Meta-Analysis of STEM Learning Using Virtual Reality: Benefits Across the Board

Jennifer G. Cromley[1] · Runzhi Chen[1] · LuEttaMae Lawrence[2]

Abstract

The use of virtual reality (VR) in formal education has burgeoned in recent years, with enthusiastic uptake by teachers and instructors across a wide range of subject areas and academic disciplines. We conducted a systematic meta-analysis of effects of VR on Science, Technology, Engineering, and Mathematics learning from middle school through postsecondary education. Eighteen published journal articles met inclusion criteria, yielding 52 effects from 2214 participants. VR has an overall positive effect on learning of $g = .33$, with the largest significant moderator effects for redesign of VR, classroom settings, science learning, desktop displays, and all types of learning outcomes (factual, conceptual, and transfer). Results depart somewhat from Howard's (2019) and Wu et al.'s (2020b) meta-analyses of VR across learning and treatment, multiple domains, and ages; in our study, desktop VR showed larger effects than head-mounted display, and we found positive effects for all learning outcome types. One trend within studies showing the largest effects is the inclusion of active learning techniques, which may shift learners' focus from interesting but irrelevant details to the most learning-relevant aspects of the VR learning environment.

Keyword Affordances of virtual reality

The use of virtual reality (VR) in formal education has burgeoned in recent years, with enthusiastic uptake by teachers and instructors across a wide range of subject areas and academic disciplines (Parong & Mayer, 2021). One possible reason is that VR has shifted from very expensive hardware plus requiring specialized programming skills to one accessible with any smartphone plus a cardboard viewer (cost from 0$ for DIY up to $30) and free learning environments. Does the research base on learning with virtual reality support this level of use? Does VR actually help learning, and if so, for whom and under what conditions? To what extent have theory-driven constructs such as presence been tested in studies of VR efficacy for learning? Does active learning help learners get more from VR? As part of a larger

meta-analysis of learning with multimedia, we systematically reviewed published VR studies conducted in Science, Technology, Engineering, and Mathematics (STEM) topics with learners from middle school through postsecondary education to try to answer these questions.

Defining VR

VR has been defined in a number of ways, and different ways of defining VR may lead to different conclusions about its effects in STEM education. We have adopted a definition that educational users and designers would recognize. First, VR is a computer-based animated environment allowing learner control to engage in free inspection in every dimension—top to bottom 360° (akin to nodding the head up and down), right to left 360° (akin to turning the head fully clockwise or counter-clockwise), and leaning over to the left and right 360° (i.e., to the point of turning the head upside down). From an engineering psychology perspective, the learning environment is mapped onto ego-referenced frame dimensions of right-left, front-back, and up-down (Wickens et al., 2021). Second, VR learning environments have typically

✉ Jennifer G. Cromley
  jcromley@illinois.edu

1 Department of Educational Psychology, College of Education, University of Illinois at Urbana-Champaign, 1310 S. 6th St. MC-708, Champaign, IL 61820, USA

2 Emma Eccles Jones College of Education and Human Services, Utah State University, 2830 Old Main Hill, Logan, UT 84322-2830, USA

given learners a very high level of control. That is, the learner can choose where to look, what to look at, for how long, driven by their own goals or interests. In addition, some VR learning environments include interactive features such as pop-up information.

The CAMIL Model of Learning with Immersive VR

A recent model of learning with VR highlights the affordances of virtual reality. Makransky and Mayer (2022) and Makransky and Petersen (2021) Cognitive Affective Model of Immersive Learning (CAMIL) relates technological features of VR to its learning affordances, cognitive and motivational mediators, and various learning outcomes. The model explicitly posits that affordances of the medium affect learning and that these affordances interact with learner characteristics (i.e., aptitude x treatment interactions). More specifically, the immersive character of VR (e.g., completely or partially shutting out stimuli from the outside world), the level of learner control, and the realism/refresh rate of the representations (representational fidelity) are the technology features in CAMIL that affect affordances of self-reported presence and the extent of agency (control of one's own actions) that the learner can exercise. These affordances—when VR is designed using multimedia design principles such as the signalling principle and the split attention principle—are then hypothesized to affect learning via a number of motivational (interest, intrinsic motivation, and self-efficacy) and cognitive (perception of embodiment in the virtual environment, working memory load, and regulation of effort) mediators. Finally, these are associated with learning outcomes from Radianti's et al. (2021) VR review-based typology: factual, conceptual (including applying principles), and procedural knowledge, together with transfer of learning to new contexts.

Each path in the CAMIL model is supported by specific research studies; in addition, some specific aspects of the model were tested empirically by Makransky and Mayer (2022) in a classroom-based study with 102 middle school students learning about climate change via a VR lesson on Greenland. Participants were assigned to an HMD VR condition (high-immersion) or an analogous 2D projected video condition (low-immersion), and completed measures of presence, intrinsic motivation, interest, and both immediate post-video and delayed post-lesson measures of factual knowledge. Students in the high-immersion condition significantly outscored those in the low-immersion condition on immediate posttest ($d = .61$) and delayed posttest ($d = .71$). A path analysis testing full mediation of effects of immersion on learning by the presence and motivation fit the data well and supported the claim in the model that VR has its effects on factual knowledge because it increases presence, interest, and intrinsic motivation.

The CAMIL model also describes how the defining characteristics of VR technology-−1080° view and learner control, together with interactivity, make VR similar to simulations (e.g., a simulation of molecular movement in a vessel whose pressure, volume, and/or temperature can be adjusted). As with simulations, many have argued that VR has added benefits for learning content that poses inherent dangers to the learner (e.g., certain reactions in a chemistry laboratory; Makransky et al., 2019a, b) or dangers to others (as in medical VR for learning about processes in vivo); uses costly or fragile, easily damaged materials (Kyriakou & Hermon, 2019); require traveling a great distance and hence expense (Klippel et al., 2019); allow viewing microscopically small (Dunnagan et al., 2020), astronomically large (Huang et al., 2019), or non-visible (e.g., flow of electrical charge; Barata et al., 2015) phenomena, or other obstacles. Thus, VR could allow learning about the Bayeux tapestry without traveling to see it, and without posing any risk to the precious historical artefact. VR could allow viewing flow of electrical conduction in the heart—in a healthy or disordered condition—without risk to a patient.

It is important to note that achievement-focused researchers and instructors might recommend VR only if it leads to *better* learning than other instructional approaches (e.g., classroom lecture, animation), as suggested by the CAMIL model. However, more equity-focused researchers and instructors might recommend VR based on its technological characteristics—e.g., arguments about VR allowing for access to otherwise-inaccessible learning—even if it leads to *equal* learning compared to other instructional approaches.

Previous Meta-analytic and Synthetic Reviews of VR

Howard (2019) meta-analysed the effect of VR on multiple outcomes, from 192 published and unpublished studies through 2014 in formal learning, workplace, psychotherapy, and medical rehabilitation settings. Of these, 84 studies reported on cognitive outcomes (i.e., learning or training such as VR for surgical training). The 84 studies included 9 wait-list control designs ($d = 1.41$) and 75 comparisons to non-VR conditions ($d = .48$), which would both be classified as media comparison studies in our coding. When multiple cognitive effects were reported, Howard averaged these within each study. In comparisons to non-VR conditions, VR had non-significant effects on declarative knowledge ($d = .20$) but significant effects on procedural skills ($d = .59$). For cognitive outcomes, studies using head-mounted (immersive) technology had a larger effect ($b = .68$) on learning than did computer monitor technology. For cognitive outcomes, studies using mouse versus paddle or other input technology did not differ. For cognitive outcomes, using VR in a narrative game-based context had a significant effect on learning ($b = .73$). Results suggest that

the more modern full-immersion VR headsets lead to better learning than from desktops, input paddles vs. mouse do not affect learning, factual learning is not improved though procedural learning is, and games—not included in the present meta-analysis—out-perform VR on average learning outcomes. Year of publication did not affect the size of learning outcomes in Howard's meta-analysis.

To synthesize the effectiveness of HMD-based immersive virtual learning, Wu et al. (2020a, b) conducted a meta-analysis on 35 articles published from 2013 to 2019 involving 1847 learners from kindergarten through adulthood. They included only studies that compared HMD learning conditions with non-HMD learning conditions and yielded a learning outcome as a dependent variable. Of these, only 31.4% of the studies concerned science learning; the remainder concerned learning in medical, special education, and physical activity domains. Wu et al. conducted moderator analyses on educational level, learning domain, region of the study, learning application types (simulation, serious educational game, or representation [all other VR environments]), HMD hardware, testing-format, immediate vs. delayed test, nature of the control group (lecture, real-world practice, or DVR), and learning duration. They found a small but significant effect of HMDs on learning outcome ($g = 0.24$). Regarding moderators, the effect of HMDs was larger for K-12 learners ($g = 0.80$) than post-secondary ($g = -0.02$) and mixed-age learners ($g = 0.68$). They found a medium effect of using HMDs in simulation approaches ($g = 0.45$). VR using HMD had no significant benefit over non-HMD learning for representing content ($g = 0.31$) or for serious educational games ($g = -0.002$). HMDs were more effective than lecture ($g = 0.78$), somewhat more effective than real-world practice ($g = 0.39$) or DVRs ($g = 0.12$). Other moderators were not found to significantly moderate the effect of VR on learning. The larger effect of simulation HMDs over content-representation HMDs (e.g., 3D model, and spherical video) could indicate that the active learning required when using simulations outperforms what could be more passive learning when using representation VR. We return to this point when we describe effects of redesigned VR, which in some cases asked learners to play a more active role while learning with VR.

Another recent review (Suh & Prophet, 2018) included many different outcomes from VR, only 8 of which were learning outcomes, and also included multiple domains, many outside of STEM. Another synthetic review (Merchant et al., 2014) mostly summarized studies of simulations or serious educational games, although a few of those articles are included in this meta-analysis. Makransky and Petersen (2021) conducted a synthetic review of a wide range of VR studies to develop the CAMIL aptitude × treatment model described above.

Rationale for Moderators

Across multiple different types of multimedia (e.g., animation, simulations, games), a number of moderators have been found to significantly influence learning. For example, effect sizes for computer-based learning have decreased over time, with early studies showing larger mean effects possibly due to novelty effects but later studies showing smaller mean effects (e.g., Kulik et al., 1985). Animations benefit middle school and high school students more than undergraduates. The self-explanation strategy helps transfer of learning, but not factual learning with text-and-diagrams. Simulations show positive effects on factual learning in classroom studies, but negative effects in laboratory studies. Concrete illustrations show benefits for transfer when text-and-diagrams is compared to single media (media comparison studies; Meyer et al., 2019), but abstract illustrations show benefits for transfer when redesigned text-and-diagrams is compared to prior-to-redesign text-and-diagrams using active control group (AC) research designs. Based on these differences in effectiveness for various types of multimedia, we coded for and planned to analyse the data for moderation by year of publication, education level of learners, type of dependent variable/learning outcome measured, classroom vs. laboratory context for the study, level of immersion (IVR vs. other VR formats), media comparison research designs (VR vs. non-VR) vs. active control group research designs (redesign VR and compare it to prior-to-redesign VR), and domain (learning content; e.g., mathematics vs. science).

Method

Search Criteria

We searched for quantitative articles on learning STEM topic with VR published in peer-reviewed journals from 2000 to 2020. As noted above, VR was defined as 1080° view with learner control, not using real-life backgrounds (Makransky & Petersen, 2021). The goal of our larger project was also to inform schools and teachers who might use multimedia for STEM learning, so the VR learning environment had to focus on content from Science, Technology, Mathematics, and/or Engineering. Learners had to be in middle school through undergraduate education levels, but students in undergraduate health professions were excluded (e.g., medical, dental, pharmacy). The language skills of our research team led us to select articles published in English. We selected studies using a learning outcome that we could categorize under the CAMIL categories of factual, conceptual, procedural, or transfer (see below). For example, we excluded Coan et al. (2020) because their learning outcome measure mixed factual

and conceptual questions. To enable meta-analysis, sufficient statistics had to be provided for each learning outcome, such as posttest means and standard deviations, or had to be calculable from provided statistics. This meant that when researchers calculated only "gain scores" ($M_{time2} - M_{time1}$; e.g., Makransky et al., 2019a, b), the study had to be excluded from the meta-analysis.

In line with the CAMIL model, we define two types of VR technology studied in this literature: head-mounted displays (high-immersion) and desktop displays (low-immersion). Head-mounted VR displays refer to wearing goggles or a headset to view the virtual environment, which fully blocks the view of the physical surroundings. Desktop VR displays refer to a similar 1080° view displayed on a screen.

Separate from VR, there are augmented reality systems that superimpose one set of visualizations (e.g., the shape of a hawk) over an actual scene (e.g., looking at a bird on the branch of a tree near the learner). In the present study, we did not include AR systems that use the actual scene, but when a manipulable 3D image was overlaid on a blank wall or table, we did categorize this as VR. Likewise, we do not include narrative goal-driven games that happen to use VR technology as VR because of the importance of the narrativity and goal-driven nature of learning in games.
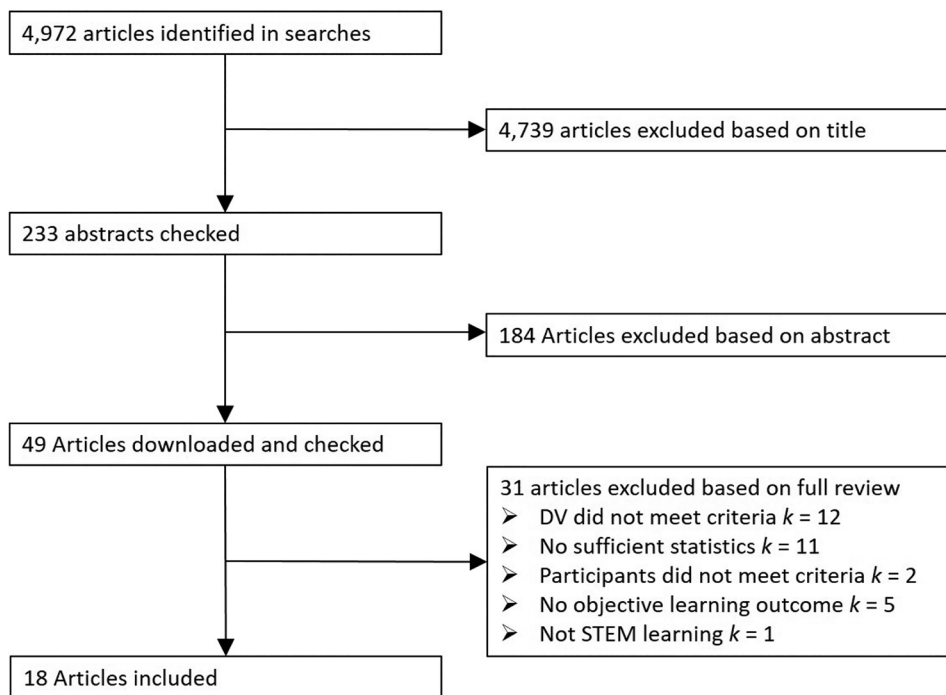
Search Strategy

We searched the ERIC, PsycINFO, and Social Sciences Citation Index databases for published articles using the free text search string ("virtual reality" AND (learn* OR school* OR grade*) NOT (game* OR videogame* OR autis* OR elementary OR teacher*)) and limiting results to 2020 or earlier (see Fig. 1 for the complete flow diagram). Search results were then hand checked to identify candidate articles published in English; the abstracts of these articles were then checked, and articles meeting criteria were printed and screened for learning outcomes. All screening was done twice, once by the second author and once by the first author until 100% agreement was reached.

Coding of Articles

Learning outcome measures were categorized as factual knowledge, conceptual knowledge, procedural knowledge, or transfer, and could use any response format such as multiple choice, open response, and drawn response. All effects were coded from each article, so, for example, if two factual knowledge measures and a conceptual knowledge measure were given at post-test, three separate effects would be coded for the same article. Factual knowledge measures tested verbatim information presented in the VR learning environment, such as definitions or recognition of statements, or calculations using steps exactly as shown in the environment. Conceptual knowledge measures required the learner to apply an instructed principle to a normal situation and/ or to draw conclusions from information presented in the VR environment. Procedural knowledge measures require a learner to put or describe steps in the correct order, such as

Fig. 1 Flow diagram of search and screening procedures

4,972 articles identified in searches

4,739 articles excluded based on title

233 abstracts checked

184 Articles excluded based on abstract

49 Articles downloaded and checked

31 articles excluded based on full review
➢ DV did not meet criteria $k = 12$
➢ No sufficient statistics $k = 11$
➢ Participants did not meet criteria $k = 2$
➢ No objective learning outcome $k = 5$
➢ Not STEM learning $k = 1$

18 Articles included

the sequence of steps in cell mitosis, or the sequence of steps in statistical hypothesis testing. Transfer measures require application of an instructed principal to an unfamiliar and different situation (e.g., blood flow through a heart with a septal wall defect) in order to predict an outcome that differs somewhat from what was instructed.

Effects were coded for 6 potential moderators: whether they were a media comparison research design (a VR condition vs. a non-VR condition) or an active control research design (a redesigned VR condition vs. a prior-to-redesign VR condition), and whether head-mounted display (immersive) or desktop VR was used. In addition, the year of publication, domain (science vs. mathematics), educational level of learners (and whether they were majoring in the topic represented in the VR environment, e.g., biology), and classroom vs. laboratory setting were also used. Additionally, the country where the study took place, APA style reference, and sufficient statistics were entered into an Excel spreadsheet. Although we were interested in whether sense of presence would moderate effects, too few studies measured this variable to allow us to test for it as a moderator. In addition, too few studies were conducted with students younger than postsecondary to test for it as a moderator. All coding was done twice, once by the second author and once by the first author until 100% agreement was reached.

Other than reverse scoring an accuracy outcome that was reported in degrees of angle (small angle is higher accuracy; Stull et al., 2009), we did not need to conduct any deletions, transformations, imputation, or other preparation of the data for analysis.

Analyses

We analysed Hedges' g effect sizes on posttest control and treatment scores using the R meta-analysis package *metafor* (Viechtbauer, 2010), which weights effect sizes by sample size. The rma.mv command was used to produce robust estimates, since multiple effects can be nested within each study (specifying study ID as the random effect). We used the Q statistic to assess heterogeneity of effect sizes; a significant QE statistic suggests heterogeneity across studies such that a moderator might be able to explain the variance in effect sizes. We conducted single-moderator analyses; *metafor* reports the QM test of the set of moderators, where a significant QM test suggests that the hypothesized moderator explains a significant amount of variance in the effect sizes. Analyses for categorical moderators used a no-intercept model. Screening for publication bias used the *funnel* command in *metafor* and the fail-safe N. Two-tailed tests and an alpha level of 0.05 were used for all analyses.

Results

Descriptives on Studies

Descriptive statistics on the set of 18 studies based on 2214 learners reporting 52 effects are shown in supplementary Table 1.

Checking for Publication Bias

Only very mild asymmetry in the funnel plot (Fig. 2) suggests little evidence of publication bias, and the fail-safe $N = 2095$ suggests that a very large number of studies with $g = 0$ would have to be added to make the overall effect of VR on learning non-significant.

Overall Effect of VR on Learning

The overall effect of VR on learning—averaging across both types of comparisons (MC and AC) and the three learning outcomes found—is a small but significant $g = 0.33$ ($p < 0.001$, QE [51] = 143.66, $p < 0.001$). The large QE statistic (see Fig. 3) reflects the wide range of effects seen in these published articles, which range from.

Moderator Analyses

A series of single-moderator analyses was conducted for year of publication, media comparison research design (VR condition vs. non-VR condition) vs. active control group research design (redesigned VR vs. prior-to-redesign VR), type of learning outcome (factual, conceptual, or transfer
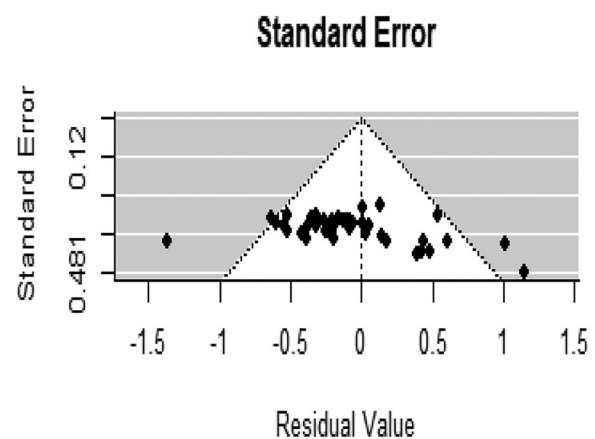
Fig. 2 Funnel plot

Fig. 3 Forest plot of results. Effects are sorted from smallest to largest within each DV/study research design combination; the dashed line marks $g = 0$; squares symbolize the estimated effect; horizontal lines symbolize the CI 95 around each estimated effect. DV/research study design combinations use the following abbreviations: F, factual; C, conceptual; T, transfer; AC, active control group research design, MC, media comparison group research design. VR harming learning ($g = -1.1$) to helping learning ($g = +1.5$). The significant heterogeneity warranted testing for moderators
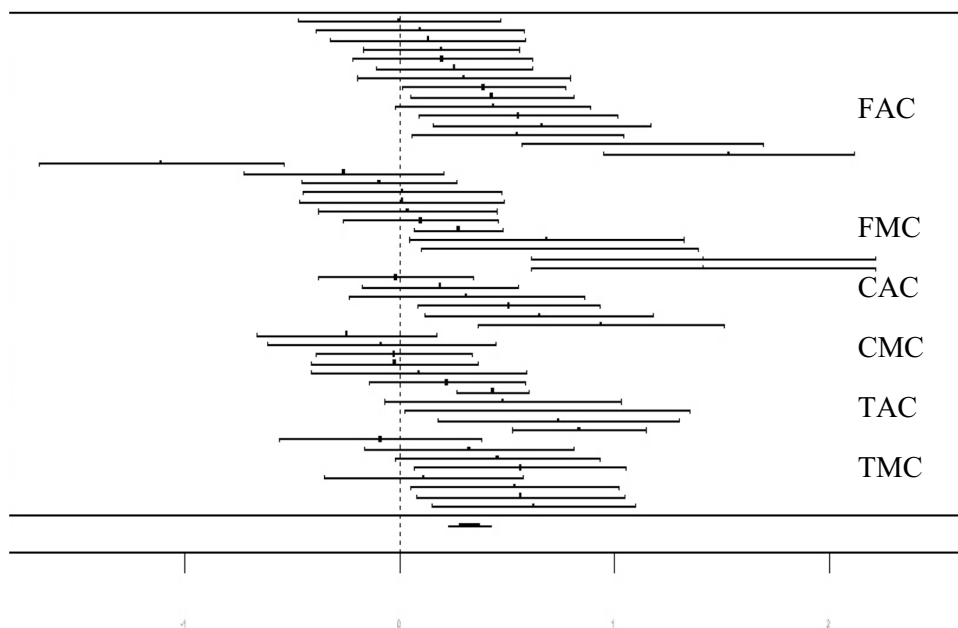
[procedural measures were not found in this set of studies]), classroom vs. laboratory setting, high-immersive VR vs. low-immersive VR, science vs. mathematics content, and US sample vs. non-US sample. All univariate moderator analyses explained a significant amount of variance in effects (see supplementary Table 2), except for the analysis of year of publication. Media comparison designs showed a mean effect of $g = 0.32$ and active control research designs yielded a significant $g = .49$. Clearly, redesign of VR produces better learning results than prior-to-redesign VR, and VR is indeed better for learning than non-VR comparison conditions. Effects are significant on factual knowledge learning outcomes ($g = 0.38$), on conceptual knowledge outcomes ($g = 0.37$), and on transfer outcomes ($g = 0.59$). Both laboratory ($g = 0.39$) and classroom ($g = 0.39$) studies yield significant benefits for learning. Effects are significant for both low-immersive ($g = 0.53$) and high-immersive ($g = .30$) systems. Effects are significant for science content matter ($g = 0.38$) but not for mathematics content. Effects are significant for studies conducted in the USA ($g = 0.47$) and countries other than the USA ($g = 0.36$). Thus, the typical multimedia study characteristics that are associated with larger or smaller effects do indeed explain the variability in VR effects. Interestingly, the advantage of HMDs found by Howard (2019) and Wu et al. (2020a, b) across multiple domains and outcomes was not found here, where only STEM learning was included.

Discussion

The overall effect of VR on STEM learning, taking into account no moderators, was a significant $g = 0.33$, with significant variability in effect sizes. VR can help STEM learning in the middle school through postsecondary age range, albeit to a modest extent. From the moderator analyses, we can see that in media comparison research designs, VR produces better learning than non-VR. Since the technological aspects of VR can also allow access to learning spaces that are inaccessible due to potential for danger (Barata et al., 2015), potential for harm (Zinchenko et al., 2020), or expense (Petersen et al., 2020), this argues for using VR despite the small effect size because VR permits access, and especially equitable access. On equity grounds, even in cases where VR does not perform better on average than a lecture, a PowerPoint, or traveling to an archaeological dig (Shackelford et al., 2019); it merely needs to perform as well as those other instructional delivery methods to warrant its use.

Redesign of VR yields significantly better learning with a large effect size ($g = 0.49$). This suggests that redesign can help learners take advantage of the affordances of VR, thereby supporting the CAMIL model. The redesigns were based on a very wide range of theories, however, and cannot be described as a coherent set of learning supports. In some cases, the redesign of VR comprised adding active learning components, such as summarizing (Parong & Mayer, 2018, Experiment 2),

explaining to a peer (Klingenberg et al., 2020), or planning out a series of moves in the environment (Wu et al., 2020a, b). In one media comparison study (Barata et al., 2015), it appears that the VR condition added active learning by requiring application of abstract principles to specific realistic scenarios. These interventions shifted learners from passive recipients of the VR display, to ones who actively gathered information, similar to passively vs. actively watching an instructional video or passively vs. actively visiting a museum exhibit (Bamberger & Tal, 2008). In one study (Stull et al., 2009), the positive effect of redesign came from adding a spatial cue (cueing principle: orientation reference) in a highly spatial task. In another case, the positive effect of redesign emerged from the CAMIL principle of high-immersion vs. low-immersion VR on a transfer task (Makransky et al., 2019a, b). In other cases, the positive effect of redesign emerged from viewing a female avatar or one matched to the learner's own sex (Makransky et al., 2019a, b) or a *non*-sexist instructor (compared to a sexist instructor) in the VR environment (Chang et al., 2019), all of which suggest a qualitatively different mechanism for the positive effects found in studies that tested redesign of VR compared to prior-to-redesign VR.

A different redesign mechanism may be in play in Klippel et al. (2019), where the success of the VR field trip over an actual field trip could be due to many possible factors, including less distraction and fatigue when staying on campus (vs. riding in a bus with other undergraduates) or replayable guided narration in VR vs. no replay in an actual field trip. Yet another redesign mechanism may be in play, in Kim et al. (2019) and Zinchenko et al. (2020) where the advantage of VR over PowerPoint was found among non-STEM undergraduates (humanities and graphic design students, respectively) completing memorization tasks; perhaps novelty effects explain these findings for these participants on this task. Zinchenko et al. (2020) found the greatest effects of VR on factual learning from those who began with the lowest prior knowledge, i.e., an expertise reversal effect (Kalyuga, 2014) for VR.

The strongest trend among the set of positive VR redesign effects is for active learning—learners should be given a specific, constructive task while learning with VR and/or told they will have to answer difficult transfer questions after learning. Learning tasks should require them to transform the information in the VR learning environment, such as by summarizing during learning, giving an explanation to a peer during or after learning, planning out learning ahead of time, or applying what was just learned in an abstract sense to specific concrete situations. This is highly consistent with Mayer's (1996) select, organize, integrate model because asking learners to engage in a constructive task should force them to choose what the most important information and details are and attend to them in the virtual environment (Select), and it should also force them to connect (Organize)

what they have selected in order to create a coherent summary or explanation. Integration is best assessed with measures of transfer, and we found substantial and significant average transfer effects ($g = 0.59$) for VR.

Researchers have argued that the promise of VR rests on its visual affordances—the 1080 view (top to bottom, right to left, tilt left down [upside down] to right down) which makes all aspects of a phenomenon inspectable—and on learner control that maps naturally to the body, and perhaps on motivational benefits from these visual affordances and learner control. The results from this meta-analysis of 18 studies based on 2214 learners which report 52 effects on learning suggest that the visual affordances, learner control, and any motivational benefits by themselves are indeed enough to yield better learning, but some redesigns of the learning environment yield a large amount more. That is, the unique features of VR do add some extra benefits beyond non-VR, but much more is to be gained when learners are actively engaged in learning.

One possible explanation for this finding is that the "wow" factor of VR is itself distracting—perhaps without a constructive task, the learner wanders in the environment looking at fascinating or beautiful but learning-irrelevant features. For example, in Zinchenko and colleagues' (2020) study using a circulatory system VR environment, perhaps learners were distracted by comparing the mitral valve with its tendon cords to the aortic valve which has a similar function but very different appearance. In this way, perhaps the 1080° view provides many more 'seductive details' to learners—that is, factually accurate details but ones which are irrelevant to the major learning goals (Park et al., 2015).

Research on learning with VR is still in its early stages, so perhaps the effects on factual learning (Howard, 2019)—recognizing things that were explicitly taught in the VR learning environment—are not surprising. The finding of the strongest trend in transfer learning for modified VR indicates the importance of redesign of the VR learning process. Active learning conditions, a form of redesign, focused on combining presented material (e.g., via a summary) so perhaps this prompts drawing conclusions from presented material and/or prior knowledge, and perhaps it facilitates generalizing to new situations (transfer). It may also prevent learners from being distracting by seductive details in VR environment that are irrelevant for learning. A different mechanism for the effect of redesign might be at play when female avatars or sex-matched avatars help middle school students transfer their knowledge from the learning environment to new problem settings; the goal of these avatar manipulations is to reduce stereotype threat, which is known to put pressure on working memory and thereby affect performance (Beilock et al., 2007).

The finding of equal effects in classroom settings as in Laboratory settings (both $g = .39$) is different from findings

with text + diagram multimedia. In classroom settings in these VR studies, students were majoring in the subject that was the focus of the course and the focus of the VR (e.g., engineering majors in a power supply course learning about high-voltage transformer stations in Barata et al., 2015). In the more-highly controlled and less distracting Laboratory settings, students were from university subject pools (e.g., psychology, education).

The finding of larger effects from desktop VR than HMDs might also be explained by this "wow" factor, or perhaps by the disorientation or discomfort some participants report with HMDs (e.g., "simulation sickness" or cybersickness; Weech et al., 2019). Perhaps the familiar 3D image on screen with joystick controls helps learners take more advantage of the learning environment, without either the need to learn how to move in the environment using the immersive HMD or any disorientation that may happen.

The slightly larger effect for studies conducted in the USA might be due to the larger proportion of US studies testing redesign of VR (75% of effects) compared to non-US studies (40% of effects testing redesign of VR), given the finding that redesign of VR have larger effects than media comparison studies.

Limitations

The VR literature in STEM education at these education levels is still relatively small, and there were few studies with students younger than postsecondary education (but see Villena-Taranilla, et al., 2022). One reason may be health warnings on VR equipment restricting its use to older children. A variable related to this is *cybersickness*, not measured in any of these studies. Motivational variables were hardly measured in these studies (intrinsic motivation in 3 studies, self-efficacy in 6, expectancy in 1). An though *presence* is prominent in the gaming literature, only 6 studies measured it. Therefore, education level, cybersickness, motivation, nor presence could be tested as a moderator. Although the fail-safe *N* and funnel plot supported the robustness of our findings, more research in the years to come will likely show more nuances in what works in VR, and for whom, in STEM learning.

Implications for Research

Future research could more carefully attend to the types and variety of learning outcomes that are measured when VR is used for STEM learning. Other reviews have noted the lack of

theoretical models driving the design of VR learning environments; our results which are supportive of the CAMIL model suggest that future studies should measure and test the multiple mediators in that model. More studies are needed testing the most effective redesign principles found in the broad literature on multimedia learning (e.g., cueing).

Implications for Practice

Given the small number of studies, we tentatively suggest that instructors give learners specific, active tasks while learning with VR such as writing a summary or giving an explanation to a peer, if the VR environment does not provide these. VR can be effective across a range of learning outcomes, but might be especially effective for transfer of learning, which is difficult for teachers to obtain. Finally, instructors might consider user lower-cost desktop systems that students already know how to operate.

Adding active learning to VR can build on the large literature on strategy instruction; strategy instruction generally appears to show larger effects on learning when the usefulness of the strategy is explained, then modelled, practiced with feedback, and performance is attributed to use of the strategy (Dinsmore et al., 2020; Pressley & Harris, 2006). The instructed strategies need to be relevant to the learning objectives and during-learning task assigned to the learner (e.g., learn well enough to take a test vs. learn well enough to teach to a peer), and this instructional alignment characterizes high-quality educational technology research more broadly. Adopting such an evidence-based approach to active learning in VR should yield even stronger effects that those documented here.

Data Availability All data are publicly available in the published journal articles that were meta-analysed; all analysed data are in Supplementary Table 1 of the manuscript.

Declarations

Ethics Approval Not applicable to a literature review; published studies comprise the data analysed in this manuscript.

Consent Statement Not applicable.

Conflict of Interest The authors declare no competing interests.

Disclaimer The sponsor played no role in the design of the research, the decision to publish, or the content of this submission.

References

*Indicates a study included in the meta-analysis

Bamberger, Y., & Tal, T. (2008). Multiple outcomes of class visits to natural history museums: The students' view. *Journal of Science Education and Technology, 17*(3), 274–284.

*Barata, P. N. A., Ribeiro Filho, M., & Nunes, M. V. A. (2015). Consolidating learning in power systems: Virtual reality applied to the study of the operation of electric power transformers. *IEEE Transactions on Education, 58*(4), 255–261.

Beilock, S. L., Rydell, R. J., & McConnell, A. R. (2007). Stereotype threat and working memory: Mechanisms, alleviation, and spillover. *Journal of Experimental Psychology: General, 136*(2), 256.

*Cai, S., Wang, X., & Chiang, F. K. (2014). A case study of augmented reality simulation system application in a chemistry course. *Computers in Human Behavior, 37*, 31–40.

*Chang, F., Luo, M., Walton, G., Aguilar, L., & Bailenson, J. (2019). Stereotype threat in virtual learning environments: Effects of avatar gender and sexist behavior on women's math learning outcomes. *Cyberpsychology, Behavior, and Social Networking, 22*(10), 634–640.

Coan, H. A., Goehle, G., & Youker, R. T. (2020). Teaching biochemistry and molecular biology with virtual reality-lesson creation and student response. *Journal of Teaching and Learning, 14*(1), 71–92.

Dinsmore, D. L., Fryer, L. K., & Parkinson, M. M. (Eds.). (2020). *Handbook of strategies and strategic processing*. Routledge.

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.

*Habig, S. (2020). Who can benefit from augmented reality in chemistry? Sex differences in solving stereochemistry problems using augmented reality. *British Journal of Educational Technology, 51*(3), 629–644.

Howard, M. C. (2019). Virtual reality interventions for personal development: A meta-analysis of hardware and software. *Human-Computer Interaction, 34*(3), 205–239.

*Huang, K. T., Ball, C., Francis, J., Ratan, R., Boumis, J., & Fordham, J. (2019). Augmented versus virtual reality in education: An exploratory study examining science knowledge retention when using augmented reality/virtual reality mobile applications. *Cyberpsychology, Behavior, and Social Networking, 22*(2), 105–110.

*Jong, M. S. Y., Tsai, C. C., Xie, H., & Kwan-Kit Wong, F. (2020). Integrating interactive learner-immersed video-based virtual reality into learning and teaching of physical geography. *British Journal of Educational Technology, 51*(6), 2064–2079.

Kalyuga, S. (2014). The expertise reversal principle in multimedia learning. In R.E. Mayer (Ed.). *Cambridge Handbook of Multimedia Learning.* Cambridge University Press.

*Kim, S., Heo, R., Chung, Y., Kim, J. M., Kwon, M. P., Seo, S. C., ... & Kim, M. K. (2019). Virtual reality visualization model (VRVM) of the tricarboxylic acid (TCA) cycle of carbohydrate metabolism for medical biochemistry education. *Journal of Science Education and Technology, 28*(6), 602–612.

*Klingenberg, S., Jørgensen, M. L., Dandanell, G., Skriver, K., Mottelson, A., & Makransky, G. (2020). Investigating the effect of teaching as a generative learning strategy when learning through desktop and immersive VR: A media and methods experiment. *British Journal of Educational Technology, 51*(6), 2115–2138.

*Klippel, A., Zhao, J., Jackson, K. L., La Femina, P., Stubbs, C., Wetzel, R., ... & Oprean, D. (2019). Transforming earth science education through immersive experiences: Delivering on a long held promise. *Journal of Educational Computing Research, 57*(7), 1745–1771.

Kulik, J. A., Kulik, C. L. C., & Bangert-Drowns, R. L. (1985). Effectiveness of computer-based education in elementary schools. *Computers in Human Behavior, 1*(1), 59–74.

Kyriakou, P., & Hermon, S. (2019). Can I touch this? Using natural interaction in a museum augmented reality system. *Digital Applications in Archaeology and Cultural Heritage, 12*, e00088.

*Lee, E. A., & Wong, K. W. (2014). Learning with desktop virtual reality: Low spatial ability learners are more positively affected. *Computers & Education, 79*, 49–58.

Makransky, G., & Mayer, R. E. (2022). Benefits of taking a virtual field trip in immersive virtual reality: Evidence for the immersion principle in multimedia learning. *Educational Psychology Review, 34*(1), 71–92.

Makransky, G., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review, 33*(3), 937–958. https://doi.org/10.1007/s10648-020-09586-2

*Makransky, G., Borre-Gude, S., & Mayer, R. E. (2019a). Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. *Journal of Computer Assisted Learning, 39*(6), 691–707.

*Makransky, G., Wismer, P., & Mayer, R. E. (2019b). A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation. *Journal of Computer Assisted Learning, 35*(3), 349–358.

Mayer, R. E. (1996). Learning strategies for making sense out of expository text: The SOI model for guiding three cognitive processes in knowledge construction. *Educational Psychology Review, 8*(4), 357–371.

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

*Merchant, Z., Goetz, E. T., Keeney-Kennicutt, W., Kwok, O. M., Cifuentes, L., & Davis, T. J. (2012). The learner characteristics, features of desktop 3D virtual reality environments, and college chemistry instruction: A structural equation modeling analysis. *Computers & Education, 59*(2), 551–568.

*Meyer, O. A., Omdahl, M. K., & Makransky, G. (2019). Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment. *Computers & Education, 140*, 103603.

Park, B., Korbach, A., & Brünken, R. (2015). Do learner characteristics moderate the seductive-details-effect? A cognitive-load-study using eye-tracking. *Journal of Educational Technology & Society, 18*(4), 24–36.

*Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology, 110*(6), 785.

Parong, J., & Mayer, R. E. (2021). Cognitive and affective processes for learning science in immersive virtual reality. *Journal of Computer Assisted Learning, 37*(1), 226–241.

*Petersen, G. B., Klingenberg, S., Mayer, R. E., & Makransky, G. (2020). The virtual field trip: Investigating how to optimize immersive virtual learning in climate change education. *British Journal of Educational Technology, 51*(6), 2099–2115.

Pressley, M., & Harris, K. R. (2006). Cognitive strategies instruction: From basic research to classroom instruction. In P. A. Alexander & P. H. Winne (Eds). *Handbook of Educational Psychology.* Routledge.

Radianti, J., Majchrzak, T. A., Fromm, J., Stieglitz, S., & Vom Brocke, J. (2021). Virtual reality applications for higher educations: A market analysis. *Hawaii International Conference on System Sciences 2021.* (HICSS-54).

Shackelford, L., Huang, D. W., Craig, A., Merrill, C., & Chen, D. (2019). Relationships between motivational support and game features in a game-based virtual reality learning environment for

teaching introductory archaeology. *Educational Media International, 56*(3), 183–200.

*Stull, A. T., Hegarty, M., & Mayer, R. E. (2009). Getting a handle on learning anatomy with interactive three-dimensional graphics. *Journal of Educational Psychology, 101*(4), 803–816.

Suh, A., & Prophet, J. (2018). The state of immersive technology research: A literature analysis. *Computers in Human Behavior, 86*, 77–90.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.

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

Weech, S., Kenny, S., & Barnett-Cowan, M. (2019). Presence and cybersickness in virtual reality are negatively related: A review. *Frontiers in Psychology, 10*, 158.

Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2021). *Engineering Psychology and Human Performance (*5th Ed). Routledge.

*Wu, B., Hu, Y., & Wang, M. (2020a). How do head-mounted displays and planning strategy influence problem-solving-based learning in introductory electrical circuit design? *Educational Technology & Society, 23*(3), 40–52.

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

*Zinchenko, Y. P., Khoroshikh, P. P., Sergievich, A. A., Smirnov, A. S., Tumyalis, A. V., Kovalev, A. I., ... & Golokhvast, K. S. (2020). Virtual reality is more efficient in learning human heart anatomy especially for subjects with low baseline knowledge. *New Ideas in Psychology, 59*, 100786.

*Zhao, J., Lin, L., Sun, J., & Liao, Y. (2020). Using the summarizing strategy to engage learners: Empirical evidence in an immersive virtual reality environment. *The Asia-Pacific Education Researcher.* 1–10.