British Journal of Educational Technology *Vol 51 No 6 2020* 1991–2005
doi:10.1111/bjet.13023

Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis

Bian Wu , Xiaoxue Yu and Xiaoqing Gu

Bian Wu is an associate professor at the Department of Education Information Technology of East China Normal University. His research focuses on technology-enhanced learning, STEM education and complex problem-solving learning. Xiaoxue Yu is a master student at the Department of Education Information Technology of East China Normal University. Her research interests include learning analytics and VR in education. Xiaoqing Gu is a full professor at the Department of Education Information Technology at East China Normal University. Her main research interests include learning design, CSCL and learning analytics. Address for correspondence: Xiaoqing Gu, Department of Education Information Technology, East China Normal University, 3663 North Zhongshan Road, Shanghai, 200062, China. Email: xqgu@ses.ecnu.edu.cn

Abstract

With the availability of low-cost high-quality head-mounted displays (HMDs) since 2013, there is a growing body of literature investigating the impact of immersive virtual reality (IVR) technology on education. This meta-analysis aims to synthesize the findings on the overall effects of IVR using HMDs compared to less immersive desktop virtual reality (DVR) and other traditional means of instruction. A systematic search was carried out on the literature published between 2013 and 2019. Thirty-five randomized controlled trials (RCTs) or quasi-experimental studies were identified. We conducted an analysis using the random effects model (REM) to calculate the pooled effect size. The studies were also coded to examine the moderating effects of their characteristics, such as learner stage, learning domain, learning application type, testing format, control group treatment and learning duration, on the outcome measure. The results showed that IVR using HMDs is more effective than non-immersive learning approaches with a small effect size ($ES = 0.24$). The key findings of the moderator analysis were that HMDs have a greater impact (a) on K-12 learners; (b) in the fields of science education and specific abilities development; (c) when offering simulation or virtual world representations; and (d) when compared with lectures or real-world practices. The meta-analysis also suggested that HMDs can improve both knowledge and skill development, and maintain the learning effect over time.

Keywords: Immersive virtual reality, Head-mounted display, Learning performance, Meta-analysis

Introduction

Virtual reality (VR) technology has been widely used to create situated and realistic learning contexts that learners cannot easily access. For example, micro or macro phenomena, such as blood cells or the solar system, are difficult to observe in the physical world. Skills training like chemical experiments or clinical surgery may suffer from safety issues. Exploring environmental issues, or solving engineering problems that involve higher-order competence development, require the

Practitioner Notes

What is already known about this topic

- Head-mounted displays (HMDs) have been widely applied in various disciplines across both K-12 and post-secondary education.
- HMDs have a positive impact on learning attitudes and perceptions.
- HMDs have produced mixed results on learning performance.

What this paper adds

- Immersive virtual reality (IVR) using HMDs is more effective than non-immersive learning approaches with a small effect size.
- The critical factors of learning implementation and research design moderate the impact of HMDs on learning performance.
- HMDs can improve both knowledge and skill development and maintain the learning effect over time.

Implications for practice and/or policy

- HMD-based immersive learning appears to be a better complement to non-immersive learning approaches.
- Theory-driven learning design should be incorporated to guide HMD-based teaching and learning practice.

integration of multisensory channels and high intensity of interactions during learning. The use of VR research has justified its benefits for education and training.

However, with advancements in VR technology, head-mounted displays (HMDs) offer a novel virtual learning experience. Compared with traditional desktop VR (DVR), in which the virtual world is seen from outside, HMDs provide personal viewing inside VR, thereby offering a sense of immersion during learning. Thus, VR is now commonly divided into two categories: immersive VR (IVR) and non-immersive VR. Although non-immersive VR has been thoroughly investigated (Merchant, Goetz, Cifuentes, Keeney-Kennicutt, & Davis, 2014), understanding its impact cannot be easily generalized to HMD-based IVR.

Despite growing interest in HMD-based immersive learning, and numerous studies suggesting positive attitudes towards and perceptions of HMD (Han, 2020; Makransky, Terkildsen, & Mayer, 2019), synthesized empirical evidence of its impact on learning performance is still lacking. Besides, immersive learning through HMD is not simply a yes-no question. Rather, rigorous analysis of the various conditions that can moderate its impact is required to fully understand the design and implementation of the HMD learning environment in practice.

Echoing the call that future studies should focus on learning outcomes instead of user experience (Radianti, Majchrzak, Fromm, & Wohlgenannt, 2020), this study aims to synthesize previous research on learning performance in HMD-based immersive learning environments. Further, it conducts a moderator analysis to shed light on the inconsistent results. Two research questions are proposed:

RQ1: What is the overall learning effect of head-mounted displays (HMDs) for immersive learning compared to non-immersive learning?

RQ2: How do studies' characteristics, such as learner stage, learning domain, publication region, HMD-based learning application type, HMD hardware, testing format, long-term impact, control group treatment and learning duration, moderate the effect?

Literature review

Immersive learning using head-mounted displays

The notion of "immersive learning" refers to the cognitive states that affect learning during activities within virtual learning environments (Scoresby & Shelton, 2011). Dede (2009) has indicated the immersive learning experience, for example, suspension of disbelief that one can travel inside a blood cell is due to the combination of actional, symbolic and sensory factors of a virtual learning environment. Regarding theoretical alignment of VR affordances, previous studies (Dede, 2009; Johnson-Glenberg, 2018; Parong & Mayer, 2018) have emphasized that, through simulation of the real world, VR technology can couple with different learning theories including multimedia learning, situated learning and embodied learning to enhance learner performance. With the rapid advancement of VR technology and emergence of the HMD device, HMD-based immersive learning has received growing attention in terms of developing various educational applications and investigating the impacts of this novel immersive media on multidimensional learning outcomes in different subjects for different stages of learners.

HMDs extend the affordances of traditional VR in terms of visualization and interaction by tracking head movements to display corresponding 3D scenes in front of one's gaze and enabling natural manipulation in virtual spaces (Han, 2020). The enhanced affordances of HMDs can be aligned with three different learning application types: virtual world representation, simulation and serious games (Merchant *et al.*, 2014). First, HMDs provide virtual world representations, such as 3D models (Weyhe, Uslar, Weyhe, Kaluschke, & Zachmann, 2018), spherical videos (Chien, Hwang, & Jong, 2020) and virtual field trips (Han, 2020), that are beyond the traditional visual stimuli in 2D or DVR. These representations overcome the temporal and spatial limits of sensory perceptions to deliver a surrounding environment for concrete learning (Hwang & Hu, 2013). Second, traditional 2D simulation can be enhanced with IVR through variations in human-computer interactions. Barricelli, Gadia, Rizzi, and Marini (2016) found that the semiotic representations and more naturalistic ways of interacting in HMDs and WIMP (window, icon, menu and pointer) in 2D simulations may have different effects on the learning experience. Further, previous research has revealed that simulation using HMDs improves knowledge transfer to solve real tasks (Falloon, 2020; Ganier, Hoareau, & Tisseau, 2014). Some studies (Civelek, Ucar, Ustunel, & Aydin, 2014; Jung & Ahn, 2018) have also combined haptic interfaces, such as wearable sensors, with an HMD to offer an embodied learning experience and improve manipulation abilities during simulation. Third, the integration of serious games with HMD settings can offer role-play opportunities in more vivid scenarios to facilitate a higher degree of engagement and create more "realistic feelings" such as empathy or the sense of a professional identity (Feng, González, Amor, Lovreglio, & Cabrera-Guerrero, 2018). The situated experience in turn improves both knowledge understanding and the development of higher-order competence (Chittaro & Buttussi, 2015).

Researchers generally acknowledge that immersion is a vital element in contributing to psychological factors including a higher sense of presence, more engagement, positive attitude towards learning subjects and better learning perceptions (Buttussi, & Chittaro, 2018; Civelek *et al.*, 2014; Han, 2020). However, participants of HMD-based immersive learning also experienced higher cognitive load and motion sickness than those of non-immersive learning (Makransky, Terkildsen, *et al.*, 2019; Srivastava, Rimzhim, Vijay, Singh, & Chandra, 2019). Although consistent findings were obtained regarding the influence on psychological factors, previous empirical studies of HMDs have resulted in mixed learning outcomes (Parmar *et al.*, 2016; Srivastava *et al.*, 2019). Some researchers (Makransky, Terkildsen, *et al.*, 2019) have argued that HMDs only have a significant influence on learning perception but not learning performance. Others, for example, Jensen and Konradsen (2018), have found that HMDs have more influence on skills training than knowledge learning or the development of higher-order competence. Besides the need to investigate multidimensional learning outcomes, the abovementioned learning application types in HMD settings must be compared with non-immersive learning to provide evidence of their effectiveness (Kozhevnikov, Gurlitt, & Kozhevnikov, 2013). Numerous studies have investigated the impact of HMDs on learning in comparison with various conventional instruction conditions, such as lectures, PowerPoint slides, e-textbooks, hands-on practice and DVRs (Leder, Horlitz, Puschmann, Wittstock, & Schütz, 2019; Meyer, Omdahl, & Makransky, 2019; Parmar *et al.*, 2016). Nonetheless, their empirical findings have left open questions on the conditions under and extent to which HMDs can lead to better learning performance.

Previous reviews

A search of the literature resulted in one meta-analysis of desktop VR-based learning (Merchant *et al.*, 2014) and five systematic reviews IVR-related learning (Concannon, Esmail, & Roduta Roberts, 2019; Feng *et al.*, 2018; Jensen & Konradsen, 2018; Radianti *et al.*, 2020; Suh & Prophet, 2018). Suh and Prophet (2018) proposed a stimulus-organism-response framework to review research on the antecedents and consequences of immersive technology use. They argued that the technological affordances of immersive technology should be investigated more to advance our knowledge of the mechanisms that explain user performance in immersive environments. It had been unclear how and in what ways immersive technology affects user performance. Besides, their literature review focused not only on IVR but also on augmented virtual reality (AR) and mixed reality (MR), and covered studies not specifically within the education field.

Merchant *et al.* (2014) conducted a meta-analysis of VR-based instruction. However, their work focused on the impact of traditional DVR. Due to the difference between DVR and HMDs in terms of visual stimuli and interactivity, which in turn leads to different levels of immersion, their findings might not easily be generalized to HMD settings. After Merchant *et al.*, several systematic reviews specifically focused on HMD technology in education (Concannon *et al.*, 2019; Feng *et al.*, 2018; Jensen & Konradsen, 2018; Radianti *et al.*, 2020). However, these reviews primarily emphasized the use of HMD for skills training or post-secondary education (Concannon *et al.*, 2019; Jensen & Konradsen, 2018; Radianti *et al.*, 2020). There has been a dearth of studies systematically investigating HMD-based learning in education and training settings and focused on improving knowledge understanding, skills development and learning transfer in various disciplines. In addition, there has been a lack of meta-analyses synthesizing the empirical findings on HMD-based immersive learning, especially comparing HMD-based immersive learning and DVR-based learning or other instructional approaches. This has hindered understanding of the impact of HMDs on learning performance, the gap of which aims to be filled by the study.

Method

Study search

Studies published from January 2013 to December 2019 were searched from the Educational Resources Information Clearinghouse (ERIC), ISI Web of Science and Google Scholar databases. We used the ISI Web of Science to retrieve journal articles in the Social Science Citation Index. ERIC was included because of the VR topic in the education field. Google Scholar was used to retrieve all the available literature including articles in press. The keywords were “virtual reality” or “head-mounted display” and “learning” or “training.” An initial search based on the keywords yielded 3967 articles.

Criteria for inclusion and exclusion

Studies meeting the following criteria were included: (a) peer-reviewed journal article; (b) randomized controlled trial (RCT) or quasi-experimental study; (c) compared head-mounted display (HMD) condition with non-HMD conditions; (d) used test or task performance as a dependent variable; (e) was written in English. The excluded criteria were: (a) rehabilitation studies; (b) single group non-comparative studies; (c) both research and control groups included the use of HMDs; (d) immersive virtual reality technology was Cave Automatic Virtual Environment (CAVE), AR or MR rather than HMDs in the experimental groups; (e) dependent variables included only self-reported data; (f) presented data did not allow calculation of effect size.

Scanning of the titles and abstracts was effective to identify unrelated studies, such as VR learning using not HMDs, in the field of rehabilitation, non-experimental studies or focusing on learner experience but without measurement of learning performance, which resulted in 239 articles for further consideration. Subsequently, the first two authors read each full-text article to assess its eligibility for meta-analysis. Finally, a total of 26 articles met all of the inclusion and exclusion criteria.

For the selected articles with multiple measures of one outcome or more than one experimental or control condition, the following rules were applied to the calculation of effect size (*ES*): (a) *ES*s were averaged if more than one measure of one outcome was reported; (b) when the measurement of different learning outcomes was reported (eg, using MCQ for knowledge retention and task performance for skill level), one *ES* was extracted for each dependent variable; (c) when there was more than one experimental or control group (eg, comparison between HMD, DVR and lecture conditions in a study), one *ES* was extracted for each comparison between the HMD and a type of non-immersive VR treatment. As a result, the total number of independent studies extracted was more than the number of articles selected ($k = 35$).

Research quality assessment

Jensen and Konradsen (2018) found poor quality among the studies in this field. Therefore, unlike their review work, which included conference papers, we only selected peer-reviewed journal articles. Further, following a suggestion to estimate the methodological quality of the studies used for meta-analysis (Gegenfurtner, Quesada-Pallarès, & Knogler, 2014), we conducted post hoc analysis of the research quality of the selected articles based on the Medical Education Research Study Quality Instrument (MERSQI). The MERSQI consists of six domains, with a total possible score for a study of 18, including study design, sampling, type of data, validity of evidence for evaluation instrument scores, data analysis and outcome (Reed *et al.*, 2007). This instrument has also been used to assess the quality of non-medical education research because it has been proven to be discipline neutral (Jensen & Konradsen, 2018). The mean MERSQI score for the selected studies was 12.6 ($SD = 1.16$), with a range of 10.5 to 14.5. In another educational study using MERSQI (Cook & Reed, 2015), the researchers found that across 26 articles, the average score was 11.3, with a range of 8.9 to 15.1. Thus, our selected studies had relatively high quality.

Coding procedures

The following moderator variables were generated by inspecting the studies reviewed and also based on the framework of previous meta-analysis studies (Merchant *et al.*, 2014; Sung, Chang, & Liu, 2016). The learner stage consisted of K-12, post-secondary and mixed-stage learners (ie, both students and adults). The learning domain included medical education, science education, physical education and specific abilities, such as spatial thinking, creativity, public speaking skills and analogical reasoning. The HMD hardware included Oculus Rift, HTC Vive and other brands. The research design included RCT and quasi-experimental studies. The control group treatment included DVR, lectures, real laboratory practices and other instructional approaches, such as self-directed learning through PPT, video, textbooks/e-textbooks, or 2D simulation. The learning duration was divided into four categories: less than 0.5 hour, between 0.5 and 1 hour, over 1 hour and not reported.

The HMD-based learning application types included representation, simulation and serious games. The classification followed the idea of Merchant *et al.*'s (2014) meta-analysis of desktop VR. In this study, we define representation as a type mainly for visual perception, for example, spherical video or virtual field trip; simulation as a type that requires intensive interactions, for example, skill training or virtual lab tasks; and serious games as a type that includes elements such as narrative plots, rules of the game, interactive cues and feedback and/or non-player characters (NPC) to inform the context of the game. Although the latter learning application type may contain features of the former type, we used the distinctive features of the latter type to differentiate applications of the former one. Therefore, we assigned each study to one of the three types based on the coding steps of game, simulation and then representation.

Besides learning outcomes, most studies also evaluated affection and the perceptions of VR learners using questionnaire surveys to measure such things as a sense of presence, self-efficacy, flow/engagement, sickness and mental workload. Due to the subjective bias induced by self-reporting results, we coded the learning outcome for only two types of learning performance: knowledge tests and task performance. We followed the previous work of Merchant *et al.* (2014) to investigate the long-term impact of HMDs on learning with respect to knowledge retention or learning transfers from simulation to real-world settings or a new context. Studies that included delayed tests were coded to compare the difference in effect size between immediate and delayed test performance. The first two authors independently coded all of the data. The agreement between the two coders was tested using Cohen's kappa, which was above 0.9 across all categories. Any disagreements between the two coders were discussed until a consensus was reached.

Analysis methods

The meta-analysis was carried out as follows: (a) calculation of effect size, (b) homogeneity analysis, (c) moderator analysis. The *ES* for each separate study was estimated based on sample size, mean, standard deviation or *t*-test, *F*-test and *p*-value using comprehensive meta-analysis (CMA) version 2.0. Hedges' g (Hedges, 1981) was chosen as the standardized measure of effect size for the continuous variables, rather than Cohen's d (Cohen, 1992), because the former corrects for smaller sample (less than 20) bias (Borenstein, Hedges, Higgins, & Rothstein, 2010). The ESs were then checked for potential outliers by examining whether an ES had a larger than average effect size by three standard deviations. All samples meet the three-sigma rule. When all of the *ES*s were calculated, the random effects model (REM) was used to calculate the overall mean *ES* for all of the studies considering the variability of the research design. An effect size of 0.2, 0.5, 0.8 was treated as small, medium and large respectively (Cohen, 1992).

Homogeneity analysis was performed to investigate the variance in *ES*s across studies. We calculated the Q-statistic to test whether all of the studies in the analysis shared a common effect size (Borenstein *et al.*, 2010). Larger Q-statistics corresponded to heterogeneous effect sizes. This was the precondition for moderator analysis. In such cases, REM was used for the calculation. Homogeneity was also assessed by I^2, which described the percentage of variation across studies resulting from true heterogeneity rather than from sampling error (Higgins & Thompson, 2002). Based on the rule of thumb proposed by Higgins and Thompson, the cut-off values of I^2 for low, medium and high heterogeneity were 0.25, 0.50 and 0.70 respectively.

The heterogeneity of the studies indicated that further grouping of individual ESs was needed to search for potential moderators that could account for the variance between the studies. Hence, moderator analysis was conducted through a mixed effect analysis (MEA) to explain systematic ES heterogeneity.

Results

Descriptive results of the included studies

We performed meta-analysis on 35 experimental or quasi-experimental studies conducted in Asian countries, including China, South Korea, India, Israel and Turkey (40%); European countries, including the UK, Italy, Germany, Ireland, Denmark and the Netherlands (37.1%); and American countries, including the US and Argentina (22.9%). The participants in the selected studies were 1847 learners covering K-12 and post-secondary education. The mean sample size was 54 participants (*SD* = 43.34). The largest proportion of studies took place in post-secondary education (65.7%), followed by K-12 settings (20%) and mixed learning stage groups (14.3%). Science was the most frequently studied learning domain (31.4%), followed by safety training (22.9%), specific abilities (22.9%), medical education (17.1%) and physical education (5.7%). More studies used RCT (65.7%) than quasi-experiments (34.3%). The numbers of studies for using three different learning application types, ie, simulation (28.6%), serious games (37.1%) and representation (34.3%), were similar. The most often used HMD hardware was Oculus Rift (40%), followed by HTC Vive (17%) and others. Knowledge tests (68.6%) were adopted over two times more than the evaluation of tasks for learning performance (31.4%). The largest proportion of studies used DVR as their control group treatment (37.1%), followed by real-world practice (22.9%), lectures (14.3%) and other non-immersive learning methods (25.7%). Apart from the studies with no reporting of learning duration (31.5%), most studies had a learning duration of<0.5 hour (34.3%), followed by >0.5 hour and <hour (17.1%) and >1 hour (17.1%). Regarding long-term learning impact, we identified nine studies that measured learning performance through both immediate and delayed tests, which comprised 25.7% of the included studies.

Overall effectiveness

We calculated the effect size of each study and the overall mean effect size of all 35 studies. Figure 1 represents the point estimates of the effect size and confidence interval in the forest plot format. In terms of independent effect size, the values ranged from −0.94 to 1.936 with 23 (66%) positive effects, ie, in favor of the HMD condition rather than other non-immersive learning interventions. Nine (26%) were negative and three (8%) had no effect. The REM analysis result showed that using HMDs had a small effect on the learning outcome ($g = 0.236$, $SE = 0.112$, $CI = [0.017, 0.454]$, $p = .035$). Our Q-value was 204.595 with $p < .01$, indicating that there were differences among the effect sizes resulting from factors other than subject-level sampling error. The I^2 for the overall model showed high heterogeneity ($I^2 = 0.83$), indicating that one or more moderators could account for this heterogeneity (Borenstein *et al.*, 2010).

Moderator analysis

The moderator analysis results are shown in Table 1. The Q-statistics revealed significant variance in effect size in terms of the learning stage of the participants as a moderator variable ($Q_B = 18.855$, $p < .01$). On average, the effect of HMDs over non-HMDs is significantly larger on the K-12 students ($g = 0.796$, *95%CI* [0.489, 1.103], $k = 7$) than on the mixed-stage ($g = 0.684$, *95%CI* [−0.176, 1.544], $k = 5$) or post-secondary students ($g = −0.015$, *95%CI* [−0.231, 0.2], $k = 23$), judging by the 95%*CI*s.

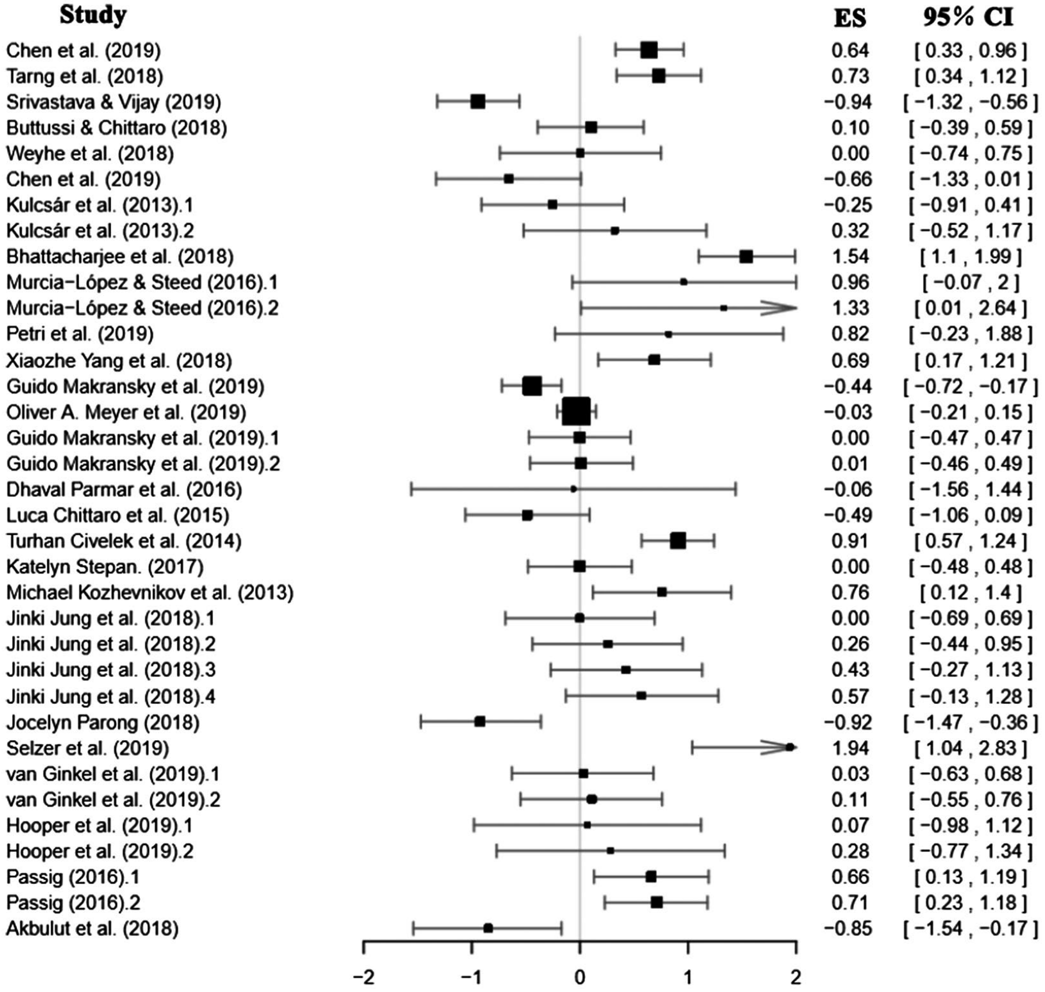

Figure 1: Forest plot of the selected studies

For HMD-based learning application types, the Q-statistics revealed significant variance in effect size ($Q_B = 6.439$, $p < .05$). Judging by the 95%*CI*, HMDs, as compared to non-HMDs, had an averaged medium effect on the learning performance of simulation approach ($g = 0.45$, *95%CI* [0.192, 0.708], $k = 10$), but the HMDs might or might not have effects over non-HMDs in studies using representation ($g = 0.305$, *95%CI* [−0.176, 0.786], $k = 12$) and serious game approach ($g = -0.002$, *95%CI* [−0.243, 0.239], $k = 13$), given that the 95%*CI* for the latter two educational states contained zero.

With respect to the treatment of the control group, the Q-statistics revealed significant variance in effect size ($Q_B = 11.165$, $p < .05$). On average, the effect of HMDs over lectures ($g = 0.782$, *95%CI* [0.358, 1.207], $k = 5$) is significantly larger than the effects of HMDs over real-world practice ($g = 0.387$, *95%CI* [−0.02, 0.793], $k = 8$) or DVR ($g = 0.122$, *95%CI* [−0.256, 0.499], $k = 13$), judging by the 95%*CI*s. The Q-statistics revealed no significant variance in effect sizes of all other moderator variables.

Table 1: Moderator analysis of learning performance

Moderator	*k*	*g*	*95%CI*	Q_B	*p-value*
Learner stage				18.855**	.000
K-12	7	0.796[c]	[0.489, 1.103]		
Post-secondary	23	−0.015	[−0.231, 0.200]		
Mixed	5	0.684[c]	[−0.176, 1.544]		
Learning domain				2.858	.582
Medical education	6	0.018	[−0.275, 0.311]		
Science education	11	0.383[b]	[−0.064, 0.830]		
Physical education	2	0.021	[−1.429, 1.471]		
Safety training	8	0.064	[−0.140, 0.267]		
Specific abilities	8	0.375[b]	[−0.195, 0.944]		
Publication region				0.883	.643
America	8	0.451[b]	[−0.06, 0.962]		
Asia	14	0.173[a]	[−0.15, 0.496]		
Europe	13	0.187[a]	[−0.214, 0.587]		
Experimental design				3.25	.071
Quasi	12	0.490[b]	[0.123, 0.857]		
RCT	23	0.091	[−0.141, 0.322]		
Learning application type				6.439*	.04
Simulation	10	0.450[b]	[0.192, 0.708]		
Serious games	13	−0.002	[−0.243, 0.239]		
Representation	12	0.305[a]	[−0.176, 0.786]		
HMD hardware				0.243	.886
Oculus Rift	14	0.289[a]	[−0.118, 0.696]		
HTC vive	6	0.124[a]	[−0.409, 0.657]		
Other	15	0.252[a]	[−0.068, 0.572]		
Testing format				0.034	.854
Knowledge test	24	0.255[a]	[0.003, 0.507]		
Task performance	11	0.205[a]	[−0.256, 0.667]		

Table 1: (Continued)

Moderator	*k*	*g*	*95%CI*	Q_B	*p-value*
Control group treatment				11.165*	.011
DVR	13	0.122[a]	[−0.256, 0.499]		
Lecture	5	0.782[c]	[0.358, 1.207]		
Real-world practice	8	0.387[a]	[−0.02, 0.793]		
Other	9	−0.053	[−0.338, 0.232]		
Learning duration				0.037	.998
>1 hour	6	0.246[a]	[−0.248, 0.741]		
0.5 ~ 1 hour	6	0.251[a]	[−0.198, 0.701]		
<0.5 hour	12	0.208[a]	[−0.102, 0.518]		
NA	11	0.252[a]	[−0.327, 0.831]		
Long-term impact				0.217	.641
Immediate test	9	0.23[a]	[0.011, 0.45]		
Delayed test	9	0.305[a]	[0.083, 0.526]		

Note k = number of independent studies; g = mean effect size; CI = confidence interval; Q_B = between-group homogeneity.
[a]Small effect, [b]Medium effect, [c]Large effect.
$^*p < .05$, $^{**}p < .01$.

Publication bias

Publication bias was assessed through Funnel pot, Classic fail-safe *N* test (Rosenthal, 1979), Egger's test (Egger, Smith, Schneider, & Minder, 1997) and Kendall's Tau rank test (Begg & Mazumdar, 1994). The funnel plot is a scatter plot of effect sizes estimated from the individual studies in a meta-analysis against a measure of study precision measured by the standard error. Generally speaking, a symmetric funnel plot suggests the absence of publication bias in a meta-analysis (Duval & Tweedie, 2000). As shown in Figure 2, possible publication bias was found in this study. The Classic fail-safe N test showed that 195 additional studies would be required to nullify the overall effect size found in this meta-analysis. These calculations showed that the number of null or additional studies needed to nullify the overall effect sizes found in this meta-analysis would be larger than the 5 *k* (*k* = 35 in this study) + 10 (Rosenthal, 1995). Egger's test ($p = .339$) and Kendall's Tau rank test ($p = .513$) were statistically non-significant, suggesting that there was no evidence of publication bias. Thus, overall, the results suggested the absence of publication bias.

Discussion

Discussion of overall effectiveness

Although previous studies (Chen *et al.*, 2020; Makransky, Terkildsen, *et al.*, 2019) have produced mixed results, the primary meta-analysis in this study indicated that HMDs are more effective than non-immersive learning approaches, with a small effect size. This finding corroborates the argument that HMD-supported immersive learning can improve learning performance (Concannon *et al.*, 2019). However, with 34% of the studies having no effect or a negative effect, we cannot exaggerate the advantage of HMD-based IVR. HMDs are certainly not a panacea to replace traditional ways of learning, as television, the Internet and mobile devices were once expected to be (Spector, 2013). Rather, they are a promising complement that can diversify learning experiences, even if they only achieve a learning performance that is comparable to that achieved with traditional ways of learning. Solid evidence has already been accumulated of the great advantage

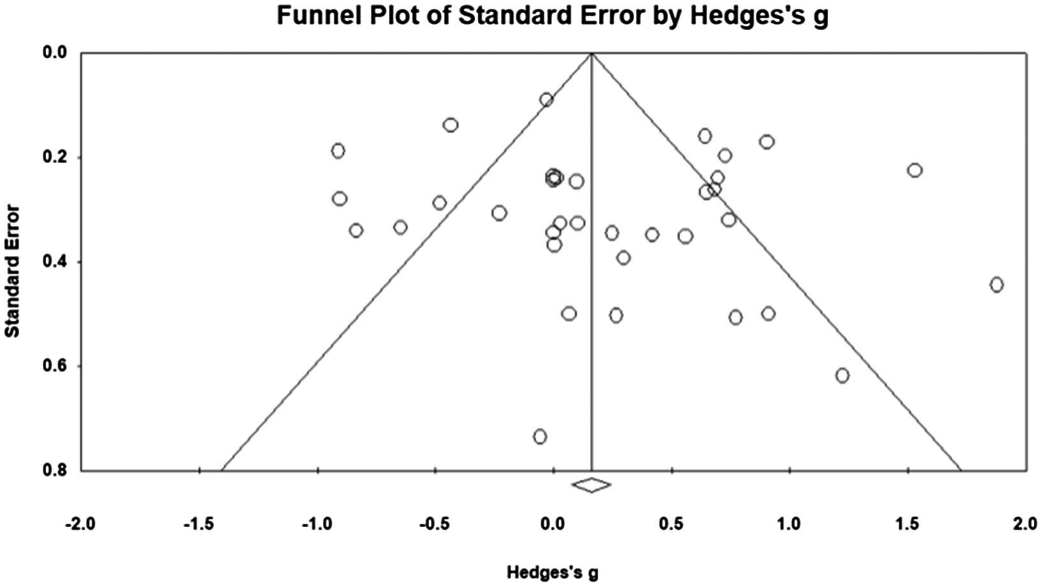

Figure 2: Funnel plot

of HMDs for important non-cognitive learning factors such as engagement, motivation and affection (Han, 2020; Makransky, Borre-Gude, & Mayer, 2019).

Discussion of moderating factors

Regarding the learning stage of HMD users, previous literature reviews have focused more on HMDs in post-secondary settings (Concannon *et al.*, 2019; Radianti *et al.*, 2020). However, this study suggests that this innovative technology has a greater impact on K-12 students than on post-secondary students' learning performance. HMDs provide high-immersive multi-sensory stimuli that have a direct influence on the affective factors of youth learners. In contrast, adult learning focuses on complex knowledge and skills and is more goal-oriented so that an HMD environment may not easily achieve its objectives. Nonetheless, even if there is less effect at the post-secondary stage, it does not necessarily mean that the application of HMDs should not be expanded to this group of learners. In contrast, widespread use in professional training, including in the medical and engineering fields, suggests that more customized HMD learning applications with the sound instructional design are needed to improve its usefulness.

In terms of learning application types, we categorized the studies into three types incorporated into HMDs: simulation, representation and serious games. It is interesting to compare the moderator effects of the learning application types in this study to the DVR settings in previous meta-analyses (Merchant *et al.*, 2014). In this study, the HMD and DVR settings had similar medium level effects (all around 0.4) for both simulation and virtual world representation respectively. However, the significant effect of games in DVR was not found in the HMD settings. This may be due to the limited amount of HMD-compatible serious gaming applications and suggests a design challenge for educational games with HMDs. Nevertheless, following the finding that game-based learning has shown distinct benefits in DVR, HMD-related educational games merit further research and development.

This study compared the effect of different control group treatments and showed that HMD outperformed lectures the most, followed by real-world practice and DVR. It was unsurprising that HMD had a better impact than lectures, given that the former is a more engaging and active way to learn. However, it was unexpected to find that a larger effect occurred when HMDs were compared with real-world practice as opposed to DVR. Given the non-significant effects in these two conditions, the control treatments deserve further investigation to clarify the conditions under which HMDs work better than DVR or real-world practice. In any event, these findings provide evidence that VR technology has unique media benefits that traditional learning in the physical world does not.

We also found a non-significant difference in effect size with variation in learning duration, which suggests that the novelty effect found in previous studies (Makransky, Borre-Gude, *et al.*, 2019; Merchant *et al.*, 2014) might not have been a severe issue biasing the effectiveness results. Adequate warm-up sessions can help students or first-time users become familiar with the HMD environment. Conversely, the similar effect size across different lengths of HMD exposure corroborates the argument that more usage can diminish the learning effect due to fatigue or sickness (Chen *et al.*, 2019).

Another noteworthy finding relates to learning assessment with respect to the testing format and long-term impact. This study reveals a comparably small effect between knowledge testing and task performance. This finding in contrast to the previous argument that HMDs only have an advantage in improving cognitive, psychomotor and affective skills (Jensen & Konradsen, 2018). A possible explanation is that the affordance of high immersiveness and interactivity in HMDs benefits both knowledge understanding and skills development (Radianti *et al.*, 2020).

The finding of similar significant small effects between immediate and delayed tests is partially consistent with previous findings from DVR settings (Merchant *et al.*, 2014). In their research, Merchant *et al.* (2014), found that learning performance was retained longer in DVR-based serious games but that it declined in simulations. In contrast, this study's promising results show that the benefits of HMDs are maintained over time or transferred to other real-world contexts in general which provides evidence of learning transfer.

We also found that quasi-experimental studies had a medium effect, whereas RCT had a small effect. Although quasi-experimental studies are usually conducted in more natural learning settings, they may suffer from an ecological validity issue and easily produce inflated effect sizes (Makransky, Borre-Gude, *et al.*, 2019). Nonetheless, the practical value of this type of research design can be greater because it contributes to our knowledge of integrating HMDs into formal education. Therefore, more replicate and scale-up quasi-experimental studies are welcomed to overcome the generalizability issues.

Conclusion

This study conducted a meta-analysis of empirical research on the use of HMDs in educational interventions published in peer-reviewed journals. It contributes to VR research by revealing that HMD-based immersive learning has an overall better effect on learning performance than non-immersive learning approaches. By analyzing the studies' characteristics as moderator variables, we found that HMDs have encouraging distinct effects on K-12 learners, improving both their knowledge and skill development. This is especially the case for science education and specific abilities development when simulation or virtual world representations were compared with lectures or real-world practices and maintained their learning effects over time. This study also provides suggestions for implementing HMD-based learning, such as selecting HMD devices and determining HMD exposure duration and research design considerations, including experimental designs and evaluation approaches.

However, this meta-analysis also suffers from two limitations. First, some of the empirical studies lacked adequate statistical information for effective size calculation. Therefore, they could not be included in the meta-analysis, which might have affected the results. Second, most previous studies failed to consider the theory-driven instructional design with respect to HMD-based immersive learning (Parong & Mayer, 2018). This limited us to conducting a more fine-grained analysis of technological-pedagogical symbiotic relations. To conclude, immersive learning through HMDs has yet to appear as a mature field. This meta-analysis encourages us to explore more effective ways to implement HMD-based teaching and learning practices in the future.

Acknowledgements

This study is supported in part by the "Flower of Happiness" Pilot Research Fund (No. 2019ECNU-XFZH014) and Peak Discipline Construction Project of Education from East China Normal University.

Statements on open data, ethics and conflict of interest

The survey data are available upon request through email contact.

No ethical issues exist in this study.

No conflict of interest exist in this study.

References

Barricelli, B. R., Gadia, D., Rizzi, A., & Marini, D. L. R. (2016). Semiotics of virtual reality as a communication process. *Behaviour & Information Technology*, *35*(11), 879–896.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*, 1088–1101.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, *1*(2), 97–111.

Buttussi, F., & Chittaro, L. (2018). Effects of different types of virtual reality display on presence and learning in a safety training scenario. *IEEE Transactions on Visualization and Computer Graphics*, *24*(2), 1063–1076. https://doi.org/10.1109/tvcg.2017.2653117

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

Chen, X. M., Chen, Z. B., Li, Y., He, T. Y., Hou, J. H., Liu, S., & He, Y. (2019). ImmerTai: Immersive motion learning in VR environments. *Journal of Visual Communication and Image Representation*, *58*, 416–427. https://doi.org/10.1016/j.jvcir.2018.11.039

Chien, S. Y., Hwang, G. J., & Jong, M. S. Y. (2020). Effects of peer assessment within the context of spherical video-based virtual reality on EFL students' English-Speaking performance and learning perceptions. *Computers & Education*, *146*, 103751. https://doi.org/10.1016/j.compedu.2019.103751

Chittaro, L., & Buttussi, F. (2015). Assessing knowledge retention of an immersive serious game vs. a traditional education method in aviation safety. *IEEE Transactions on Visualization and Computer Graphics*, *21*(4), 529–538. https://doi.org/10.1109/tvcg.2015.2391853

Civelek, T., Ucar, E., Ustunel, H., & Aydin, M. K. (2014). Effects of a Haptic augmented simulation on K-12 students' achievement and their attitudes towards physics. *EURASIA Journal of Mathematics, Science and Technology Education*, *10*(6). https://doi.org/10.12973/eurasia.2014.1122a

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Concannon, B. J., Esmail, S., & Roduta Roberts, M. (2019). Head-mounted display virtual reality in post-secondary education and skill training: A systematic review. *Frontiers in Education*, *4*(80), 1–23.

Cook, D. A., & Reed, D. A. (2015). Appraising the quality of medical education research methods: The medical education research study quality instrument and the Newcastle-Ottawa scale-education. *Academic Medicine*, *90*(8), 1067–1076.

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, *323*(5910), 66–69.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634.

Falloon, G. (2020). From simulations to real: Investigating young students' learning and transfer from simulations to real tasks. *British Journal of Educational Technology*, *51*(3), 778–797. https://doi.org/10.1111/bjet.12885

Feng, Z., González, V. A., Amor, R., Lovreglio, R., & Cabrera-Guerrero, G. (2018). Immersive virtual reality serious games for evacuation training and research: A systematic literature review. *Computers & Education*, *127*, 252–266.

Ganier, F., Hoareau, C., & Tisseau, J. (2014). Evaluation of procedural learning transfer from a virtual environment to a real situation: A case study on tank maintenance training. *Ergonomics*, *57*(6), 828–843.

Gegenfurtner, A., Quesada-Pallarès, C., & Knogler, M. (2014). Digital simulation-based training: A meta-analysis. *British Journal of Educational Technology*, *45*(6), 1097–1114.

Han, I. (2020). Immersive virtual field trips in education: A mixed-methods study on elementary students' presence and perceived learning. *British Journal of Educational Technology*, *51*(2), 420–435.

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

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558.

Hwang, W. Y., & Hu, S. S. (2013). Analysis of peer learning behaviors using multiple representations in virtual reality and their impacts on geometry problem solving. *Computers & Education*, *62*, 308–319.

Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies*, *23*(4), 1515–1529.

Johnson-Glenberg, M. C. (2018). Immersive VR and education: Embodied design principles that include gesture and hand controls. *Frontiers in Robotics and AI*, 5. https://doi.org/10.3389/frobt.2018.00081

Jung, J. K., & Ahn, Y. J. (2018). Effects of interface on procedural skill transfer in virtual training: Lifeboat launching operation study. *Computer Animation and Virtual Worlds*, *29*(3–4), e1812. https://doi.org/10.1002/cav.1812

Kozhevnikov, M., Gurlitt, J., & Kozhevnikov, M. (2013). Learning relative motion concepts in immersive and non-immersive virtual environments. *Journal of Science Education and Technology*, *22*(6), 952–962. https://doi.org/10.1007/s10956-013-9441-0

Leder, J., Horlitz, T., Puschmann, P., Wittstock, V., & Schütz, A. (2019). Comparing immersive virtual reality and powerpoint as methods for delivering safety training: impacts on risk perception, learning, and decision making. *Safety Science*, *111*, 271–286.

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

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, *60*, 225–36. https://doi.org/10.1016/j.learninstruc.2017.12.007

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.

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. https://doi.org/10.1016/j.compedu.2019.103603

Parmar, D., Bertrand, J., Babu, S. V., Madathil, K., Zelaya, M., Wang, T. W., ... Frady, K. (2016). A comparative evaluation of viewing metaphors on psychophysical skills education in an interactive virtual environment. *Virtual Reality*, *20*(3), 141–157. https://doi.org/10.1007/s10055-016-0287-7

Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology*, *110*(6), 785–797. https://doi.org/10.1037/edu0000241

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, *147*, 103778.

Reed, D. A., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., & Wright, S. M. (2007). Association between funding and quality of published medical education research. *JAMA*, *298*(9), 1002–1009.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*(2), 183–192.

Scoresby, J., & Shelton, B. E. (2011). Visual perspectives within educational computer games: Effects on presence and flow within virtual immersive learning environments. *Instructional Science*, *39*(3), 227–254.

Spector, J. M. (2013). Trends and research issues in educational technology. *Malaysian Online Journal of Educational Technology*, *1*(3), 1–9.

Srivastava, P., Rimzhim, A., Vijay, P., Singh, S., & Chandra, S. (2019). Desktop VR is better than non-ambulatory HMD VR for spatial learning. *Frontiers in Robotics and AI*, *6*. https://doi.org/10.3389/frobt.2019.00050

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

Sung, Y. T., Chang, K. E., & Liu, T. C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, *94*, 252–275.

Weyhe, D., Uslar, V., Weyhe, F., Kaluschke, M., & Zachmann, G. (2018). Immersive anatomy atlas—Empirical study investigating the usability of a virtual reality environment as a learning tool for anatomy. *Frontiers in Surgery*, *5*. https://doi.org/10.3389/fsurg.2018.00073