta-Analysis of Virtual, Augmented, and Mixed Reality

Ayşe Gül ÖZAŞKIN-ARSLAN*
Primary Education, Afyon Kocatepe University, Afyonkarahisar, Türkiye
ORCID: 0000-0002-9018-5525

Şafak ULUÇINAR-SAĞIR
Primary Education, Amasya University, Amasya, Türkiye
ORCID: 0000-0003-3383-5330

Article history

Received:
19.01.2026

Received in revised form:
01.03.2026

Accepted:
12.04.2026

Key words:
virtual reality (VR); augmented reality (AR); science education; meta-analysis; immersive learning environments

This study evaluates the effectiveness of Extended Reality (XR) technologies, including Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), on student achievement in science education. Despite the increasing integration of immersive tools in classrooms, a comprehensive comparative synthesis remains limited. To address this gap, three parallel meta-analyses were conducted on 218 unique studies, yielding 225 independent effect sizes, published between January 2000 and June 2021. Following PRISMA guidelines, a systematic search was performed across major electronic databases and academic repositories. Statistical analysis using random-effects models revealed that all three technologies significantly enhanced student academic achievement, with AR and VR showing more pronounced impacts compared to the emerging domain of Mixed Reality. However, high levels of heterogeneity across the results indicate that the success of these tools appears to be highly context-dependent. Moderator analyses suggest that effectiveness varies significantly according to subject discipline, educational level, and specific instructional design features such as simulation types and gamification. These findings were interpreted through the lenses of Cognitive Load Theory, the Gartner Hype Cycle, and Embodied Learning. The results provide robust evidence for the transformative potential of XR in science instruction while highlighting the necessity of sound pedagogical and contextual planning for successful classroom implementation.

Introduction

The 21st-century classroom is undergoing a significant transformation, moving beyond static textbooks and passive lectures into dynamic, interactive learning environments. This transformation reflects a century-long pursuit of meaningful science instruction. As early as 1902, Armstrong's 'heuristic approach' emphasized discovery-oriented science learning, arguing that students should construct knowledge through active investigation rather than mere information transfer (Armstrong, 1902). This view aligns with the idea that science is

* Correspondency: aysegulozaskn@gmail.com

fundamentally investigative, requiring students to engage with scientific processes and theoretical explanations simultaneously (McFarlane & Sakellariou, 2002).

This shift also mirrors broader constructivist paradigms, where learning is understood as an active process of meaning-making rather than passive reception. However, integrating new media into education raises a long-standing theoretical debate. Clark (1985) argued that media are merely delivery vehicles that do not influence learning, whereas Kozma (1994) contended that the attributes of a medium can interact with cognitive processes to facilitate learning. XR, a suite of immersive technologies including VR, AR, and MR, offers a particularly useful test case for this debate because its affordances are explicitly designed to influence the cognitive and affective dimensions of learning (Christensen et al., 2018; Lainema et al., 2019). VR immerses learners in fully digital environments, AR overlays digital information onto the physical world (Azuma, 1997), and MR enables virtual objects to interact with real environments, creating hybrid spaces for learning (Milgram & Kishino, 1994). These tools can be positioned along the reality–virtuality continuum, where the physical world lies at one end and fully virtual environments at the other.

The potential of XR may be particularly significant in science education, where learning is often constrained by the challenge of representing abstract concepts, invisible forces, and complex phenomena. As Driver et al. (1994) noted, many foundational scientific concepts—such as atoms, genes, or gravitational fields—cannot be directly discovered through simple observation and must instead be constructed through carefully designed learning experiences. Although traditional laboratory activities are essential, they often face constraints such as high costs, safety concerns, time limitations, and limited opportunities for repeated experimentation (Zacharia, 2007; Schrum & Levin, 2009). XR technologies may address these barriers by offering safe, cost-effective, and repeatable environments where learners can manipulate variables and observe causal relationships in real time. In this way, XR applications can bridge symbolic scientific representations with experiential understanding by enabling visualization and manipulation of abstract phenomena (Hennessy et al., 2007), supporting conceptual change (Shelton & Stevens, 2004), linking scientific ideas to authentic contexts (Klopfer & Squire, 2008), and providing interactive environments that may facilitate personalized learning (Dede, 1998).

In parallel with growing adoption, literature has expanded rapidly, as reflected by a surge in systematic reviews exploring XR from multiple perspectives. Recent syntheses have examined usability and learning experience (Ramli et al., 2024), pedagogical design features (Gil Parga et al., 2024), and multimodal interaction in VR environments (Hu et al., 2025). Global market trends further signal the increasing relevance of XR, with investments exceeding $25 billion in 2022 across education, health, and engineering (Statistica, 2022). Despite this growing body of evidence, a notable gap persists. Existing reviews typically focus on a single technology (most often AR) or a specific feature such as interaction modalities in VR (e.g., Gil Parga et al., 2024; Ramli et al., 2024; Hu et al., 2025), whereas broader XR reviews often remain largely descriptive or qualitative (e.g., Fernández-Cerero et al., 2025). As a result, a comprehensive meta-analysis that quantitatively synthesizes and directly compares the effectiveness of VR, AR, and the largely underexplored domain of MR on student achievement specifically within science education remains limited.

To address this gap, the present study had a twofold objective. First, it aimed to synthesize existing research to determine the overall effectiveness of VR, AR, and MR on student achievement in science. Second, it sought to systematically explore the influence of key

moderator variables to identify the conditions under which these applications are most effective. Accordingly, this meta-analysis addressed the following research questions:

(1) What is the overall effect of VR applications on student achievement in science education?
(2) What is the overall effect of AR applications on student achievement in science education?
(3) What is the overall effect of MR applications on student achievement in science education?
(4) To what extent do study characteristics (such as educational level, science discipline, application type, and instructional design features) moderate the effectiveness of VR, AR, and MR applications?

Beyond addressing a clear synthesis gap, the present meta-analysis also adopts an explicit comparative logic across the XR spectrum. Although VR, AR, and MR are often discussed under the same umbrella term, they differ substantially in terms of immersion level, interaction affordances, and the nature of the learning experience they generate. These differences suggest that combining all XR modalities into a single pooled analysis may obscure technology-specific patterns and lead to overgeneralized conclusions. Therefore, the study conducts three parallel meta-analytic syntheses—one for VR, one for AR, and one for MR—using consistent inclusion criteria, coding procedures, and statistical models. This structure enables both (a) robust estimation of the overall achievement effects of each technology within science education and (b) systematic exploration of moderators to clarify the instructional and contextual conditions under which each modality is most effective. In doing so, the study aims to provide evidence-informed guidance for researchers, curriculum designers, and educators seeking to make pedagogically meaningful decisions about when and how immersive technologies can best support science learning outcomes. Overall, the study contributes a technology-sensitive, context-aware quantitative synthesis that advances beyond single-technology reviews by directly comparing VR, AR, and MR within a unified analytic framework focused exclusively on science education.

Method

This study employed a systematic review and meta-analytic approach to synthesize quantitative evidence on the effectiveness of immersive technologies (VR, AR, and MR) in science education. To ensure transparency and reproducibility, all procedures were conducted in line with the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The overall synthesis was structured according to Cooper's (2017) seven-step protocol for research synthesis, including problem formulation, literature search, study evaluation, data extraction and coding, statistical integration, interpretation, and reporting.

Search strategy and study selection

A comprehensive literature search was performed to identify relevant peer-reviewed journal articles, as well as dissertations published between January 2000 and June 2021. Searches were conducted across major electronic databases, including ISI Web of Knowledge, Scopus, ERIC, ScienceDirect, EBSCO, ProQuest Digital Dissertations, and Google Scholar. To capture research conducted in Türkiye and minimize language-related publication bias, national academic repositories were also searched, including YÖK Tez Merkezi and ULAKBİM.

The time frame was selected for two reasons. First, XR-related technologies began to appear more systematically in empirical science education research after 2000, alongside the wider availability of computer-based simulations, interactive multimedia, and early immersive platforms. Second, limiting the search to this period allowed the synthesis to capture the contemporary development trajectory of immersive learning environments while maintaining methodological comparability across studies in terms of research design, achievement measurement, and reporting practices.

The search strategy relied on Boolean operators and keyword combinations aligned with both (a) the technology type and (b) achievement-related learning outcomes. For example, the primary English search string used for Virtual Reality was:

(("virtual reality" OR "VR") AND (achievement OR success OR learning OR performance OR gain) AND ("experimental design" OR "experimental study"))

Equivalent strings were developed for AR and MR. In addition, Turkish adaptations of all search strings were applied to broaden geographical representation and to identify potentially eligible studies not indexed in international databases.

Given that the aim of this study was to compare effect patterns across the XR spectrum, the screening process was conducted through three parallel and independent selection streams, corresponding to VR, AR, and MR. Each stream followed PRISMA 2020 (Page et al., 2021) stages of identification, screening, eligibility assessment, and inclusion. In the identification stage, a total of 14,908 records were retrieved for VR, 8,094 for AR, and 384 for MR. After duplicate removal and title/abstract screening, full texts were assessed for eligibility using the predefined inclusion and exclusion criteria described below.

The PRISMA 2020 flow diagrams provide a detailed overview of the identification, screening, eligibility assessment, and inclusion stages for each technology: Figure 1 (VR), Figure 2 (AR), and Figure 3 (MR). Following this process, the final dataset consisted of 218 unique publications providing 225 independent effect sizes. All included studies provided sufficient statistical information for effect size calculation and met the predefined eligibility criteria, including the use of an experimental or quasi-experimental design with a control or comparison group. Some publications contributed more than one independent effect size only when they involved distinct samples, separate intervention conditions, or clearly independent comparisons. To ensure statistical validity, special attention was given to the independence of effect sizes. Each effect size represented a unique comparison between an experimental group using an XR intervention and a control/comparison group receiving traditional instruction or a non-immersive alternative. A single publication was allowed to contribute more than one effect size only when the effect sizes were based on distinct samples, separate intervention conditions, or clearly independent comparisons. When multiple achievement outcomes, repeated measurements, or multiple statistical indices were reported for the same sample and intervention condition, only one effect size was retained, or the available effects were combined to avoid overweighting that sample. In such cases, selection decisions were guided by the outcome most directly aligned with academic achievement in science education. This procedure preserved the independence of effect sizes while allowing multiple contributions from the same publication only when they represented independent comparisons.

Inclusion and exclusion criteria

To enhance the quality and comparability of the synthesized data, studies were screened against a predefined set of inclusion criteria:

- Publication language: Only studies published in English or Turkish were included.
- Research design: The study employed an experimental or quasi-experimental design with a formal control or comparison group, enabling the calculation of standardized effect sizes.
- Intervention type (independent variable): The intervention involved an immersive technology application explicitly categorized as VR, AR, or MR and used for instructional purposes.
- Outcome measure (dependent variable): The study reported academic achievement outcomes in science education contexts (e.g., physics, chemistry, biology) measured through a structured instrument.
- Target population: Participants were students enrolled in K–12 education (primary, middle, or high school) or higher education.
- Data reporting: The study provided sufficient quantitative information to compute effect sizes, including means, standard deviations, and sample sizes, or equivalent inferential statistics such as t or F values.

Although the inclusion of Turkish-language studies was intended to broaden geographical representation and reduce language-related bias, the restriction to English and Turkish publications may still have introduced some degree of language bias.

Studies were excluded if they met one or more of the following conditions:

- The study was purely qualitative or descriptive and did not include a control/comparison group.
- The study focused on professional training or adult workforce development outside formal education settings.
- The study lacked sufficient statistical information to calculate an effect size.
- The study was published as a conference proceeding.

The last exclusion criterion was applied only to the VR and AR analyses because the initial searches in these two streams retrieved a large number of conference papers with limited methodological and statistical reporting. These records were excluded to improve consistency in reporting standards and data completeness within the VR and AR datasets. No equivalent exclusion was applied to the MR stream because of the smaller and still emerging evidence base for MR.

Data coding and reliability

A detailed coding framework was developed to extract and classify information from each eligible study. Coding categories were organized into three primary dimensions:

- Publication details (e.g., year, country, publication type)
- Methodological characteristics (e.g., design type, sample size, educational level), and
- Instructional design characteristics (e.g., XR technology type, science discipline, simulation/application type, learning environment features).

To reduce coder bias and assess coding reliability, a second researcher independently coded a subset of studies using a random number generator (N = 60, approximately 28% of the final sample). Inter-coder agreement was calculated using the Miles and Huberman (1994) reliability formula. Initial agreement was 87%, and after a second review round focused primarily on statistical data extraction and category clarifications, agreement increased to 98%. All disagreements were resolved through discussion until full consensus (100%) was reached, supporting the consistency of the coding process.

Statistical analysis and model selection

The primary effect size metric for this meta-analysis was Cohen's d (standardized mean difference). Effect sizes were calculated directly from descriptive statistics or estimated from inferential statistics (e.g., t or F values). Effect sizes were interpreted using Cohen's (1988) benchmarks: 0.2 (small), 0.5 (medium), and 0.8 (large).

All meta-analytic calculations were conducted using Comprehensive Meta-Analysis (CMA) software. A random-effects model was applied across all analyses, as educational technology effects may vary depending on implementation quality, learning context, instructional integration, and participant characteristics (Field & Gillett, 2010; Hedges & Vevea, 1998). Between-study heterogeneity was assessed using Cochran's Q and the I^2 statistic.

To evaluate potential publication bias, multiple complementary procedures were used. These included Rosenthal's and Orwin's Fail-safe N calculations as well as visual inspection of funnel plot symmetry. These procedures provided complementary statistical and graphical evidence regarding potential publication bias and the possible impact of missing studies.

Moderator selection

Because heterogeneity is common in educational technology research, a comprehensive set of potential moderator variables was identified to explore sources of variance. The selection was guided by the framework proposed by Merchant et al. (2014) and supplemented by variables identified in recent meta-analyses (Villena-Taranilla et al., 2022; Wu et al., 2020). These included subject discipline (e.g., physics, chemistry), educational level (K–12, higher education), instructional mode (presentation, application, independent use), and hardware type (2D display vs. wearable tech). Subgroup analyses were performed only for categories containing a sufficient number of studies (typically $k \geq 3$) to maintain statistical power and interpretability.

Results

This section presents the findings from the three separate meta-analyses conducted on VR, AR, and MR. For each technology, results were presented for the study selection process, overall mean effect size, heterogeneity, publication bias, and moderator analyses.

Results for VR

Study selection and characteristics

Database searches initially yielded 14,908 records. The screening process, based on predefined inclusion and exclusion criteria, is detailed in the PRISMA 2020 flow diagram (Figure 1). After removing duplicates, screening titles and abstracts, and assessing full texts, a

total of 98 studies (k = 98) that quantitatively examined the effect of VR use in science education on academic achievement and provided suitable data for meta-analysis were included in the meta-analysis.

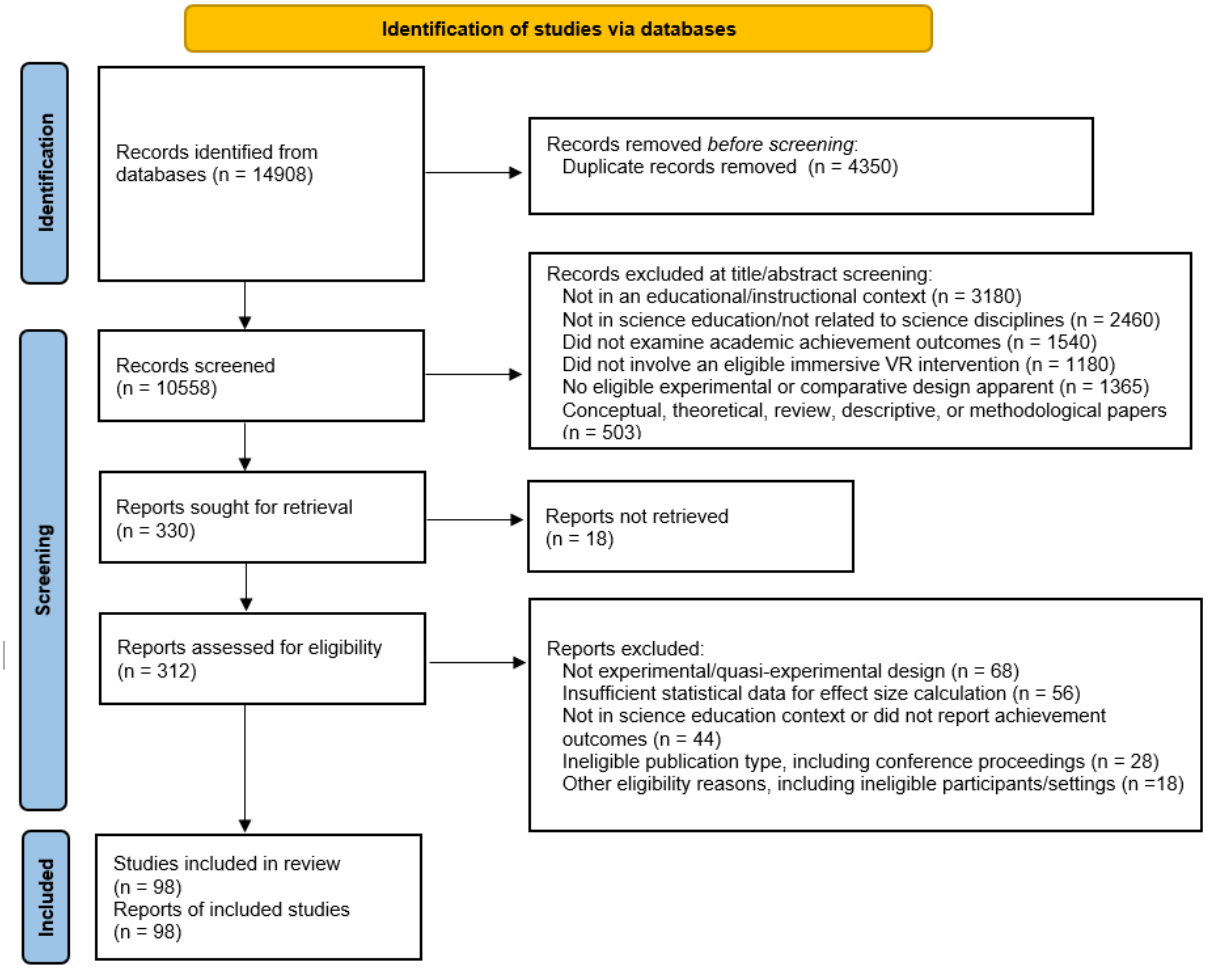

Figure 1. PRISMA 2020 flow diagram for VR studies included in meta-analysis

Overall effect size and heterogeneity

The synthesis of 98 studies revealed a statistically significant and positive overall effect of VR applications on science achievement, as detailed in Table 1. Under the random-effects model, the weighted mean effect size was Cohen's d = 0.531 (95% CI [0.403, 0.658], p < .001). According to Cohen's (1988) conventions, this value represents a medium-sized effect, suggesting that, on average, students receiving VR-based instruction demonstrated science achievement about half a standard deviation higher than students in control or comparison groups. The individual effect sizes for all 98 studies are visualized in the forest plot presented in Supplementary File 1.

Table 1. *Overall Effect Size and Heterogeneity for VR Studies*

Model	**k**	**Cohen's *d***	**95% CI**	**Q**	***I²* (%)**
Random-Effects	98	0.531*	[0.403, 0.658]	853.57	88.6

*Note. *p < .001. k = number of effect sizes; d = mean effect size (Cohen's d); CI = confidence interval; Q = Cochran's Q test for heterogeneity; I² = percentage of variability due to heterogeneity.*

The analysis also revealed a high degree of heterogeneity among the effect sizes across studies (Q(97) = 853.57, p < .001). The I^2 statistic of 88.6% indicates that the vast majority of the observed variance is due to true differences between studies rather than sampling error. This substantial heterogeneity suggests that the effectiveness of VR varies across different contexts and justifies the subsequent moderator analyses to explore the sources of this variation.

Publication bias analyses

To assess the potential for publication bias, a multifaceted approach was employed. First, a visual inspection of the funnel plot (Supplementary Material - Figure S1) was conducted. The plot suggested a generally symmetrical distribution of studies around the mean effect size, with most studies clustering at the top where precision is highest. This symmetry provides an initial indication of a low risk of publication bias.

To supplement this visual analysis with quantitative metrics, Rosenthal's (1979) Fail-safe N was calculated. The result (N = 1184) indicated that 1,184 undiscovered or unpublished studies with a null effect would be required to render the overall effect size statistically non-significant. This number far exceeds the tolerance criterion of 5k + 10 (5 × 98 + 10 = 500), suggesting that the findings are relatively stable against potential publication bias. Furthermore, Orwin's (1983) Fail-safe N test revealed that 9,505 studies with a null effect would be needed to reduce the effect size to a trivial value (e.g., d = 0.001), further supporting the stability of the result. Collectively, these analyses suggest that publication bias is unlikely to be a significant threat to the validity of the findings for the VR meta-analysis.

Moderator analysis

Given the substantial heterogeneity observed (I^2 = 88.6%), we conducted moderator analyses to explore potential sources of this variance. We examined 15 study characteristics as potential moderators. The analyses revealed that four of these variables significantly explained the differences in effect sizes across studies. The results of these significant moderators are summarized in Table 2.

Table 2. *Summary of significant moderator analyses for VR*

Moderator Variable	***k***	***Q_B***	***df***	***p***
Subject Area	94	7.21	2	**.027**
Type of VR Application	98	20.66	3	**.000**
Control Group Method/Tool	97	80.21	4	**.000**
Learning Environment	96	36.38	1	**.000**

Note. k = number of effect sizes included in the analysis. Q_B = between-groups heterogeneity statistic. df = degrees of freedom. Full details are in the Supplementary Material, Table S1.

Several key patterns emerged from these findings. First, the effectiveness of VR varied significantly by *subject area*. The largest effect was observed in Physics (d = 0.758), while the effects in Biology (d = 0.340) and Chemistry (d = 0.394) were smaller.

Second, the *type of VR* was a strong moderator. Simulations (d = 0.634) and virtual or educational games (d = 0.550) yielded significantly larger effects than virtual worlds (d = 0.132) and representation-based applications like 360-degree videos (d = 0.196).

Third, the nature of the *control group* significantly influenced the relative effect of VR. The largest advantages were seen when VR was compared to traditional instruction methods (d = 0.682) and physical laboratory activities (d = 0.607). By contrast, when the control group used concrete 3D models, the effect of VR was negative (d = -0.798), indicating that the relative advantage of VR may depend on the comparison condition.

Finally, the *learning environment* was a significant factor. VR applications implemented within a classroom setting showed a moderate positive effect (d = 0.562), whereas those conducted in remote learning contexts had a negligible, slightly negative effect (d = -0.001).

Conversely, several variables were not found to be significant moderators of the effectiveness of VR ($p > .05$). These included publication type, study design, educational level, specific middle school grade, data collection instrument, task implementation (individual or collaborative), task type, feedback provision, learner control, instructional mode, and immersion hardware (2D display vs. wearable tech). The detailed statistical results for all 15 moderator analyses, including the non-significant findings, are presented in the Supplementary Material (Table S1).

Results for AR

Study selection and characteristics

The systematic search for AR studies identified an initial 8,094 records. Following the screening protocol detailed in the PRISMA 2020 flow diagram (Figure 2), 94 studies, which reported 100 distinct effect sizes, met the inclusion criteria and were included in the final meta-analysis.

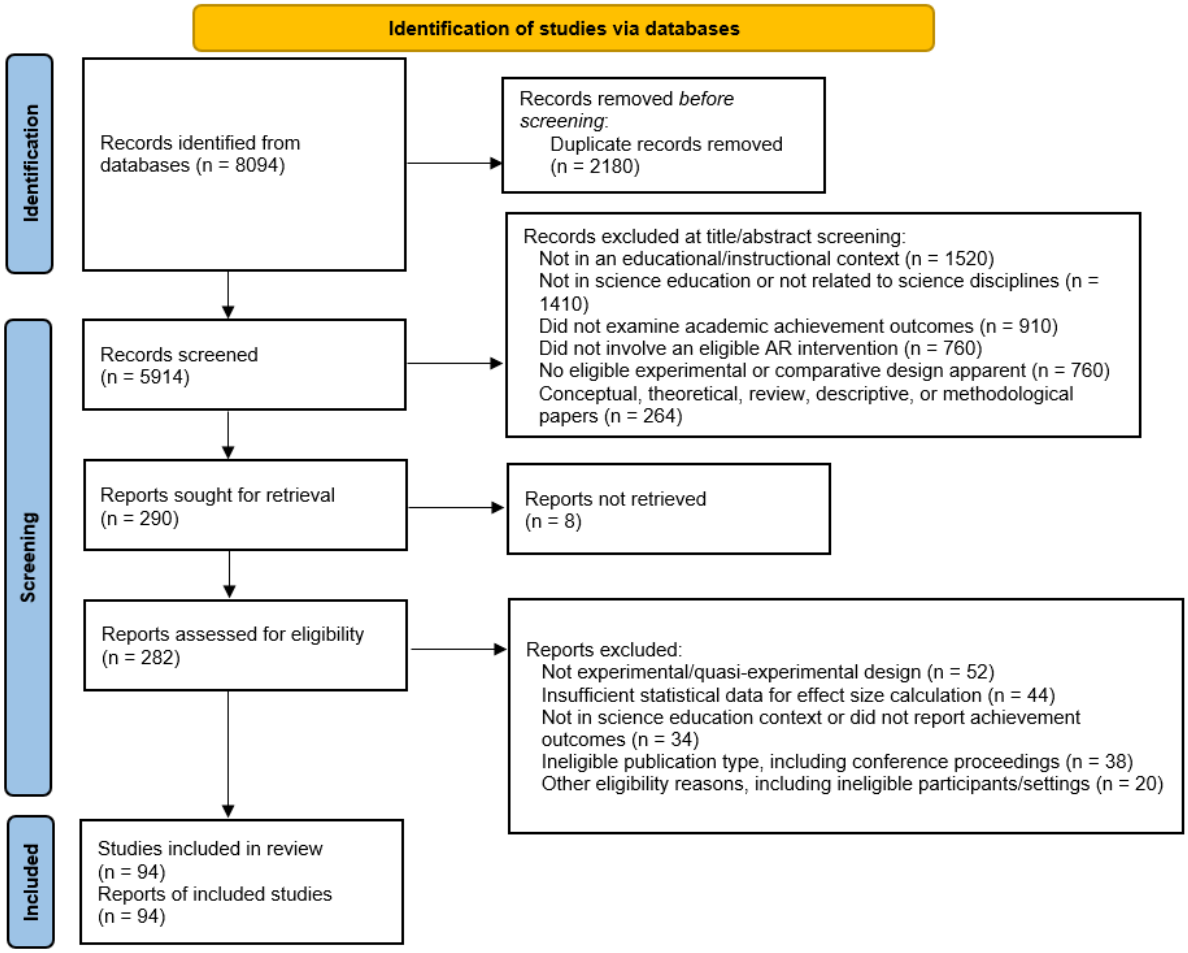

Figure 2. PRISMA 2020 flow diagram for the selection of AR studies

Overall effect size and heterogeneity

The synthesis of 100 effect sizes demonstrated that AR had a statistically significant and positive effect on science achievement. As detailed in Table 3, the random-effects model yielded a weighted mean effect size of Cohen's d = 0.641, representing a medium-to-large effect. The analysis also identified substantial heterogeneity across the studies (I^2 = 80.5%). This indicates that the effectiveness of AR varied across contexts and supports the need for moderator analyses. The individual effect sizes were visualized in the forest plot in Supplementary File 2.

Table 3. *Overall effect size and heterogeneity for AR studies*

Model	**k**	**Cohen's *d***	**95% CI**	**Q**	***I²* (%)**
Random-Effects	100	0.641*	[0.526, 0.756]	508.87	80.5

*Note. *p < .001. k = number of effect sizes; d = mean effect size (Cohen's d); CI = confidence interval; Q = Cochran's Q test for heterogeneity; I^2 = percentage of variability due to heterogeneity.*

Publication bias analyses

The potential for publication bias was assessed through multiple methods. The funnel plot (Supplementary Material - Figure S2) appeared largely symmetrical. This visual assessment was supported by quantitative tests, including Rosenthal's (1979) Fail-safe N (N =

4813) and Orwin's (1983) Fail-safe N (N = 7783). Furthermore, Begg and Mazumdar's (1994) rank correlation test was non-significant ($\tau = 0.095$, $p = .158$). These combined results suggest a low risk of publication bias for the AR meta-analysis.

Moderator analysis

Given the substantial heterogeneity ($I^2 = 80.5\%$), we analyzed 13 potential study characteristics as moderators. As summarized in Table 4, four variables significantly moderated the effect of AR on student achievement.

Table 4. *Summary of significant moderator analyses for AR*

Moderator Variable	***k***	***Q_B***	***df***	***p***
School Level	94	19.57	3	**.000**
Subject Area	96	7.28	2	**.026**
Control Group Method	95	6.64	2	**.036**
Data Collection Tool	98	7.16	1	**.007**

Note. k = number of effect sizes included in the analysis. Q_B = between-groups heterogeneity statistic. df = degrees of freedom. Full details are in the Supplementary Material, Table S2.

Several key patterns emerged from these findings. First, the effectiveness of AR was significantly moderated by school level, with the largest effect observed in middle school (d = 0.891). Second, the subject area mattered; AR was significantly more effective in Chemistry (d = 0.774) and Physics (d = 0.745) than in Biology (d = 0.442). Third, the control group method was a significant factor, with AR showing a larger advantage relative to traditional methods (d = 0.701). Finally, the data collection tool was a significant moderator; studies using researcher-developed tests (d = 0.706) reported higher effect sizes than those using standardized tests (d = 0.317).

Variables such as publication type, study design, and AR feature did not significantly moderate the effect of AR. The complete statistical results for all 13 moderator analyses, including non-significant findings, are presented in the Supplementary Material (Table S2).

Results for MR

Study selection and characteristics

The initial literature search for MR yielded 384 records. After the PRISMA screening process (Figure 3), a total of 26 studies, reporting 27 distinct effect sizes, were included in the final meta-analysis.

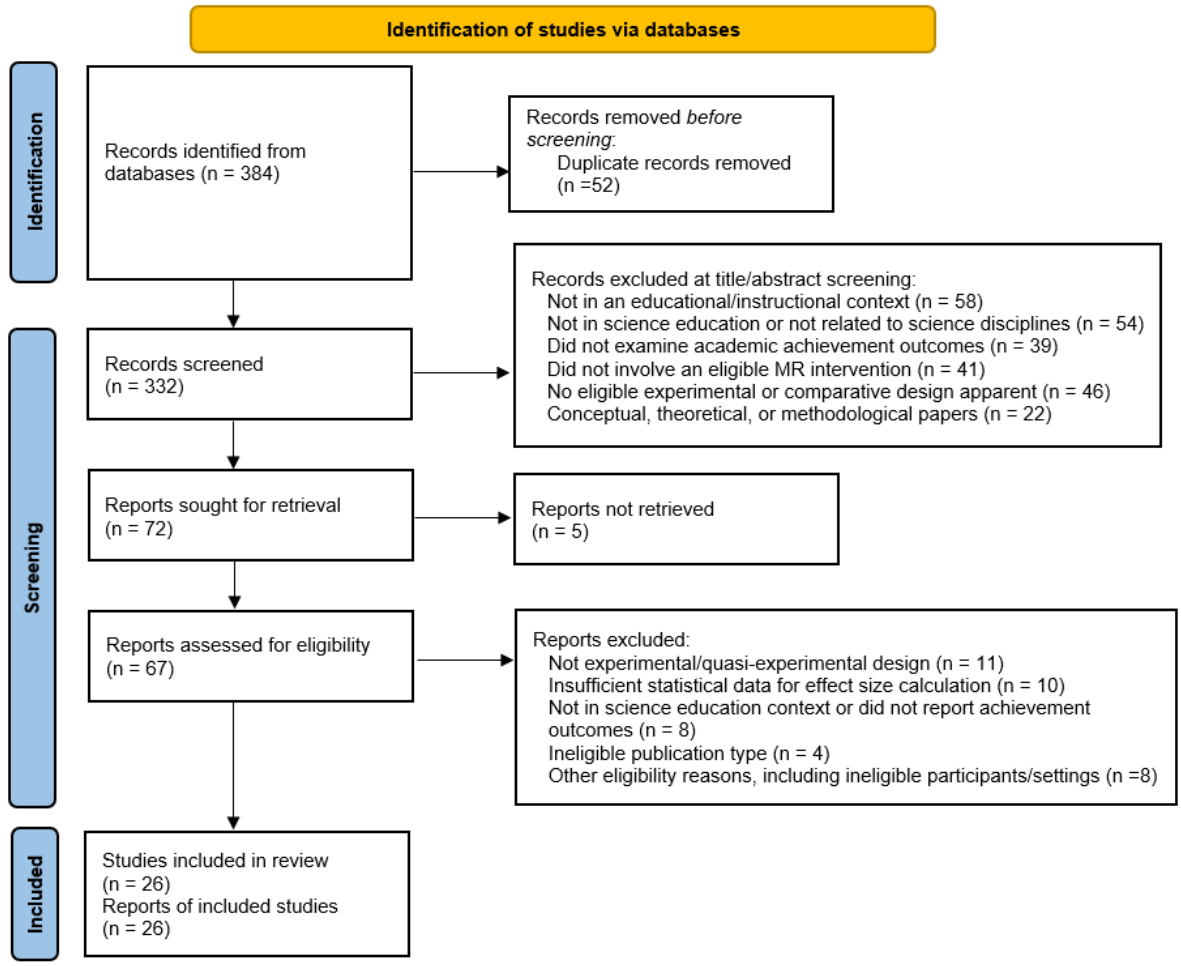

Figure 3. PRISMA 2020 flow diagram for the selection of MR studies

Overall effect size and heterogeneity

The synthesis of 27 effect sizes revealed that MR had a statistically significant but small positive effect on science achievement. As shown in Table 5, the random-effects model yielded a weighted mean effect size of Cohen's d = 0.299. The analysis also identified substantial heterogeneity across the studies (I^2 = 81.0%), supporting the need for moderator analysis. The forest plot showing individual study effects is available in Supplementary File 3.

Table 5. *Overall effect size and heterogeneity for MR studies*

Model	**k**	**Cohen's *d***	**95% CI**	**Q**	***I²* (%)**
Random-Effects	27	0.299*	[0.064, 0.533]	137.00	81.0

*Note. *p < .05. k = number of effect sizes; d = mean effect size (Cohen's d); CI = confidence interval; Q = Cochran's Q test for heterogeneity; I² = percentage of variability due to heterogeneity.*

Publication bias analysis

The potential for publication bias was evaluated using several methods. Visual inspection of the funnel plot (Supplementary Figure S3) revealed a generally symmetric distribution. This was supported by Egger's regression test (Egger et al., 1997), which was non-significant (p = .401), and Rosenthal's Fail-safe N (N = 213), which exceeded the

tolerance criterion of 5k + 10 (145). These results suggest that publication bias is unlikely to pose a significant threat to the validity of the findings for the MR meta-analysis.

Moderator analysis

Given the significant heterogeneity (I^2 = 81.0%), we analyzed 12 potential moderators to identify sources of variance. As summarized in Table 6, four of these variables significantly moderated the effect of MR on student achievement.

Table 6. *Summary of significant moderator analyses for MR*

Moderator Variable	***k***	***Q_B***	***df***	***p***
Educational Level	25	16.08	2	.000
Subject Area	25	7.38	2	.025
Gamification	27	11.46	1	.001
Control Group Method	27	4.26	1	.039

Note. k = number of effect sizes included in the analysis. Q_B = between-groups heterogeneity statistic. df = degrees of freedom. Full details for each category are in the Supplementary Material, Table S3.

Several important patterns emerged from the significant moderators. The effectiveness of MR varied significantly by educational level, with the largest effect observed in high school (d = 0.802), compared to much smaller effects in middle school and higher education. The subject area was also a significant factor, with MR demonstrating a large effect in Chemistry (d = 0.744) but only small effects in Biology and Physics.

Furthermore, the inclusion of gamification elements was a powerful moderator; gamified MR applications (d = 0.817) were substantially more effective than non-gamified ones (d = 0.200). Finally, the control group method was significant, with MR showing a moderate advantage over traditional methods (d = 0.502) but a negligible effect when compared to alternative methods (d = 0.006).

Other study characteristics, such as sensory modalities and instructional mode, did not significantly moderate the effect of MR. The complete results for all 12 moderator analyses are presented in the Supplementary Material (Table S3).

Discussion

This study synthesized a large body of research to provide a comparative, meta-analytic overview of the effectiveness of VR, AR, and MR in science education. The primary finding was that all three immersive technologies were associated with improvements in student achievement, although their effects varied in magnitude. The medium-to-large effects of AR (d = 0.641) and VR (d = 0.531) align with previous meta-analyses (e.g., Merchant et al., 2014; Garzón & Acevedo, 2019), but this study extends the literature by providing a direct, science-specific comparison across the XR spectrum, including the emerging field of MR.

However, the high heterogeneity (I^2 > 80) across all three meta-analyses is a critical finding in itself, underscoring that the effectiveness of XR is not monolithic but highly context-dependent. This variability aligns with the complex interplay between media and method. As

Clark (1985) and Kozma (1994) debated, the medium itself may not be the sole cause of learning; rather, its unique attributes, when combined with sound pedagogical strategies, can influence cognitive processes and learning outcomes. The observed heterogeneity indicates that variance in effect sizes is driven by more than sampling error and is likely shaped by contextual and implementation-related factors that are not always captured consistently in primary studies. One such critical factor is usability. The ease of use, interface clarity, and technical stability of an XR application can profoundly impact the learning experience and, consequently, academic achievement (Ramli et al., 2024). For instance, Huang and Lee (2022) identified usability-related factors such as interactive quality and dynamic compatibility as crucial for positive learning experiences in VR. Although such characteristics are rarely reported in standardized ways across studies, they likely represent an important source of the heterogeneity observed in this synthesis.

The critical role of context: Interpreting moderator effects

The influence of subject area was a consistent and powerful moderator. The strong effect of VR on Physics, and the significant effectiveness of AR and MR in Chemistry, suggests that immersive technologies are particularly advantageous for visualizing the abstract, dynamic, and often invisible phenomena prevalent in the physical sciences (Hennessy et al., 2007). This pattern can be theoretically interpreted through Mayer's (2009) multimedia learning theory and Sweller's (1988) cognitive load theory. In physical sciences, where students must manipulate variables and observe micro-level interactions, XR reduces extraneous cognitive load by making complex spatial relationships more comprehensible (Chen, 2006). Conversely, the comparatively smaller effect in Biology may indicate that some biology topics rely more on descriptive or classification-based learning goals, which may reduce the added value of high-level abstract visualization in certain contexts.

Furthermore, the findings on the importance of instructional design—where interactive simulations and games significantly outperform passive applications—are consistent with recent literature emphasizing pedagogical strategy. This result provides quantitative support for the conclusions of Gil Parga et al. (2024), whose systematic review argued that the educational success of AR is inextricably linked to its underlying pedagogical design. The results suggest that when XR applications are designed as active, constructivist learning tools, as seen in our findings for VR simulations ($d = 0.634$) and gamified MR ($d = 0.817$), they yield substantially larger learning gains. The field's progression towards empowering educators in this process is further highlighted by the development of authoring tools like VR-Peas, which enable teachers to create their own pedagogically-sound VR scenarios (Oubahssi et al., 2024).

Educational level also emerged as a crucial factor for AR (middle school) and MR (high school). This may reflect a developmental stage in which students possess the necessary abstract reasoning skills to engage with technology. Specifically, the effectiveness of MR in high school can be interpreted through the lens of embodied learning. As the phenomenologist Merleau-Ponty (1962) argued, there is an inseparable unity between the mind and the body. MR can involve students' bodies in the physical environment while interacting with digital holograms, allowing for a more profound connection between the learner and scientific phenomena (Ali et al., 2019; Johnson-Glenberg et al., 2014).

Furthermore, the instructional design of the XR application is paramount. For VR, interactive simulations and games were significantly more effective than passive applications,

reinforcing the principles of active, constructivist learning (Aiello et al., 2012). Similarly, gamification dramatically increased MR's effectiveness, likely by enhancing motivation and engagement. However, the relatively smaller overall effect of MR (d = 0.299) can be interpreted through Gartner's hype cycle (Steinert & Leifer, 2010), suggesting that as this nascent technology matures toward the 'slope of enlightenment,' more refined and effective pedagogical implementations may emerge. A particularly insightful finding was the negative effect of VR relative to concrete 3D models (d = -0.798), suggesting that physical manipulatives may offer advantages in contexts requiring direct tactile feedback and hands-on spatial exploration (Barrett et al., 2015).

Limitations

This meta-analysis has several limitations. First, the search was restricted to studies in English and Turkish, which may introduce a language bias. Second, due to the emerging nature of the field, particularly for MR, the number of studies available within certain moderator subgroups was small, limiting statistical power and the stability of subgroup estimates. Third, high heterogeneity suggests that other unmeasured variables (e.g., quality of implementation, teacher training) are likely to contribute to the variance. Finally, we acknowledge the blurring lines between AR and MR in the primary literature, where the frequent misuse of terms can hinder accurate classification (Chytas et al., 2022).

Implications for practice and future research

The findings offer several practical implications. For educators and instructional designers, the message is clear: the choice of technology should be deliberate and context-aware. The adoption of XR alone is unlikely to be sufficient unless it is supported by thoughtful pedagogical integration. We recommend tailoring the technology to the discipline (e.g., VR for physics, interactive AR for middle school chemistry) and prioritizing interactive, pedagogically sound designs over passive experiences.

For researchers, this study highlights several avenues for future inquiry. Further primary research on MR is needed to build a stronger evidence base, as also noted in recent reviews (Fernández-Cerero et al., 2025). In addition, future studies should adopt more consistent terminology to differentiate AR from MR. Future meta-analyses should also explore other learning outcomes beyond achievement, such as motivation and critical thinking. Furthermore, we recommend research that moves beyond the mere presence or absence of features (e.g., embodiment) to investigate their qualitative implementation and conceptual relevance. For example, future work could specifically investigate the impact of different multimodal interactions (e.g., haptic feedback) on learning, as explored by Hu, Liu, and Xie (2025). Moreover, synthesizing research based on the intended pedagogical goals of the application, perhaps using a framework like Bloom's Revised Taxonomy as demonstrated by Gil Parga et al. (2024), could reveal a more nuanced understanding of how XR technologies support different levels of cognitive complexity. Investigating these qualitative and pedagogical dimensions is crucial for developing and testing nuanced models of Cognitive Load Theory specifically within XR contexts, thereby contributing to a deeper understanding of the conditions and mechanisms through which these technologies support learning.

Conclusion

This comprehensive meta-analysis provides evidence suggesting that XR technologies can serve as potentially effective tools for enhancing student achievement in science

education. By synthesizing two decades of quantitative research, this study showed that immersive technologies generally outperformed traditional instructional methods, although their effectiveness was not uniform across the XR spectrum. AR emerged as the most impactful tool (d = 0.641), followed by VR (d = 0.531), while MR currently demonstrates a smaller but promising effect (d = 0.299).

The findings suggest that XR technologies may help connect abstract scientific concepts with students' experiential understanding. The stronger effects observed in Physics and Chemistry across the three technologies suggest that immersive tools may be particularly useful for visualizing invisible, micro-level, or dynamic phenomena. This pattern may be interpreted through cognitive load theory and multimedia learning principles; by providing intuitive, 3D representations of complex spatial relationships, XR effectively reduces the extraneous cognitive load that often hinders learning in the physical sciences.

A key implication of this synthesis is that technology alone does not guarantee improved learning outcomes; rather, its effectiveness is closely linked to pedagogical design and contextual factors. The superior performance of interactive simulations and gamified applications over passive representation-based tools indicates that the most significant learning gains occur when students are active participants in their own knowledge construction. Furthermore, the stages associated with stronger effects, namely middle school for AR and high school for MR, suggest that these technologies may align with specific stages of cognitive development. In particular, the effectiveness of MR in higher grades highlights the potential of embodied learning; by integrating the physical body and environment into digitally mediated experiences, hybrid spaces foster a deeper, more holistic engagement with scientific inquiry.

The relatively smaller effect size of MR can be interpreted as a reflection of its current position in the Gartner Hype Cycle. As a nascent technology characterized by technical complexity and evolving terminology, MR is still moving toward a more mature stage of pedagogical implementation. As hardware becomes more accessible and educators gain access to user-friendly authoring tools, the potential of MR to support more integrated, hybrid science learning environments may increase.

In conclusion, while XR technologies offer a valuable instructional resource, their successful implementation requires a deliberate, context-aware approach. Educators and designers should prioritize interactive, constructivist designs and tailor technology choices to the specific demands of the scientific discipline. By supporting the visualization of abstract and invisible phenomena, XR may contribute to more immersive, interactive, and meaningful forms of science learning.

Declarations

Acknowledgments

This article is derived from the doctoral dissertation of the first author. We would like to thank the thesis committee members and academic mentors who provided feedback during the research process.

Funding:

The authors received no financial support for the research, authorship, and/or publication of this article.

Ethics Statements:
This study does not contain any studies with human participants and/or animals performed by any of the authors.
Conflict of Interest:
The authors declare that they have no competing interests.
Informed Consent:
Not applicable. This meta-analysis used only data from previously published studies and did not involve recruitment of participants or collection of any new individual-level data.
Data availability:
The dataset used in this study consists of coded information from published articles. All data supporting the findings of this study are provided in the Supplementary Materials.

References

Aiello, P., D'elia, F., Di Tore, S., & Sibilio, M. (2012). A constructivist approach to virtual reality for experiential learning. *E-Learning and Digital Media, 9*(3), 317–324. http://dx.doi.org/10.2304/elea.2012.9.3.317.

Ali, A. A., Dafoulas, G., Augusto, J. C., & Ibrahim, S. (2019). Collaborative educational environments incorporating mixed reality technologies: A systematic mapping study. *IEEE Transactions on Learning Technologies, 12*(3), 321–332.

Armstrong, H. E. (1902). The heuristic method of teaching. *School Science and Mathematics, 1*(8), 395–401.

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators Virtual Environments, 6*(4), 355–385. https://doi.org/10.1162/pres.1997.6.4.355

Barrett, A., Pack, A., Guo, Y., & Terlecki, M. (2015). Constrained interactivity for the enhancement of learning: An empirical study. *Computers & Education, 83*, 196–205. https://doi.org/10.1016/j.compedu.2014.12.009

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101. https://doi.org/10.2307/2533446

Chen, C. C. (2006). Are spatial visualization abilities relevant to virtual reality? *E-Journal of Instructional Science and Technology, 9*(2), 1–16.

Christensen, R., Eichhorn, K., Prestridge, S., Petko, D., Sligte, H., Baker, R., Alayyar, G., & Knezek, G. (2018). Supporting learning leaders for the effective integration of technology into schools. *Technology, Knowledge and Learning, 23*(3), 457–472. https://doi.org/10.1007/s10758-018-9385-9

Chytas, D., Piagkou, M., Demesticha, T., Tsakotos, G., & Natsis, K. (2022). Are extended reality technologies (ERTs) more effective than traditional anatomy education methods? *Surgical and Radiologic Anatomy, 44*, 1215–1218. https://doi.org/10.1007/s00276-022-02998-5

Clark, R. E. (1985). Confounding in educational computing research. *Journal of Educational Computing Research, 1*(2), 137–148.

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

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Thousand Oaks, CA: Sage.

Dede, C. (1998). Introduction. In C. Dede (Ed.), *Learning with Technology: The 1998 ASCD Yearbook* (pp. v–x). Alexandria, VA: Association for Supervision and Curriculum Development.

Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher*, *23*(7), 5–12. https://doi.org/10.2307/1176933

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Fernández-Cerero, J., Fernández-Batanero, J. M., & Montenegro-Rueda, M. (2025). Possibilities of extended reality in education. *Interactive Learning Environments, 33*(1), 1–15. https://doi.org/10.1080/10494820.2024.2342996

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665–694.

Garzón, J., & Acevedo, J. (2019). Meta-analysis of the impact of augmented reality on students' learning gains. *Educational Research Review, 27*, 244–260. https://doi.org/10.1016/j.edurev.2019.04.001

Gil Parga, S., Singh, U., Gutierrez, J., & Marks, S. (2024). Pedagogical design in education using augmented reality: A systematic review. *Interactive Learning Environments, 32*(8), 4219–4236. https://doi.org/10.1080/10494820.2023.2195445

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504.

Hennessy, S., Wishart, J., Whitelock, D., Deaney, R., Brawn, R., La Velle, L., McFarlane, A., Ruthven, K., & Winterbottom, M. (2007). Pedagogical approaches for technology-integrated science teaching. *Computers & Education*, *48*(1), 137–152. https://doi.org/10.1016/j.compedu.2006.02.004

Hu, H., Liu, G., & Xie, T. (2025). Multimodal interaction in virtual reality supported education: A systematic review. *Interactive Learning Environments, 33*(1), 170–191. https://doi.org/10.1080/10494820.2024.2342993

Huang, H., & Lee, C. F. (2022). Factors affecting usability of 3D model learning in a virtual reality environment. *Interactive Learning Environments, 30*(5), 848–861. https://doi.org/10.1080/10494820.2019.1691605

Johnson-Glenberg, M. C., Birchfield, D., Tolentino, L., & Koziupa, T. (2014). Collaborative embodied learning in mixed reality motion capture environments: Two science studies. *Journal of Educational Psychology, 106*(1), 86–104. https://doi.org/10.1037/a0034008

Klopfer, E., & Squire, K. (2008). Environmental Detectives—the development of an augmented reality platform for environmental simulations. *Education Technology Research and Development, 56*, 203–228. https://doi.org/10.1007/s11423-007-9037-6

Kozma, R. B. (1994). A reply: Media and methods. *Educational Technology Research and Development, 42*, 11–14.

Lainema, K., Lainema, T., Hämäläinen, R., & Heinonen, K. (2019). Going beyond technological affordances: Assessing organizational and socio-interactional affordances. In *Proceedings of the 16th International Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2019)* (pp. 323–330). IADIS. https://doi.org/10.33965/celda2019_2019111040

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

McFarlane, A., & Sakellariou, S. (2002). The role of ICT in science education. *Cambridge Journal of Education, 32*(2), 219–232. https://doi.org/10.1080/03057640220147568

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: a meta-analysis. *Computers & Education, 70*, 29–40. https://doi.org/10.1016/j.compedu.2013.07.033.

Merleau-Ponty, M. (1962). *Phenomenology of perception* (C. Smith, Trans.). Routledge. *(Original work published 1945.)*

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2. ed.). SAGE Publications.

Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information Systems, 77*(12), 1321–1329.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157–159. https://doi.org/10.2307/1164923

Oubahssi, L., Piau-Toffolon, C., & Mahdi, O. (2024). VR-Peas: A Virtual Reality PEdAgogical Scenarisation tool. *Interactive Learning Environments, 32*(10), 7212–7229. https://doi.org/10.1080/10494820.2024.2308094

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Ramli, R. Z., Wan Husin, W. Z., Elaklouk, A. M. S., & Sahari @ Ashaari, N. (2024). Augmented reality: A systematic review between usability and learning experience. *Interactive Learning Environments, 32*(10), 6250–6266. https://doi.org/10.1080/10494820.2023.2255230

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Schrum, L., & Levin, B. B. (2009). *Leading 21st century schools: Harnessing technology for engagement and achievement.* Thousand Oaks, CA: Corwin Press.

Shelton, B. E., & Stevens, R. (2004). Using coordination classes to interpret conceptual change in astronomical thinking. In Y. B. Kafai, W. A. Sandoval, N. Enyedy, A. S. Nixon, & F. Herrera (Eds.), *Embracing Diversity in the Learning Sciences: Proceedings of the Sixth International Conference of the Learning Sciences* (p. 634). Routledge. https://doi.org/10.4324/9781410611017-133

Statista. (2022). *Extended reality (XR): AR, VR, and MR in the United States - statistics & facts.* Retrieved from https://www.statista.com/topics/7524/extended-reality-xr-ar-vr-and-mr-in-the-us/

Steinert, M., & Leifer, L. J. (2010). Scrutinizing Gartner's hype cycle approach. In *PICMET 2010 Technology Management for Global Economic Growth* (pp. 1–13). IEEE.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Villena-Taranilla, R., Tirado-Olivares, S., Gutiérrez, R. C., & González-Calero, J. A. (2022). Effects of virtual reality on learning outcomes in K-6 education: A meta-analysis. *Educational Research Review, 35*, 100434. https://doi.org/10.1016/j.edurev.2022.100434

Wu, B., Yu, X., & Gu, X. (2020). Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis. *British Journal of Educational Technology, 51*(6), 1991–2005. https://doi.org/10.1111/bjet.13023

Zacharia, Z. C. (2007). Comparing and combining real and virtual experimentation: An effort to enhance students' conceptual understanding of electric circuits. *Journal of Computer Assisted Learning, 23*(2), 120–132. https://doi.org/10.1111/j.1365-2729.2006.00215.x