

UC Santa Barbara

UC Santa Barbara Previously Published Works

Title

Confounded or Controlled? A Systematic Review of Media Comparison Studies Involving Immersive Virtual Reality for STEM Education

Permalink

<https://escholarship.org/uc/item/8zv0283x>

Journal

Educational Psychology Review, 36(3)

ISSN

1040-726X 1573-336X

Authors

Lawson, Alyssa P

Marchand Martella, Amedee

LaBonte, Kristen

et al.

Publication Date

2024-07-03

DOI

10.1007/s10648-024-09908-8

Peer reviewed



Confounded or Controlled? A Systematic Review of Media Comparison Studies Involving Immersive Virtual Reality for STEM Education

Alyssa P. Lawson^{1,2}  · Amedee Marchand Martella² · Kristen LaBonte² · Cynthia Y. Delgado² · Fangzheng Zhao² · Justin A. Gluck³ · Mitchell E. Munns² · Ashleigh Wells LeRoy⁴ · Richard E. Mayer²

Accepted: 12 June 2024 / Published online: 3 July 2024
© The Author(s) 2024

Abstract

A substantial amount of media comparison research has been conducted in the last decade to investigate whether students learn Science, Technology, Engineering, and Mathematics (STEM) content better in immersive virtual reality (IVR) or more traditional learning environments. However, a thorough review of the design and implementation of conventional and IVR conditions in media comparison studies has not been conducted to examine the extent to which specific affordances of IVR can be pinpointed as the causal factor in enhancing learning. The present review filled this gap in the literature by examining the degree to which conventional and IVR conditions have been controlled on instructional methods and content within the K-12 and higher education STEM literature base. Thirty-eight published journal articles, conference proceedings, and dissertations related to IVR comparison studies in STEM education between the years 2013 and 2022 were coded according to 15 categories. These categories allowed for the extraction of information on the instructional methods and content characteristics of the conventional and IVR conditions to determine the degree of control within each experimental comparison. Results indicated only 26% of all comparisons examined between an IVR and conventional condition were fully controlled on five key control criteria. Moreover, 40% of the comparisons had at least one confound related to instructional method and content. When looking at the outcomes of the studies, it was difficult to gather a clear picture of the benefits or pitfalls of IVR when much of the literature was confounded and/or lacked sufficient information to determine if the conditions were controlled on key variables. Implications and recommendations for future IVR comparison research are discussed.

Alyssa P. Lawson and Amedee Marchand Martella co-first authorship.

Extended author information available on the last page of the article

Keywords Control-treatment equivalency · Immersive virtual reality · Media comparison · STEM education · Systematic review

Objective and Rationale

Media comparison research involves comparing the learning outcomes of students who learn with different instructional media such as with immersive virtual reality versus conventional modes of instruction (e.g., slideshow presentations, textbook readings, or video lessons). In other words, students in one condition are taught the same content with one type of medium, and students in another condition are taught with a different type of medium, and the learning outcomes of the two are compared (Warnick & Burbules, 2007). Despite the prevalence of these kinds of studies, a persistent challenge in media comparison research is to implement control and treatment conditions that differ with respect to the instructional medium but are equivalent with respect to the instructional methods and content they contain (Clark, 1983, 1994a, 1994b, 2012).

Our goal is not to contribute to the long-standing debate on the merits of media comparison research but rather is to pinpoint ways to improve the methodological rigor of this literature base. Our particular interest is in research on immersive virtual reality (IVR) for STEM education which has been the focus of recent media comparison research. IVR refers to a computer-supported device that visually transports and immerses a learner in a new computer-generated environment and allows users to feel they are present in an environment that is different from their physical environment (Immersive Virtual Reality, 2008). IVR differs from augmented reality wherein augmented reality presents computer-generated items overlaid in a learner's physical environment (Carmigniani & Furht, 2011).

The use of IVR has been rising in popularity within the last decade, with strong claims made about the effectiveness of using IVR in STEM education. However, the extent to which the specific affordances of IVR can be pinpointed as the causal factor in enhancing learning has yet to be systematically investigated. As Cromley et al. (2023) found in their meta-analysis on the use of virtual reality (VR) in STEM learning, the strongest effects of VR were for those conditions that included active learning techniques—that is, benefits were found when specific instructional methods were embedded in VR. Similarly, Conrad et al. (2024) found that IVR is advantageous as compared to other media when learners are actively involved as compared to passively involved. It is important to highlight that active learning techniques are not necessarily unique to the VR device and can be used in conventional media, thus potentially conflating the usefulness of VR itself. Therefore, the aim of this systematic review was to examine whether conclusions about the effectiveness of the IVR technology itself can confidently and appropriately be made or are limited by the confounding of other instructional methods and content that is non-specific to the technology. This systematic examination of the literature provides important insights into the degree of confidence we can have in conclusions about the effectiveness of the IVR technology itself.

Background on Technology in Education

The use of technology in education has been popular for over a century, with new waves of learning media occurring with the introduction of technologies such as motion pictures, radio, educational television, overhead projectors, programmed teaching machines, video, personal computers, the Internet, and extended reality devices (Cuban, 1993; Purdue Online, 2024; Saettler, 1990). As noted in the 1980s, the introduction of each new medium brings with it advocates who argue that students will experience learning improvements *because* of these new technologies (Clark, 1983). In fact, much of the promotion of educational technology stems from the perspective that incorporating new technologies into the learning process will result in learning outcomes that are educationally significant (Reeves & Oh, 2017). These claims have led to subsequent media comparison studies to determine the effectiveness of a new instructional medium as compared to conventional media (Clark, 1983).

Media comparison studies are popular in educational psychology and educational technology despite methodological concerns that have been raised for over four decades (Buchner & Kerres, 2023; Honebein & Reigeluth, 2021). In the 1960's and 70's, when televised and computerized instruction were prominent, some researchers expressed that media comparison studies were not fruitful and that learning objectives could be achieved using a variety of different instructional media (Clark, 1983; Levie & Dickie, 1973). These concerns highlighted questions of whether media comparison studies were worthwhile and if media could influence learning. These questions became the focus of a special issue (Vol. 42, No. 2) in the journal *Educational Technology Research and Development* (ETRD) during 1994. This special issue on the media comparison debate involved researchers such as Clark (1994a, 1994b), Jonassen et al. (1994), Kozma (1994a, 1994b), Morrison (1994), Reiser (1994), and Shrock (1994).

One overarching issue raised in the debate was that studies involving the comparison of different media could be confounded. More specifically, Clark (1994a) argued that underlying all media are instructional methods, and it is these instructional methods that are the important ingredient rather than the medium itself. When he examined meta-analyses comparing audiotutorial and conventional instruction or comparing computerized and conventional college instruction, he identified uncontrolled methods and content (including differences in the time it took to complete lessons) that made it difficult to know whether the results could be attributed to the medium specifically or to other elements of the intervention (Clark, 1983). In more recent reviews, similar confounds have been identified. For example, Honebein and Reigeluth (2021) reviewed 39 media comparison articles (years 1980–2019) in the journal *ETRD* and 41 media comparison articles (years 2009–2018) in other journals and found that the majority of comparative articles confounded the instructional media and the instructional methods.

As stated by Clark (1983), "It was Mielke (1968) who reminded us that when examining the effects of different media, only the media being compared can be different. All other aspects of the treatments, including the subject matter content and

method of instruction, must be identical” (p. 448). This notion of isolating and controlling variables is a hallmark of rigorous, unconfounded experimental research and is critical for making valid, causal conclusions (Chen & Klahr, 1999; Martella et al., 2023). Revisiting the media comparison debate, there are some, like Clark (1994a, 1994b), who argue that media do not influence learning and are mere delivery trucks whereas there are others, like Kozma (1994a, 1994b), who argue that media have unique attributes and that we should be focusing on how the media and methods work together to facilitate meaning-making and knowledge construction. Regardless of the side of the debate a researcher is on—or if they fall somewhere in the middle—if we are to determine whether media influence learning, one or more instructional components should be selectively contrasted and studied in a systematic fashion to determine which components are most important for learning (e.g., see De La Paz, 2007).

One issue for research on instructional practices, such as those within science education, is the use of “baggage-laden terms” without the accompaniment of clear operational definitions of what different instructional conditions entail (Klahr, 2013, p. 14076). When key features of different types of instruction are not outlined, it is difficult to determine *how* they differ. Without an understanding of how they differ, pinpointing *why* one approach was more or less effective than another becomes challenging, if not impossible. In this effort of exploring the “why,” it is important to identify features that are unique to the medium itself—as some capabilities cannot be replicated/recreated via other media (Hastings & Tracey, 2005)—and those that can be controlled between instructional conditions, with the goal of shedding light on the role a specific medium plays in learning. These systematic approaches to media comparison research are as relevant and needed as ever with the wave of new technologies surfacing in the educational market.

Immersive Virtual Reality

Of the new technologies entering the market, IVR has emerged as a tool “poised to revolutionize education” (AlGerafi et al., 2023, p. 1). In recent years, there has been a large push to use IVR as a tool for teaching students new content, particularly STEM content. In fact, the global market for virtual reality in education is projected to reach 28.70 billion USD by 2030 (Fortune Business Insights, 2023). Companies such as *Meta* have been promoting the idea that the IVR technology they have been building is creating new opportunities for student learning and have even partnered with different universities to teach instructors about how to use immersive technology for learning in their classroom (Clegg, 2023). Suggestions for practitioners within the IVR literature are often of the form “using IVR can improve learning” (e.g., Coban et al., 2022; Villena-Taranilla et al., 2022).

This push for the use of IVR in classrooms stems, at least partly, from the generally positive results that the literature is presenting. When examining meta-analyses for information on the effectiveness of IVR in education as compared to more traditional or non-immersive approaches, the results often show positive effects on learning (e.g., Coban et al., 2022; Conrad et al., 2024; Villena-Taranilla et al., 2022;

Wu et al., 2020). Despite these positive effects on learning, media comparison studies can present challenges in determining whether it was the media itself that influenced learning, the methods themselves that were attributable to the results, or the two working in tandem to produce learning outcomes. As such, researchers have expressed the need for a “thorough, scientific discussion” of the designs and methods used in IVR research (Buchner, 2023, p. 1).

Systematic Review Research Questions

Considering the media comparison debate and the issues that can arise in the design of conditions within these studies, it is important to establish the extent to which the instructional methods and content are controlled between instructional conditions in IVR comparative studies. This examination would provide insight into whether the results of comparative studies could be attributed to specific affordances of the particular medium. Therefore, the present review was conducted to examine the extent to which the instructional methods and content were controlled (labeled throughout paper as the “degree of control”) between IVR and conventional conditions in STEM education research with students in K-12 and higher education.

Operationally, we defined IVR for the present review as involving a head-mounted display (such as from Oculus Quest or HTC Vive). Our focus was on STEM education given that the media comparison literature is abundant in this area and has not yet been systematically examined with regard to the designs and methods used. Further, the importance of ensuring diverse populations of students are attracted to and retained in STEM fields cannot be overstated (Palid et al., 2023). As noted in the visionary report from the National Science Foundation, students should have “an equitable opportunity to acquire foundational STEM knowledge” wherein an “understanding of how people learn with modern technology [is needed] to create more personalized learning experiences, to inspire learning, and to foster creativity from an early age” (National Science Foundation, 2020, p. 5).

There were two primary research questions that guided our review:

1. Are the instructional conditions in IVR comparison studies controlled on instructional method and content?
2. When looking at different study outcomes, how confounded are the comparisons between IVR and conventional conditions?

Method

The complete pre-registered search strategy and screening process are available on the Open Science Framework (OSF)¹. The OSF pre-registration protocol follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The only PRISMA guidelines that were not incorporated were those

¹ https://osf.io/3zbj6/?view_only=f8aa26cd27ed4bf889cf538143a0de59

outside the scope of the present review. A condensed version of our methodology is presented below.

Literature Search

We used two primary approaches to locate published articles, dissertations, and conference proceedings for consideration after an iterative search process that is detailed on OSF¹. First, seven databases were searched using a variety of search terms. Second, articles included in recently published meta-analyses and reviews of the related literature were examined. For our database searches, *OpenDissertations*, *ProQuest Dissertations and Theses*, *Compendex*, *Web of Science*, *ERIC*, *PubMed*, and *PsycInfo* were included as they cover STEM subjects. Document types were limited to dissertations (from *OpenDissertations* and *ProQuest Dissertations and Theses*), conference articles and proceeding papers (from *Compendex* and *Web of Science*, respectively), and peer-reviewed scholarly journal articles (from *Compendex*, *ERIC*, *PubMed*, *PsycInfo*, and *Web of Science*). Dissertations and proceeding papers were incorporated to include research that had not yet been published in peer-reviewed journals. The dates of inclusion were from January 1, 2013 to December 31, 2022. The starting year of 2013 was chosen for three reasons: (1) related, previous meta-analyses and critical analyses have often used 2013 as their start date for admission to their review (e.g., Hamilton et al., 2021; Wu et al., 2020); (2) the Oculus Rift HMD-VR headset (specifically the Oculus development kit V1) first debuted in 2013; and (3) the iterative search process (described in a document on OSF¹) determined that searches prior to a 2013 publication year did not result in relevant articles.

The final search string used in these databases was: [IN ABSTRACT] (“virtual reality” OR “head mounted display” OR “head-mounted display” OR “immersive learning” OR “immersive VR” OR “educational VR” OR “VR learning”) AND (training OR education OR educational OR learn OR learning OR classroom OR instruction OR teach OR teaching) AND (intervention* OR experiment* OR empirical OR control OR treatment* OR quantitative OR group)) NOT [IN TITLE] (“systematic review” OR “meta analysis” OR meta-analysis OR “case study” OR rehabilitati* OR elderly OR animal* OR “intellectual disability*” OR “physical disability*”). The search was limited to English language documents. The literature searches yielded 11,432 articles. However, after the first deduplication process in *Zotero*, the number was reduced to 9076 articles. After the second and final deduplication process in *Rayyan*, there were a total of 8973 unique articles (the PRISMA flow chart is presented in Fig. 1).

Criteria for Inclusion

The primary aim of this systematic review was to examine articles that compared learning outcomes involving STEM content of K-12 or higher education students who learned with IVR (e.g., involving a head-mounted display) versus with more traditionally used learning media (e.g., video lectures, textbook readings, or online

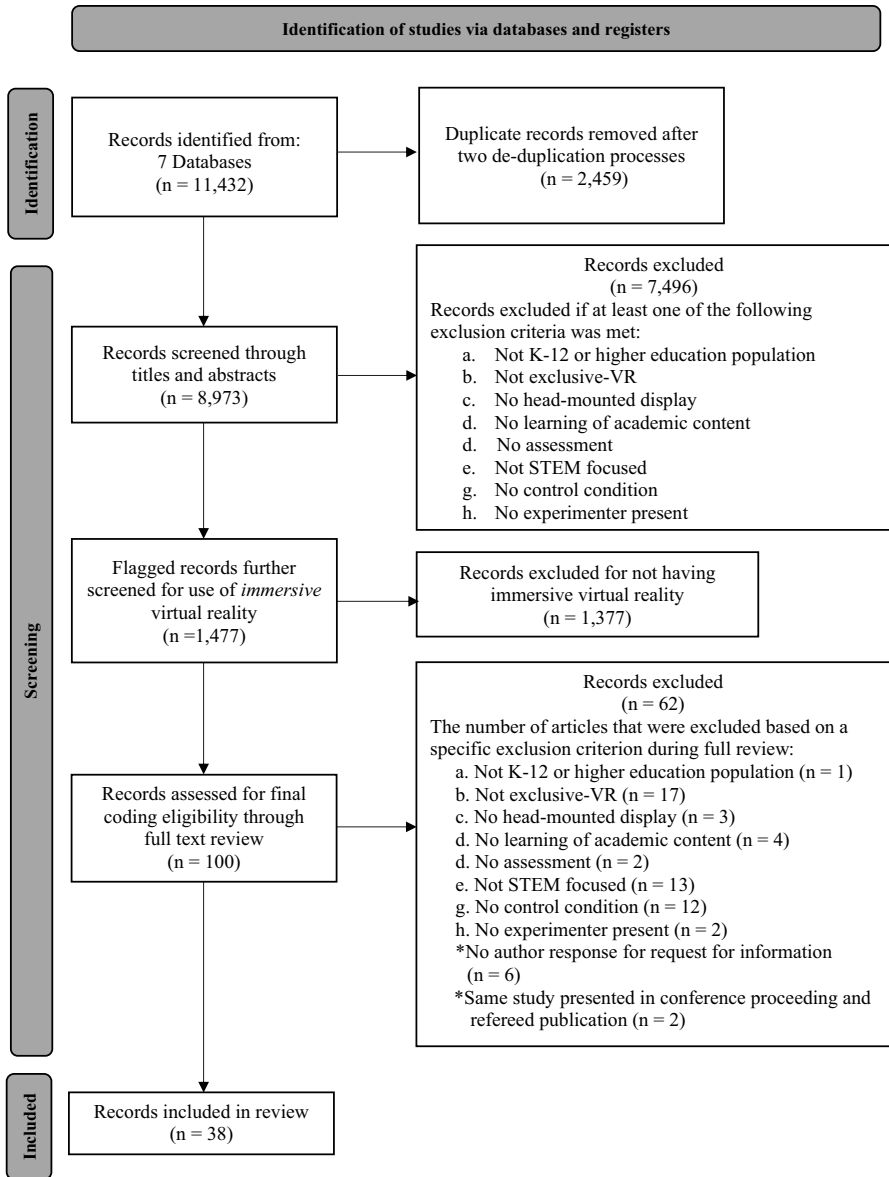


Fig. 1 PRISMA 2020 flow diagram

lessons). To be considered for inclusion in the present review, the following inclusion criteria had to be met: The articles needed to include (a) a non-clinical sample of children or adults in a K-12 or higher education environment; (b) a virtual reality learning experience that was fully immersive; (c) the use of a head-mounted display for the IVR device (i.e., not using a CAVE system); (d) a focus on presenting

learners with new *academic content* with an associated posttest assessment rather than practicing a skill, refining body movements, developing social or emotional skills, or developing spatial skills; (e) a lesson on STEM-related content (note: this criterion is different from the pre-registered protocol and was decided upon during screening due to the volume of articles and refined research questions); (f) an IVR condition where the main introduction to new content took place in IVR (i.e., students needed to *learn* new content exclusively within an IVR lesson but could *practice* content outside of IVR); (g) a conventional condition that reflected more traditional educational experiences (i.e., a learning environment that did not use IVR, desktop VR, simulations, or games-for-learning); and (h) an experimenter and/or instructor who was physically present to monitor participants (i.e., not a remote experiment) during the duration of the experiment in both conditions. The articles also needed to be original empirical work and be written in English.

Screening

Both co-first authors as well as the third and fourth authors conducted title-abstract screening for the 8973 articles (see Fig. 1). For purposes of interrater agreement, 20% of the articles were double screened (blinded review) by the third and fourth authors (12% and 8%, respectively). The interrater agreement level was 99.1%. The nine conflicts were discussed among the screeners, and resolutions were reached for each. After the screening process, 1477 articles were reviewed to determine whether the use of the term “virtual reality” referred to *immersive* virtual reality. After this process, 100 articles moved on to full-text review. The same screening authors were tasked with reviewing a random subset of these articles (percentage reviewed varied across screening authors), alongside the fifth, sixth, and seventh authors. Authors were contacted for clarification on any of the inclusion criteria. If the information received did not meet the inclusion criteria or authors did not respond, the article was excluded. After full-text review, the number of articles that were included in the final set was 38. The final list of articles is listed in OSF. PDFs for included articles were obtained from the University’s library holdings, through Interlibrary Loan, or through contacting authors.

Coding Procedure

Articles were coded across 12 primary categories (see Table 1) and three supplemental categories (see supplemental Table S1). The supplemental categories provided a deeper look into what the conditions involved (e.g., whether the conventional condition involved technology, what type of headset was used in the IVR condition). The primary categories were designed for the extraction of information needed to answer our two primary research questions. Each column in the spreadsheet represented a coding category, and each row, our unit of analysis, represented a comparison between one IVR condition and one conventional condition from each article. Articles could have multiple rows if they involved multiple comparisons (e.g., IVR 1 vs. conventional and IVR 2 vs. conventional; IVR vs. conventional 1 and IVR vs.

Table 1 Number (and percentage) of experimental comparisons in each coding designation across 12 coding categories

Coding description	N=50 Comparisons
Foundational details	
<i>Category 1: Were Operational Definitions of Conventional Condition Provided?</i>	
Yes, Complete	21 (42.00%)
Yes, Partial	14 (28.00%)
No	15 (30.00%)
<i>Category 2: Were Operational Definitions of IVR Condition Provided?</i>	
Yes, Complete	25 (50.00%)
Yes, Partial	14 (28.00%)
No	11 (22.00%)
<i>Category 3: Were Activities Involved in Conventional Condition?</i>	
Yes	10 (20.00%)
Likely Yes	0 (0.00%)
No	23 (46.00%)
Likely No	9 (18.00%)
Difficult to determine	8 (16.00%)
<i>Category 4: Were Activities Involved in IVR Condition?</i>	
Yes	25 (50.00%)
Likely Yes	0 (0.00%)
No	16 (32.00%)
Likely No	0 (0.00%)
Difficult to determine	9 (18.00%)
<i>Category 5: Did IVR Condition Involve Non-IVR Activities?</i>	
Yes	4 (8.00%)
No	42 (84.00%)
Difficult to determine	4 (8.00%)
<i>Category 6: What Were the Learning Outcome Results of the Comparisons?</i>	
IVR	12 (24.00%)
Conventional	1 (2.00%)
Tied	27 (54.00%)
Mixed	7 (14.00%)
No Stats/Questionable Stats	3 (6.00%)
Research Question 1	
<i>Category 7: Were Conditions Controlled on Activities?</i>	
Yes	18 (36.00%)
Likely Yes	2 (4.00%)
No	14 (28.00%)
Likely No	3 (6.00%)
Difficult to determine	13 (26.00%)
<i>Category 8: Were Non-IVR Activities Controlled Between Conditions?</i>	
Yes	3 (6.00%)
No	1 (2.00%)
Not Relevant	42 (84.00%)

Table 1 (continued)

Coding description	<i>N</i> = 50 Comparisons
Difficult to determine	4 (8.00%)
<i>Category 9: Did Conditions Receive Same Amount of Practice with Dependent Measure?</i>	
Yes	32 (64.00%)
No	6 (12.00%)
Difficult to determine	12 (24.00%)
<i>Category 10: Was Time Spent Learning Controlled Between Conditions?</i>	
Yes	28 (56.00%)
Likely Yes	1 (2.00%)
No	0 (0.00%)
Likely No	7 (14.00%)
Difficult to determine	14 (28.00%)
<i>Category 11: Was Content Matched?</i>	
Yes	33 (66.00%)
No	1 (2.00%)
Difficult to determine	16 (32.00%)
<i>Category 12: What Was Degree of Control Across Five Control Criteria?</i>	
<i>Approach 1:</i>	
Fully Controlled	13 (26.00%)
Mostly Controlled	11 (22.00%)
Somewhat Controlled	8 (16.00%)
Somewhat Not Controlled	10 (20.00%)
Mostly Not Controlled	6 (12.00%)
Fully Not Controlled	2 (4.00%)
<i>Approach 2:</i>	
Fully Controlled	13 (26.00%)
Mostly Controlled	6 (12.00%)
Somewhat Controlled	1 (2.00%)
Somewhat Not Controlled	3 (6.00%)
Mostly Not Controlled	0 (0.00%)
Fully Not Controlled	0 (0.00%)
Difficult to Determine	27 (54.00%)

conventional 2) across the same experiment or across multiple experiments. For articles that included factorial designs, we chose to include only those comparisons that were controlled on one of the two variables. For example, for a factorial design that compared media (IVR vs. conventional learning) with generative learning strategies (summarizing vs not summarizing), we only included the comparisons (1) IVR without summarizing vs traditional without summarizing and (2) IVR with summarizing vs traditional with summarizing. The total number of comparisons was 50 across the 38 articles (five articles included multiple comparisons, either with one IVR condition compared to two different conventional conditions or two IVR conditions compared to one conventional condition).

Coding Categories

The left column of Table 1 lists the names and possible designations for each of 12 primary coding categories. Each coding category was represented by a question with specific codes that addressed each question. These categories serve as the foundation for answering the two primary research questions.

Foundational Details of IVR and Conventional Conditions to Answer Research Questions 1 and 2

Categories 1 and 2: Were All Features of the Conventional Condition and the IVR Condition Operationally Defined? To determine how conditions differ and establish why one condition did or did not lead to greater learning than the other condition(s), specific definitions and procedures need to be provided within a study (Klahr, 2013; Martella et al., 2020). As such, we coded for whether a complete operational definition was provided for each condition. Although this variable is continuous, for the purposes of this review, we coded for whether there was a complete definition, a partial definition, or no definition of what occurred during the lesson in both the IVR and conventional conditions. A complete operational definition included specifics as to the methods, features, and/or procedures involved in each lesson. To be considered complete, the description of the condition would need to allow another researcher to be able to replicate the lesson and/or list out all essential features involved in the instructional conditions of the study (for example, see Parong & Mayer, 2020). A partial definition involved providing some information about the methods, features, and/or procedures involved in each lesson—that is, it went beyond a simple label—but did not include enough information to determine how that lesson was specifically implemented in practice (for example, see definition of conventional condition in Su et al., 2022). No definition involved simply labeling the condition (e.g., *traditional lecture*, *virtual reality lesson*) without describing what specifically occurred during the lesson or not providing any details about the condition.

Categories 3 and 4: Did the Conventional Condition and the IVR Condition Receive Activities During the Lesson? There is a large literature base on the effectiveness of including students in participatory activities during a lesson (often termed *active learning*) and allowing them to have opportunities to practice. Activities can include, for example, answering practice questions, completing class worksheets, working on a hands-on laboratory task, building concrete models, creating concept maps, and discussing content with peers. When these activities help students engage in generative learning (i.e., meaning is actively constructed through the organization of new information and the integration of this new information with prior knowledge; Fiorella & Mayer, 2015, 2016), they can be effective for student learning. We therefore coded for whether students received activities during the lesson to practice or extend their learning (coded *yes*) or did not engage in such activities (coded *no*), such as being instructed to simply watch a video or listen to a lecture.

It is important to note that the conventional condition could be coded as *likely no* for having activities if it was only described with a general term such as *traditional lecture* and no other information was provided to determine definitively whether activities were integrated into the lesson as traditional lecture is commonly referred to as passive learning (Deslauriers et al., 2019; Freeman et al., 2014; Hartikainen et al., 2019). The conventional condition could also be coded as *likely yes* for having activities if it was only described with a general term such as *active learning* and no other information was provided as active learning typically involves the use of activities to engage students in the learning process (Martella et al., 2023). When the IVR or conventional condition did not involve an operational definition or was not specifically labeled with a phrase that indicated activities were/were not likely involved (e.g., active learning condition, passive lecture condition), the condition was coded *difficult to determine*.

Category 5: Did the IVR Condition Involve Non-IVR Activities? When IVR conditions involve activities, these activities may not necessarily be implemented within the IVR lesson. Rather, these activities could be implemented outside of the IVR environment where they become non-IVR activities such as class discussions that take place in the physical classroom or paper-based worksheet activities. Therefore, we coded for whether an IVR condition included non-IVR activities (with the coding options *yes* and *no*). In the event that an IVR condition was not well operationalized and we could not determine if there were any activities that occurred outside of IVR, the condition was given the code *difficult to determine*.

Category 6: What Were the Learning Outcome Results of the Comparisons? To determine the achievement-based learning outcomes of each comparison between a conventional condition and an IVR condition, the learning outcomes were coded as *IVR better* (IVR condition had statistically significantly greater learning on all dependent measures), *conventional better* (conventional condition had statistically significantly greater learning on all dependent measures), *tied* (conditions resulted in nonsignificant differences on all dependent measures), *mixed* (the results differed based on the dependent learning measures examined [e.g., IVR was better on a measure(s) of transfer and conventional was better on a measure(s) of retention]), and *inconclusive* (statistics were not presented or were questionable).

Research Question 1: Are the Instructional Conditions in IVR Comparison Studies Controlled on Instructional Method and Content?

Category 7: Were the Conditions Controlled on Whether the Lessons Included Activities? To determine if there was a confound between the medium and the methods (e.g., IVR condition with practice activities vs. conventional condition with no practice activities), categories 3 and 4 were examined for consistency between conditions with coding options of *yes*, *likely yes*, *no*, and *likely no*. See Table 2 for how categories 3 and 4 were used to determine whether conditions in each comparison were controlled on participatory activities. If one or both of the conditions had received a code of *difficult to determine* for categories 3 and 4, the comparison would be given

Table 2 Examples of determining whether a comparison is controlled based on presence or absence of activities

Does conventional condition have activities?	Does IVR condition have activities?	Note	Are the conditions controlled on activities?
Yes	Yes	N/A	Yes
No	No	N/A	Yes
Likely No	No	Conventional condition labeled, e.g., “traditional lecture”	Likely Yes
Likely Yes	Yes	Conventional condition labeled, e.g., “active learning”	Likely Yes
No	Yes	N/A	No
Yes	Yes	Activity confound present	No
Likely No	Yes	Conventional condition labeled, e.g., “traditional lecture”	Likely No
Likely Yes	No	Conventional condition labeled, e.g., “active learning”	Likely No
Difficult to Determine	Yes	N/A	Difficult to Determine
Yes	Difficult to Determine	N/A	Difficult to Determine
Difficult to Determine	Difficult to Determine	N/A	Difficult to Determine

a code of *difficult to determine* in response to the conditions being controlled on participatory activities.

For those comparisons that met the requirements of having conditions that were controlled on the involvement of activities during the lesson, the conditions were further examined to ensure the activities were implemented in the same way between conditions. If there was a confound related to the activities (e.g., activities in one condition involved group work, and activities in the other condition involved independent work), the comparison was coded *no* for conditions being controlled on activities.

Category 8: Were Any Non-IVR Activities Controlled Between Conditions? If an IVR condition includes non-IVR activities and the conventional condition does not, it becomes difficult to strictly point to the immersive lesson as the causal factor. To determine if this variable was controlled between conditions, category 5 was examined, and for any comparisons where the IVR condition involved non-IVR activities, the conventional condition was evaluated for the presence of these activities as well. Therefore, we coded for whether any non-IVR activities were controlled between conditions with coding options of *yes* and *no*. If an IVR condition did not have non-IVR activities, the comparison was coded *not relevant*. In the event that an IVR or conventional condition was not well operationalized and we could not determine if there were any activities that occurred outside of IVR, the comparison was coded *difficult to determine*.

Category 9: Did Both Conditions Receive the Same Amount of Practice with the Dependent Measure? Receiving multiple opportunities to retrieve information from memory can be an effective way to improve students' retention of content (Roediger & Karpicke, 2006). As such, we coded for whether students in both conditions received the same amount of practice with the dependent measure (such as completing a procedural task similar to the one on the posttest or completing multiple-choice practice questions and then taking a multiple-choice posttest). The options were *yes* and *no*. In the event that the IVR and/or conventional condition was not well operationalized and we could not determine whether students in one condition receive more practice with the dependent measure than their peers in the other condition, the comparison was coded *difficult to determine*.

Category 10: Was the Time Spent Learning the Content Controlled Between Conditions? Different interventions may be afforded more time for students to engage with the content, leading to an issue surrounding whether it was the independent variable (i.e., particular teaching approach) that led to differences in student performance or a difference in time-on-task/the exposure to the content (Mason & Smith, 2020) or lesson efficiency due to more design effort given to one condition (Clark, 1983). To combat this potential confound, it is important to hold time spent learning in an experiment constant, meaning that both groups should be given the same or similar amount of time to learn the material. We therefore coded for whether students in the conditions compared in each study received the same amount of time to learn the content, within a range of 20% difference in time, with the options being *yes* and *no*. In some cases, if it was difficult to determine a strong "yes" or "no" due

to the comparison presenting ranges of times that learners could spend learning in either condition, the comparison was coded *likely yes* or *likely no* depending on if the range of times overlapped. Finally, if the authors did not provide a time range or specific duration of the lessons in the IVR and/or conventional condition and we could not determine if the time spent learning was controlled between conditions, the comparison was coded *difficult to determine*.

Category 11: Was the Content of the Lesson Matched? Lessons taught in the IVR condition and conventional condition should be matched on content (i.e., involve teaching the same concepts and general information) to ensure valid comparisons can be made. Although the degree to which the lessons were matched on content varies on a continuous scale, for the purposes of this review, the content was either deemed the same (with a code of *yes*) or different (with a code of *no*) between conditions. Examples of matched content could include statements such as “Concepts covered in the VR work were identical to those covered in each of the other conditions” (see Lamb et al., 2018, p. 21) or information that discussed consistency across conditions (see paragraph with heading “Consistency Across Conditions” in Petersen et al., 2022, p. 10). A *no* code would be given if there were differences in the topics/concepts/procedures presented in the lesson that could be identified based on the lesson or content description or if content differences were acknowledged. When details about the lesson involved in the IVR and/or conventional condition were not provided or an explicit statement that the content was identical were not made and we could not determine if the content was matched between conditions, the comparison was coded *difficult to determine*.

Category 12: What is the Degree of Control Across Five Control Criteria? To determine the degree to which the IVR and conventional conditions were controlled on instructional method and content in each article, five control criteria were assessed. Definitions of these criteria as well as examples of adherences to and violations of each criterion are shown in Table 3. These five control criteria were as follows:

1. Any activities involved in the lesson needed to be matched between the conditions (category 7). Code options for the question “were the conditions controlled on whether the lessons included participatory activities?” were *yes*, *likely yes*, *no*, *likely no*, and *difficult to determine*. The codes *yes* and *likely yes* counted as meeting this criterion.
2. Activities that were completed outside of IVR needed to be matched between conditions to isolate the effects of the immersive technology (category 8). Code options for the question “were the conditions controlled on any activities that occurred outside IVR?” were *yes*, *no*, *not relevant*, and *difficult to determine*. The codes *yes* and *not relevant* were counted as meeting this criterion.
3. Practice with the dependent measure needed to be matched between conditions (category 9). Code options for the question “did both conditions receive the same amount of practice with the dependent measure?” were *yes*, *no*, and *difficult to determine*. The code *yes* was counted as meeting this criterion.

Table 3 Definitions and example adherences and violations for the five control criteria

Control	Definition	Adherence	Violation
Matched on activities	Both conditions received the same activities, implemented in the same way OR both conditions did not receive any activities	IVR condition received a group pipetting task as did the conventional condition	IVR condition received a group pipetting task and the conventional condition received a solo pipetting task
Matched on non-IVR activities	The conventional condition received the same non-IVR activities as the IVR condition OR the IVR condition did not receive any non-IVR activities	The IVR condition received a paper-based worksheet as did the conventional condition	The IVR condition received a paper-based worksheet, but the conventional condition did not receive this worksheet activity
Matched on practice with dependent measure	Both conditions received practice with the dependent measure OR both conditions did not receive practice with the dependent measure	The IVR condition practiced retrieving information via multiple-choice questions as did the conventional condition	The IVR condition practiced retrieving information via multiple-choice questions but the conventional condition did not receive this practice
Matched on time spent learning the content	Both conditions received a lesson that was the same length (or within 20% difference)	The IVR condition lesson was 30 min in length as was the conventional condition lesson	The IVR condition lesson was 30 min in length but the conventional condition lesson was 15 min in length
Matched on the content	Both conditions received a lesson that involved the same topics, concepts, and/or procedures	The IVR condition lesson went into depth on five biology concepts as did the conventional condition	The IVR condition lesson went into depth on five biology concepts but the conventional condition lesson only briefly covered these five concepts

4. The time spent learning the material needed to be matched between conditions (category 10). Code options for the question “was the time spent learning the content controlled between conditions?” were *yes*, *likely yes*, *no*, *likely no*, and *difficult to determine*. The codes *yes* and *likely yes* were counted as meeting this criterion.
5. The content of the lessons needed to be matched (category 11). Code options for the question “was the content of the lesson matched?” were *yes*, *no*, and *difficult to determine*. The code *yes* was counted as meeting this criterion.

To determine the degree of control between IVR and conventional conditions in each article, we took two approaches. The first approach was to count how many of the five criteria were explicitly met. Across all conditions, *no*, *likely no*, and *difficult to determine* codes were counted as not meeting the control across all control criteria. If all five criteria were met, the comparison was deemed “fully controlled.” If four criteria were met, the comparison was deemed “mostly controlled.” If three criteria were met, the comparison was deemed “somewhat controlled.” If two criteria were met, the comparison was deemed “somewhat not controlled.” If one criterion was met, the comparison was deemed “mostly not controlled.” If zero of the five criteria were met, the comparison was deemed “fully not controlled.” See the top half of Table 4 for an example of how the degree of control was assigned for this first approach.

The second approach was identical to the first approach but with one difference. In this approach, *difficult to determine* codes were no longer counted as not meeting a criterion; rather, if a comparison was assigned this code for any of the five criteria, the comparison was labeled “degree of control difficult to determine.” The purpose of this approach was twofold. First, if a criterion was actually met but this information had not been presented clearly (or at all) in the article to be coded as such, the comparison would not be penalized. Second, this approach would afford insight into the number of comparisons that were affected by a lack of information provided in an article for at least one of the five criteria. See the bottom half of Table 4 for an example of how the degree of control was assigned for this second approach.

Research Question 2: When Looking at Different Study Outcomes, How Confounded Are The Comparisons Between IVR and Conventional Conditions?

To determine how confounded the comparisons were between IVR and conventional conditions when looking at different study outcomes, each comparison was examined across the five control criteria to determine if at least one of these criteria was explicitly not met (codes of *no* or *likely no*). If at least one control criterion was not met, regardless of whether another criterion was determined to be *difficult to determine*, the comparison was deemed “confounded.” If a comparison did not explicitly fail on any criterion but had at least one *difficult to determine* code, it was deemed “confounding difficult to determine.” Finally, if a comparison explicitly met all five criteria (codes of *yes*, *likely yes*, or *not relevant*), it was deemed

Table 4 Examples of degree of control based on the five criteria

Involvement in activities controlled	Non-IVR activities controlled	Same practice with dependent measure	Time spent learning controlled	Matched content	Degree of control
<i>Approach 1</i>					
Yes	Yes	Yes	Yes	Yes	Fully controlled
No	Yes	Yes	Yes	Yes	Mostly controlled
No	No	Yes	Yes	Yes	Moderately controlled
No	No	No	Yes	Yes	Moderately not controlled
No	No	No	No	Yes	Mostly not controlled
No	No	No	No	No	Fully not controlled
Difficult to determine	Yes	Yes	Yes	Yes	Mostly controlled
<i>Approach 2</i>					
Yes	Yes	Yes	Yes	Yes	Fully controlled
No	Yes	Yes	Yes	Yes	Mostly controlled
No	No	Yes	Yes	Yes	Moderately controlled
No	No	No	Yes	Yes	Moderately not controlled
No	No	No	No	Yes	Mostly not controlled
No	No	No	No	No	Fully not controlled
Difficult to determine	Yes	Yes	Yes	Yes	Degree of control difficult to determine

“not confounded.” These grouping of comparisons were then compared against category 6 (i.e., learning outcome results) for an examination of how confounded the comparisons were when looking at different study outcomes. See Table 5 for an example of how comparisons were determined to be confounded or not.

Interrater Agreement

There were two rounds of coding. In the first round, comparisons were coded according to categories 1, 2, 5, 8–11, S1, and S2. The third, fourth, fifth, sixth, and seventh authors each received a random number of articles to code across these categories. The co-first authors double coded each article under blinded conditions; the average agreement level for this first round of coding was 76.65%. Both co-first authors reviewed and resolved all discrepancies. With interrater agreement being lower than initially deemed acceptable (80% or higher), an independent research methodologist (see details on research methodologist in “Acknowledgements” section) coded all articles and compared his codes to the codes developed from the process described above. The average agreement level was 99.23%. The remaining discrepancies were reviewed and resolved by the co-first authors and the research methodologist.

To more thoroughly investigate the first research question, a second round of coding occurred wherein categories 3, 4, 6, 7, 12, and S3 were created and independently coded by the co-first authors under blinded conditions. The average agreement level was 90.60%. The remaining discrepancies were reviewed and resolved by these authors. As interrater agreement was deemed acceptable for this round of coding, the research methodologist was not brought in for coding of these additional categories.

Data Analysis

Frequency counts were determined for each code within the different coding categories. These frequencies were turned into a percentage of comparisons (out of 50) that received a particular code.

Results

Table 1 shows the number (and percentage) of experimental comparisons falling into each coding designation for each of the 12 primary coding categories (for a breakdown of results in the supplementary categories, see Table S1 in supplemental material). Results for each category follow.

Table 5 Examples of whether a comparison was deemed confounded or not

Involvement in activities controlled	Non-IVR activities controlled	Same practice with dependent measure	Time spent learning controlled	Matched content	Confounded?
Yes	Yes	Yes	Yes	Yes	Not confounded
No	Yes	Yes	Yes	Yes	Confounded
No	Difficult to determine	Yes	Yes	Yes	Confounded
Difficult to determine	Yes	Yes	Yes	Yes	Confounding difficult to determine

Foundational Details of IVR and Conventional Conditions in Service to Research Questions 1 and 2

Category 1: Were All Features of the Conventional Condition Operationally Defined?

The conventional condition had a complete operational definition in fewer than half of the comparisons (21 or 42.00%), a partial definition in 14 of the comparisons (28.00%), and no definition in 15 of the comparisons (30.00%). Therefore, 35 of the comparisons (70.00%) had at least a partial operational definition of the conventional condition.

Category 2: Were All Features of the IVR Condition Operationally Defined?

The IVR condition had a complete operational definition in half of the comparisons (25 or 50.00%), a partial definition in 14 of the comparisons (28.00%), and no definition in 11 of the comparisons (22.00%). Therefore, 39 of the comparisons (78.00%) had at least a partial operational definition of the IVR condition.

Category 3: Did the Conventional Condition Receive Activities During the Lesson?

The highest percentage of comparisons (23 or 46.00%) involved conventional conditions that did not involve participatory activities, with an additional nine (18.00%) that likely did not involve activities. Therefore, in 32 of the comparisons (64.00%), conventional conditions were coded as either *no* or *likely no* for involving activities. In these conditions, 18 were video slideshows/lectures, four were live lectures, two were lectures (format not specified), four were texts/readings, three were paper print-outs or images of the lesson, and one was projected diagrams. Only 10 of the comparisons (20.00%) had conventional conditions where participants received activities, and there were no comparisons that had these conditions coded as likely receiving activities. When the activities were explicitly named or described, they included, for example, self-explanations, lab exercises, item construction (e.g., a DNA molecule), and worksheets. Finally, in eight of the comparisons (16.00%), the conventional conditions were not labeled or described well enough to know whether the lesson involved activities.

Category 4: Did the IVR Condition Receive Activities During the Lesson?

Unlike the conventional conditions, half of the comparisons (25 or 50.00%) involved IVR conditions that did involve participatory activities. When the activities were explicitly named/described, they included, for example, self-explanations, lab exercises, practice multiple-choice questions, class discussion, and worksheets. Approximately one-third of the comparisons (16 or 32.00%) had IVR conditions in which

participants did not receive activities. In these conditions, 10 involved a passive, narrative tour through the lesson or a virtual lecture, and six had an added component where participants were able to interact with the virtual world given the immersion specifically afforded by IVR technology. These hands-on experiences were not specific activities to practice the content but did allow participants interact with the lesson using the IVR technology. For example, in Parong and Mayer (2020), participants could touch and move red blood cells that they encountered during the narrated tour of the parts and functions of a blood vessel and cell. Finally, in nine of the comparisons (18.00%), the IVR conditions were not described well enough to know whether the lesson involved activities.

Category 5: Did the IVR Condition Involve Non-IVR Activities?

Four of the IVR conditions (8.00%) involved activities that were completed outside of IVR. These included completing a worksheet, manipulating physical tools during a real-life task, engaging in a class discussion, and responding to questions asked by experimenter (i.e., providing oral self-explanations). Forty-two of the IVR conditions (84.00%) did not involve non-IVR activities, and four (8.00%) were difficult to determine for this category.

Category 6: What Were the Learning Outcome Results of the Comparisons?

In 12 of the comparisons (24.00%), the IVR condition was found to be statistically significantly better than the conventional condition. Only one comparison (2.00%) resulted in the conventional condition having statistically significantly better learning outcomes than the IVR condition. The most common outcome (27 or 54.00%) was the conditions resulting in non-statistically significant learning differences (i.e., they “tied”). The results were mixed for seven of the comparisons (14.00%) with different outcomes depending on which of the dependent learning measures were examined. Finally, the results were inconclusive in three of the comparisons (6.00%). Overall, there was not strong evidence that IVR was more or less effective than traditional media in promoting learning of STEM content.

Research Question 1: Are the Instructional Conditions in IVR Comparison Studies Controlled on Instructional Method and Content?

Category 7: Were the Conditions Controlled on Whether the Lessons Included Activities?

In 13 of the comparisons (26.00%), there was not enough information provided in the article to determine if the conditions were controlled on participatory activities. For those articles that did provide enough information: over one-third of all comparisons (18 or 36.00%) involved IVR and conventional conditions that were

controlled on activities, with an additional two comparisons (4.00%) demonstrating that this variable was likely controlled between conditions (total: 20 or 40.00%). The percentage of comparisons where the involvement of activities was not controlled between conditions was 14 (28.00%), with an additional three comparisons (6.00%) demonstrating that activities were likely not controlled between conditions (total: 17 or 34.00%).

For the 17 comparisons where activities were not controlled (or likely not controlled) between conditions, the IVR condition seemed to provide an advantage in terms of learning in 14 of them. These advantages consisted of the IVR condition having learning activities when the conventional condition did not (11 of the 14) and the IVR condition having a confound in the activity that seemed to favor it over the conventional condition (three of the 14). For example, in Lamb et al. (2018), participants in the IVR condition had to correctly complete each component of the DNA activity before they could progress but their peers in the conventional, hands-on activity condition did not receive this type of mastery criterion/feedback during their DNA activity.

The IVR condition did not seem to have an advantage in terms of learning in three of these 17 comparisons where activities were not controlled. In these three comparisons, there was a confound related to the activities that was difficult to determine whether it gave one of the two conditions an advantage over the other. For example, in Dunnagan et al. (2020), the participants in the conventional condition worked in groups of two on the lab exercise whereas their peers in the IVR condition worked one-on-one with a virtual teaching assistant (TA) who provided assistance as needed. It is difficult to say whether working alone with TA support or working as a group provided a potential learning advantage to one condition versus another. However, this difference could introduce a confound and create an advantage for one group over another. Similarly, in Petersen et al. (2022), participants in the conventional condition worked with small groups during the lesson whereas those in the IVR condition participated in the simulation individually and did not interact with peers. It is also difficult to say whether the individualized VR instruction was more or less advantageous than the interaction with group members afforded in the conventional condition; this methodological difference created a confound nonetheless (a limitation noted in their article).

Category 8: Were Any Non-IVR Activities Controlled Between Conditions?

In 42 of the comparisons (84.00%), the IVR condition did not have non-IVR activities and was given a *not relevant* code. Four of the comparisons (8.00%) involved at least one condition that was not described well enough to know if conditions were controlled on non-IVR activities. In three comparisons (6.00%), non-IVR activities were controlled between conditions. Finally, one comparison (2.00%) did not provide the conventional condition with the same kinds of activities that IVR participants completed outside of IVR.

Category 9: Did Both Conditions Receive the Same Amount of Practice with the Dependent Measure?

In the majority of comparisons (32 or 64.00%), both conditions were given the same amount of practice with the dependent measure. However, in six of the comparisons (12.00%), one condition was given more practice with the dependent measure than the other condition. Interestingly, it was always the IVR condition that received more practice with the dependent measure than the conventional condition. In 12 of the comparisons (24.00%), the description of the conditions was too ambiguous to determine whether both conditions received the same amount of practice with the dependent measure.

Category 10: Was the Time Spent Learning the Content Controlled Between Conditions?

Over half of the comparisons (28 or 56.00%) did demonstrate that the time spent with the learning materials was controlled between conditions, with one additional comparison (2.00%) coded as a “likely yes” for time being controlled. There were no comparisons where the time spent with the learning materials was not controlled between conditions; however, seven of the comparisons (14.00%) were coded as a likely no for time being controlled. Finally, 14 of the comparisons (28.00%) did not include a specification as to the length of time spent with the learning material across both conditions.

Category 11: Was the Content of the Lesson Matched?

The content was matched in 33 (66.00%) of the comparisons. However, 16 of the comparisons (32.00%) did not involve enough information to determine if the content taught was the same between the two conditions. There was only one comparison (2.00%) in which the content was not matched between the IVR and conventional conditions.

Category 12: What is the Degree of Control Across Five Control Criteria?

As previously discussed, there were two approaches to our examination of the degree of control across the five control criteria. In the first approach, *difficult to determine* codes were counted as not meeting the criterion. In the second approach, *difficult to determine* codes for any of the criteria led the comparison to be counted as degree of control difficult to determine. Figure 2 shows the number of comparisons that fit within each degree of control for the first approach, and Fig. 3 shows the number of comparisons that fit within each degree of control for the second approach. The results of each approach are presented below.

Approach 1 For the degree of control across comparisons, 13 (26.00%) were *fully controlled*, 11 (22.00%) were *mostly controlled*, eight (16.00%) were *somewhat controlled* comparisons, 10 (20.00%) *somewhat not controlled* comparisons, six (12.00%) were *mostly not controlled*, and two (4.00%) were *fully not controlled*.

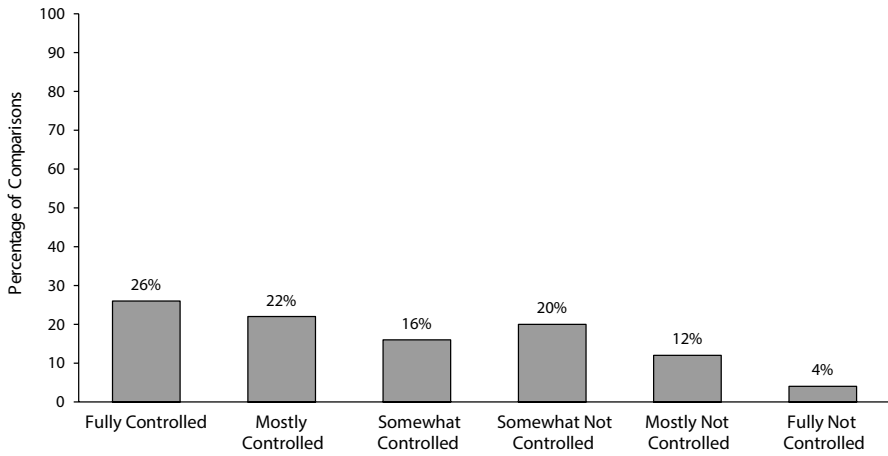


Fig. 2 Number of comparisons that were classified under each degree of control for approach 1

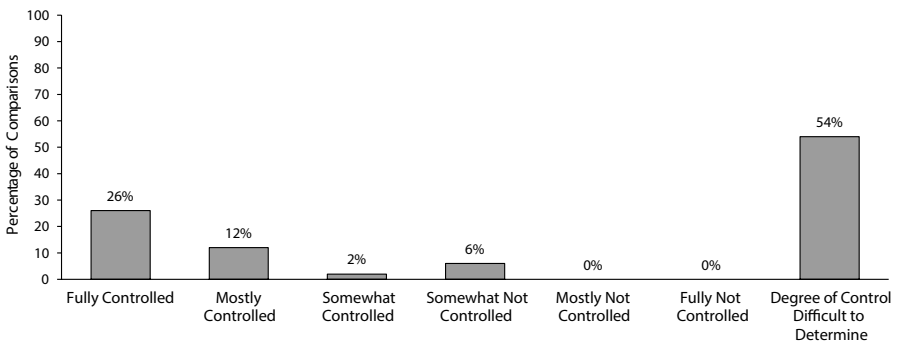


Fig. 3 Number of comparisons that were classified under each degree of control for approach 2

Approach 2 For the degree of control across comparisons, 13 (26.00%) were *fully controlled*, six (12.00%) were *mostly controlled*, one (2.00%) was *somewhat controlled*, three (6.00%) were *somewhat not controlled*, zero (0.00%) were *mostly not controlled*, and zero (0.00%) were *fully not controlled*. Finally, the number of *degree of control difficult to determine* comparisons was 27 (54.00%).

Research Question 2: When Looking at Different Study Outcomes, How Confounded Are The Comparisons Between IVR and Conventional Conditions?

As previously stated in the “[Method](#)” section, for the purposes of answering research question 2, a comparison needed to explicitly fail on at least one of the five criteria to be deemed “confounded.” There were 20 (40.00%) confounded comparisons in total. How confounded the comparisons were between IVR and conventional conditions when looking at different study outcomes are presented below and shown in Fig. 4.

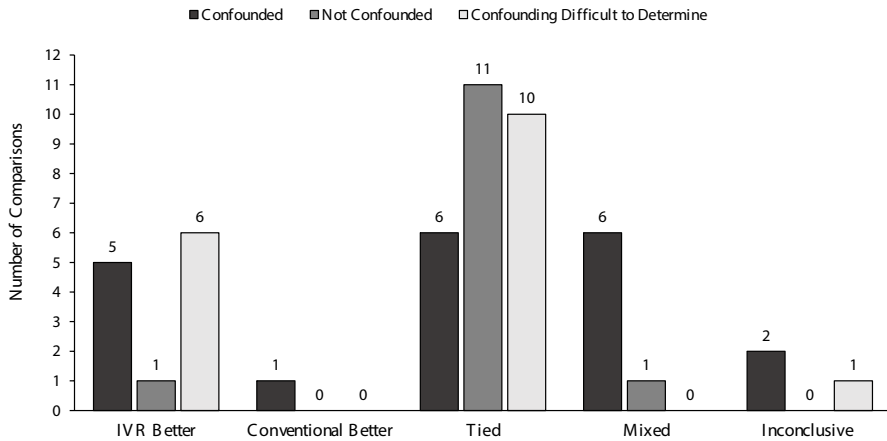


Fig. 4 Breakdown of study outcomes based on whether comparison is confounded

IVR better: For those 12 comparisons where the IVR condition resulted in statistically significantly better learning across all measures, five (41.67%) were confounded, one (8.33%) was not confounded, and six (50.00%) were difficult to determine as to whether they were confounded. **Conventional better:** There was only one comparison where the conventional condition was better across all measures, and it was confounded. **IVR and conventional tied:** For those 27 comparisons where the IVR and conventional conditions tied for all measures (i.e., were not statistically significantly different on learning outcomes, six (22.22%) were confounded, 11 (40.74%) were not confounded, and 10 (37.04%) were difficult to determine as to if they were confounded. **Mixed:** For those seven comparisons where the results were mixed, six (85.71%) were confounded, one (14.29%) was not confounded, and zero (0.00%) were difficult to determine as to if they were confounded. **Inconclusive:** For those three comparisons where the results were inconclusive (either because no inferential statistics were reported or the statistics were questionable), two (66.66%) were confounded, zero (0.00%) were not confounded, and one (33.33%) was difficult to determine as to if it was confounded.

Overall, we conclude that it is difficult to gather a clear picture of the benefits or pitfalls of IVR when much of the literature is confounded and/or lacks sufficient information to determine if the conditions are controlled on instructional methods and content.

Discussion

The present review was conducted to examine the extent to which the instructional methods and content were controlled between the instructional conditions in IVR comparison studies involving STEM content. Given the numerous methodological problems we identified in the current research base on media comparison studies involving IVR, our overarching goal is to improve the quality of media comparison

research in the field of educational psychology and educational technology, with IVR comparison studies as an example. Results of our critical analysis are discussed according to the two primary research questions.

Research Question 1: Are the Instructional Conditions in IVR Comparison Studies Controlled on Instructional Method and Content?

Degree of Control

There were five criteria on which we assessed the degree of control between IVR and conventional conditions in teaching STEM education. These criteria included controlling participatory activities, non-IVR activities, practice with the dependent measure, time spent learning the material, and the content of the lessons. For both the first and second approach to assessing the degree of control between conditions, we found that only 26% of comparisons were fully controlled—that is, the majority of comparisons did not meet all five control criteria. With the first approach where all five criteria needed to be explicitly met, 32 comparisons (64%) met three or more of the criteria and 18 comparisons (36%) met two or fewer. Therefore, a high percentage of articles had more control issues than not, leading to questions surrounding which features of the IVR conditions were attributable to the results of the study.

With the second approach, when any comparison received a difficult to determine code for one or more of the control criteria, it was labeled degree of control difficult to determine. With this approach, 20 comparisons (40%) met three or more of the criteria, three comparisons (6%) met two or fewer, and 27 comparisons (54%) lacked information for at least one of the criteria to determine the exact degree of control. This second approach lends insight into the number of comparisons that suffered from a lack of sufficient detail in the article to determine whether it was fully not controlled, mostly not controlled, somewhat not controlled, somewhat controlled, or mostly controlled. We therefore urge researchers to include and journal editors to require sufficient methodological details in IVR comparison papers.

Overall, the main findings are that just over one-quarter of the IVR comparison studies we reviewed were fully controlled across all five of our criteria and just over half lacked sufficient information on at least one of the criteria. We conclude that much work needs to be done to improve the methodological quality and reporting of media comparison studies involving IVR in STEM disciplines.

Frequent Confounds

When looking within each of the five criteria specifically, the number of comparisons that met each control criterion varied substantially. The criterion where issues were most glaring was whether the use of activities in each condition was held constant. In fewer than half of the 50 comparisons, activities were controlled (or likely controlled) between conditions whereas approximately one-third of comparisons were not controlled (or likely not controlled) on this variable. When conditions differ on the presence or absence of student engagement via participatory activities, the

instructional methods and instructional media become confounded. These confounds make it difficult to deduce why one condition outperformed (or did not outperform) the other. For example, in a study by Makransky et al. (2019), participants in the IVR condition received voice-over guidance, hands-on tasks, multiple-choice questions, and feedback (including elaboration) whereas those in the conventional condition (text condition) received a 14-page safety manual without any of these same instructional methods. Thus, not only did the use of IVR differ between conditions but so did the instructional methods used in each condition. As such, drawing conclusions about the unique benefits that IVR, as compared to conventional instructional methods, has for teaching STEM content is difficult. As noted in the meta-analysis by Cromley et al. (2013) and systematic review by Conrad et al. (2024), virtual reality conditions that included active learning techniques showed stronger effects on learning than when conditions were more passive. Therefore, it may be the case that these active learning techniques are the key ingredient for improved learning, regardless of the media that is used.

It is important to note that any active engagement that was unique to IVR conditions was not considered a confound of participatory activities. For example, allowing participants to touch and move red blood cells that they encountered during a narrated tour of the parts and functions of a blood vessel and cell in the IVR environment (see Parong & Mayer, 2020) was a unique affordance of the medium and was not counted as a confound. However, active learning strategies, particularly those that allowed students to practice or further encode the content, which were not unique to IVR, were problematic in confounding the instructional methods with the instructional media. Active learning strategies are not a unique feature of IVR as many conventional learning environments include the same types of strategies to promote learning (see Freeman et al., 2014; Martella et al., 2023; Stains et al., 2018 for discussions of the frequency of active learning in conventional STEM classrooms).

Occasional Confounds

Although not as problematic as the activity confound, there were differences between conditions on the amount of time participants spent learning the content of the lesson. We found that more than half of comparisons were controlled (or likely controlled) on the time spent learning the content whereas 14% were likely not controlled on this variable. Although the percentage of comparisons that had this confound was on the lower side, it is unclear how many of the 14 comparisons that did not provide a lesson duration had an issue with this control criterion. In order to make sound claims about the benefits of IVR, it is essential that comparison groups receive an equal amount of time engaging with the content. As such, the presence of this confound is problematic in that it is difficult to determine if it was the independent variable that caused the results or a difference in time-on-task, exposure to class content, or lesson efficiency due to more design effort in one of the conditions.

Another confound that occurred occasionally was related to how much practice each group received with the dependent measure. We found that in just under two-thirds of comparisons, both conditions were given the same amount of practice with

the dependent measure; however, 12% of comparisons were not controlled on this variable. Interestingly, it was always the IVR condition that received more practice with the dependent measure than the conventional condition. Although the percentage of comparisons that had this confound was on the smaller side, it is unclear how many of the 12 comparisons that did not provide sufficient information about what occurred in the conditions had an issue with this control criterion. However, the mere presence of this confound is problematic in that it provides students with multiple opportunities to retrieve information from memory or to familiarize themselves with the content and format of the test.

Infrequent Confounds

Not matching content between conditions was a confound that was scarcely seen within the comparisons examined in our review. More specifically, the content of the lessons was matched in just under two-thirds of the comparisons and was not matched in only one comparison. Although it is positive that we only found one instance of content not being matched, there were 16 comparisons that did not involve enough information to determine if the content taught was the same between the two conditions. Without specific details provided to ensure both conditions received identical content, it is difficult to confirm it was not differences in what students learned that affected the outcome of the study. More specifically, if the IVR condition is taught different content than the conventional condition, one cannot make valid comparisons on an assessment of their learning outcomes.

Finally, the majority of IVR conditions did not involve activities that occurred outside of IVR, meaning that any activity provided to students was conducted within the IVR environment. However, for those four comparisons that did involve non-IVR activities in the IVR condition, one did not provide the conventional condition with the same activities. If an IVR condition includes non-IVR activities and the conventional condition does not, it becomes difficult to strictly point to the immersive lesson as the causal factor. Although this confound only affected one out of four comparisons, it was still present in the literature base we reviewed and thus needs to be considered when designing future research to investigate the use of IVR in STEM education.

Research Question 2: When Looking at Different Study Outcomes, How Confounded Are The Comparisons Between IVR and Conventional Conditions?

When looking at the outcomes of the articles included in our systematic review, the IVR condition was most often found to be better for student learning when comparisons lacked sufficient information to determine if a confound was present or not (i.e., 50.00% of comparisons where IVR was better). When results were tied between conditions, the comparisons were most often found to not have a confound (i.e., 40.74%). In fact, the outcome of IVR and conventional conditions leading to similar learning was the only outcome in which there were more studies that were not confounded than were confounded. When results were mixed based on different

outcomes examined, the comparisons were most often found to be confounded (i.e., 85.71%). The results seem to suggest that the way in which the conditions have been designed and the attention given to designing studies without confounds between conditions may play a role in the outcomes of the studies—a major finding of this review. However, given that confounds were present across all learning outcomes, more rigorous, controlled research is needed to shed light on the effectiveness of IVR as compared to conventional instruction.

Recommendations Moving Forward

Given the common methodological issues identified in our systematic review, we offer three primary recommendations to researchers with the aim of improving the value of media comparison research involving IVR and STEM content.

First, we recommend researchers and journal editors ensure the IVR and conventional conditions in each study are well described, with specific methods and procedures outlined clearly. As demonstrated throughout the findings of this systematic review, many articles presented incomplete information concerning the way in which their conditions were designed and implemented. In fact, only 42% of comparisons involved conventional conditions that had a complete operational definition and only 50% of comparisons involved IVR conditions that had a complete operational definition. Without detailed knowledge of what occurred in each condition, it becomes more difficult to determine how well variables were controlled in each study. By providing more thorough descriptions of the conditions in each experiment and outlining all critical features of each instructional intervention, readers will be better able to compare between conditions and identify any potential limitations of the study.

Second, researchers need to intentionally design both the IVR and conventional conditions in media comparison research according to the research question of interest. For example, if the central question of interest is whether immersing students in a virtual world promotes greater learning, the only difference that should arise between conditions is immersion. Any other differences between the conditions (e.g., activities, time on task) would serve as confounds and prevent the researchers from drawing conclusions about the use of immersion, specifically, in learning. Given that research questions about IVR can have real-world implications (e.g., how content is taught in classrooms), it is imperative to determine the causal factor driving the results. Due to the fact that researchers, educators, and technology designers read and draw conclusions from research published in education journals, educational psychology journals, and educational technology journals, authors should emphasize what can and cannot be drawn from studies conducted in this area, particularly with an eye towards potential confounds in the research design.

Third, all critical variables outside of the independent variable, such as the ones presented in this paper, need to be controlled between conditions in order to draw causal conclusions regarding the impact of a particular medium or instructional method. Controlling variables is a basic tenant of experimental design but continues to be a persistent problem present in the media comparison research we reviewed.

The literature on the use of IVR in STEM education consists of a high percentage of comparisons between conventional and IVR conditions that are confounded. To draw conclusions about the unique affordances of IVR in STEM education, studies need to control and isolate variables such as those evaluated in this paper. This will allow for a better understanding of when and how IVR can serve as an effective instructional tool in STEM education and when it cannot.

Limitations

There were three potential limitations to the present review. First, the search strategy was limited to IVR conditions that were exclusively IVR and conventional conditions that were exclusively non-IVR. The purpose of this decision was to compare the primary method of teaching new content in IVR versus in another more traditionally used medium. The results of the present review may differ when the main learning event(s) takes place in the real-world and IVR is used as an active learning tool. Second, the search strategy was limited to studies involving STEM content, and the findings may not generalize to lessons involving non-STEM content. Third, extensive steps were taken to ensure the accuracy of the coding. Although the final levels of agreement were at or above 80%, it is possible that other researchers could code articles differently than the team of the present review. Similarly, other research teams may believe other criteria are important to include when asking a similar question. Therefore, the results and conclusions should be interpreted as based on our assessment of the articles which may or may not be representative of how other researchers would analyze them.

Conclusion

The media comparison literature involving IVR in STEM education contains a number of issues that need to be addressed if we are to determine the extent to which IVR is effective for STEM learning. Consider that just over one-fourth of comparisons involved conditions that were deemed “fully controlled” on five control criteria related to instructional method and content, and almost half of all comparisons had at least one confound related to the instructional methods and content. When confounded, IVR conditions were more likely to involve activities than the conventional condition and involve more practice with the dependent measure(s), presenting a potential confound(s) between the medium and the methods. Finally, a major concern of the reviewed literature base was that many articles did not present enough relevant information to determine whether conditions were controlled on critical variables. The present review suggests that future research should carefully address issues related to the design of conventional and IVR conditions in media comparison studies to gain a better understanding of the effects of different IVR and conventional interventions and move toward being able to provide more practical implementation recommendations.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10648-024-09908-8>.

Acknowledgements Additionally, we would like to thank Dr. Ronald C. Martella for his help as the independent research methodologist on this project. Dr. Martella has published 100 refereed journal articles and has over 200 total publication (including research methods textbooks and curricular products), and his expertise in this area was greatly appreciated.

Author Contribution We confirm that the manuscript has been read and approved by all authors; the author order has also been approved by all authors. We understand the corresponding author is the sole contact for the editorial process.

Funding One co-first author received support from the National Science Foundation STEM Education Postdoctoral Research Fellowship Program under grant number 2222208. The last author also received support from the Office of Naval Research grant N00014-21-1-2047.

Data Availability Data are available upon request and the complete search strategy is available on the Open Science Framework at https://osf.io/3zbj6/?view_only=f8aa26cd27ed4bf889cf538143a0de59.

Declarations

Competing Interests The authors declare no competing interests.

Disclaimer Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Office of Naval Research.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

*Papers analyzed in the systematic review and included in article text

- AlGerafi, M. A. M., Zhou, Y., Oubibi, M., & Wijaya, T. T. (2023). Unlocking the potential: A comprehensive evaluation of augmented reality and virtual reality in education. *Electronics*, *12*, 3953. <https://doi.org/10.3390/electronics12183953>
- Buchner, J. (2023). Beyond media comparison: investigating when and how learning with augmented reality works. *Digital Psychology*, *4*(1S), 1–3. <https://doi.org/10.24989/dp.v4i1S.2419>
- Buchner, J., & Kerres, M. (2023). Media comparison studies dominate comparative research on augmented reality in education. *Computers & Education*, *185*, 104711. <https://doi.org/10.1016/j.compedu.2022.104711>
- Carmigniani, J., & Furht, B. (2011). Augmented reality: an overview. In B. Furht (Ed.), *Handbook of Augmented Reality*. Springer. https://doi.org/10.1007/978-1-4614-0064-6_1

- 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. <https://doi.org/10.1111/1467-8624.00081>
- Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research*, 53(4), 445–459. <https://doi.org/10.3102/00346543053004445>
- Clark, R. E. (1994a). Media will never influence learning. *Educational Technology Research and Development*, 42(2), 21–29. <https://www.jstor.org/stable/30218684>
- Clark, R. E. (1994b). Media and method. *Educational Technology Research and Development*, 42(3), 7–10. <https://www.jstor.org/stable/30220073>
- Clark, R. E. (Ed.). (2012). *Learning with media* (2nd ed). Information Age Publishing.
- Clegg, N. (2023). *Metaverse technologies are creating new opportunities for teachers to inspire students*. Meta. <https://about.fb.com/news/2023/09/metaverse-technologies-education-opportunities/>
- Coban, M., Bolat, Y. I., & Goksu, I. (2022). The potential of immersive virtual reality to enhance learning: A meta-analysis. *Educational Psychology Review*, 36, 100452. <https://doi.org/10.1016/j.edurev.2022.100452>
- Conrad, M., Kablitz, D., & Schumann, S. (2024). Learning effectiveness of immersive virtual reality in education and trainings: A systematic review of findings. *Computers & Education: X Reality*, 4, 100053. <https://doi.org/10.1016/j.cexr.2024.100053>
- Cromley, J. G., Chen, R., & Lawrence, L. (2023). Meta-analysis of STEM learning using virtual reality: Benefits across the board. *Journal of Science Education and Technology*, 32, 255–364. <https://doi.org/10.1007/s10956-023-10032-5>
- Cuban, L. (1993). *Teachers and machines: The classroom use of technology since 1920* (2nd ed). Teachers College Press.
- De La Paz, S. (2007). Managing cognitive demands for writing: Comparing the effects of instructional components in strategy instruction. *Reading & Writing Quarterly*, 23(3), 249–266. <https://doi.org/10.1080/10573560701277609>
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *PNAS*, 116(39), 19251–19257. <https://doi.org/10.1073/pnas.1821936116>
- *Dunnagan, C. L., Dannenberg, D. A., Cuales, M. P., Earnest, A. D., Gurnsey, R. M., & Gallardo-Williams, M. T. (2020). Production and evaluation of a realistic immersive virtual reality organic chemistry laboratory experience: infrared spectroscopy. *Journal of Chemical Education*, 97, 258–262. <https://doi.org/10.1021/acs.jchemed.9b00705>
- Fiorella, L., & Mayer, R. E. (2015). Learning as a generative activity: Eight learning strategies that promote understanding. *Cambridge University Press*. <https://doi.org/10.1017/CBO9781107707085>
- Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, 28(4), 717–741. <https://doi.org/10.1007/s10648-015-9348-9>
- Fortune Business Insights. (2023). *Virtual reality in education market size, share & COVID-19 impact analysis, by component (hardware, software, and content), by application (k-12, higher education, and vocational training), and regional forecast, 2023–2030*. Fortune Business Insights. <https://www.fortunebusinessinsights.com/industry-reports/virtual-reality-in-education-market-101696>
- Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafo, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *PNAS*, 111(23), 8410–8415. <https://doi.org/10.1073/pnas.1319030111>
- Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education*, 8, 1–31. <https://doi.org/10.1007/s40692-020-00169-2>
- Hartikainen, S., Rintala, H., Pylväs, L., & Nokelainen, P. (2019). The concept of active learning and the measurement of learning outcomes: A review of research in engineering higher education. *Education Sciences*, 9(4), 276. <https://doi.org/10.3390/educsci9040276>
- Hastings, N. B., & Tracey, M. W. (2005). Does media affect learning: Where are we now? *TechTrends*, 49(2), 28–30. http://digitalcommons.wayne.edu/coe_aos/15
- Honebein, P. C., & Reigeluth, C. M. (2021). To prove or improve, that is the question: The resurgence of comparative, confounded research between 2010 and 2019. *Educational Technology Research and Development*, 69, 465–496. <https://doi.org/10.1007/s11423-021-09988-1>
- Immersive virtual reality. (2008). *Encyclopedia of multimedia*. Springer. https://doi.org/10.1007/978-0-387-78414-4_85

- Jonassen, D. H., Campbell, J. P., & Davidson, M. E. (1994). Learning with media: restructuring the debate. *Educational Technology Research and Development*, 42(2), 31–39. <https://www.jstor.org/stable/30218685>
- Klahr, D. (2013). What do we mean? On the importance of not abandoning scientific rigor when talking about science education. *Proceedings of the National Academy of Sciences*, 110(3), 14075–14080. <https://doi.org/10.1073/pnas.1212738110>
- Kozma, R. B. (1994a). A reply: Media and methods. *Educational Technology Research and Development*, 42(3), 11–14. <https://doi.org/10.1007/BF02298091>
- Kozma, R. B. (1994b). Will media influence learning? Reframing the debate. *Educational Technology Research and Development*, 42, 7–19. <https://doi.org/10.1007/BF02299087>
- *Lamb, R., Antonenko, P., Etopio, E., & Seccia, A. (2018). Comparison of virtual reality and hands on activities in science education via functional near infrared spectroscopy. *Computers & Education*, 124, 14–26. <https://doi.org/10.1016/j.compedu.2018.05.014>
- Levie, W. H., & Dickie, K. (1973). *The analysis and application of media*. In R. Travers (Ed.), The second handbook of research on teaching. Rand McNally.
- Makransky, G., Borre-Gude, S., & Mayer, R. E. (2019). Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. *Journal of Computer Assisted Learning*, 35(6), 691–707. <https://doi.org/10.1111/jcal.12375>
- Martella, A. M., Klahr, D., & Li, W. (2020). The relative effectiveness of different active learning implementations in teaching elementary students how to design simple experiments. *Journal of Educational Psychology*, 112(8), 1582–1596. <https://doi.org/10.1037/edu0000449>
- Martella, A. M., Martella, R. C., Yacilla, J. K., Newson, A., Shannon, E. N., & Voorhis, C. (2023). How rigorous is active learning research in STEM education? An examination of key internal validity controls in intervention studies. *Educational Psychology Review*, 35, 107. <https://doi.org/10.1007/s10648-023-09826-1>
- Mason, E. N., & Smith, R. A. (2020). Tracking intervention dosage to inform instructional decision making. *Intervention in School and Clinic*, 56(2), 92–98. <https://doi.org/10.1177/1053451220914897>
- Mielke, K. (1968). Questioning the questions of ETV research. *Educational Broadcasting Review*, 2, 6–15.
- Morrison, G. R. (1994). The media effects question: “unresolvable” or asking the right question. *Educational Technology Research & Development*, 42(2), 41–44. <https://www.jstor.org/stable/30218686>
- National Science Foundation. (2020). *STEM education for the future: a visioning report*. <https://www.nsf.gov/edu/Materials/STEM%20Education%20for%20the%20Future%20-%202020%20Visioning%20Report.pdf>
- Palid, O., Cashdollar, S., Deangelo, S., Chu, C., & Bates, M. (2023). Inclusion in practice: A systematic review of diversity-focused STEM programming in the United States. *International Journal of STEM Education*, 10. <https://doi.org/10.1186/s40594-022-00387-3> (Article 2)
- *Parong, J., & Mayer, R. E. (2020). Cognitive and affective processes for learning science in immersive virtual reality. *Journal of Computer Assisted Learning*, 37, 226–241. <https://doi.org/10.1111/jcal.12482>
- *Petersen, G. B., Klingenberg, S., & Makransky, G. (2022). Pipetting in virtual reality can predict real-life pipetting performance. *Technology, Mind, and Behavior*, 3. <https://doi.org/10.1037/tmb0000076>
- Purdue Online. (2024). *The evolution of technology in the classroom*. Purdue University. <https://online.purdue.edu/blog/education/evolution-technology-classroom>
- Reeves, T. C., & Oh, E. G. (2017). The goals and methods of educational technology research over a quarter century (1989–2014). *Educational Technology Research and Development*, 65, 325–339. <https://doi.org/10.1007/s11423-016-9474-1>
- Reiser, R. A. (1994). Clark’s invitation to the dance: as instructional designer’s response. *Educational Technology Research and Development*, 42(2), 45–48. <https://www.jstor.org/stable/30218687>
- Roediger, H. L., III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Saettler, P. (1990). *The evolution of American Educational Technology* (revised ed). Libraries Unlimited.
- Shrock, S. A. (1994). The media influence debate: Read the fine print, but don’t lose sight of the big picture. *Educational Technology Research & Development*, 42(2), 49–53. <https://doi.org/10.1007/BF02299092>

- Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., ... & Young, A. M. (2018). Anatomy of STEM teaching in American universities: A snapshot from a large-scale observation study. *Science*, 359(6383), 1468–1470. <https://doi.org/10.1126/science.aap8892>
- *Su, Y-S., Cheng, H-W., & Lai, C-F. (2022). Study of virtual reality immersive technology enhanced mathematics geometry learning. *Frontiers in Psychology*, 13, 760418. <https://doi.org/10.3389/fpsyg.2022.760418>
- Villena-Taranilla, R., Tirado-Olivares, S., Cozar-Gutierrez, R., & Gonzalez-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>
- Warnick, B. R., & Burbules, N. C. (2007). Media comparison studies: Problems and possibilities. *Teachers College Record*, 109(11), 2483–2510. <https://doi.org/10.1177/016146810710901102>
- 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>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Alyssa P. Lawson^{1,2}  · Amedee Marchand Martella² · Kristen LaBonte² · Cynthia Y. Delgado² · Fangzheng Zhao² · Justin A. Gluck³ · Mitchell E. Munns² · Ashleigh Wells LeRoy⁴ · Richard E. Mayer²

✉ Alyssa P. Lawson
alyssapl Lawson@gmail.com

¹ Landmark College Institute for Research and Training, 19 River Rd S, Putney, VT 05346, USA

² Department of Psychological & Brain Sciences, University of California, Santa Barbara, Santa Barbara, CA, USA

³ Department of Psychology and Neuroscience, University of Colorado, Boulder, Boulder, CO, USA

⁴ Pacific Science & Engineering, San Diego, CA, USA