The Effects of Virtual Reality, Augmented Reality, and Mixed Reality as Training Enhancement Methods: A Meta-Analysis

Alexandra D. Kaplan, **Jessica Cruit**, University of Central Florida, Orlando, USA, **Mica Endsley**, SA Technologies, Gold Canyon, AZ, USA, **Suzanne M. Beers**, MITRE Corporation, Colorado Springs, CO, USA, **Ben D. Sawyer**, and **P. A. Hancock**, University of Central Florida, Orlando, USA

Objective: The objective of this meta-analysis is to explore the presently available, empirical findings on transfer of training from virtual (VR), augmented (AR), and mixed reality (MR) and determine whether such extended reality (XR)-based training is as effective as traditional training methods.

Background: MR, VR, and AR have already been used as training tools in a variety of domains. However, the question of whether or not these manipulations are effective for training has not been quantitatively and conclusively answered. Evidence shows that, while extended realities can often be time-saving and cost-saving training mechanisms, their efficacy as training tools has been debated.

Method: The current body of literature was examined and all qualifying articles pertaining to transfer of training from MR, VR, and AR were included in the meta-analysis. Effect sizes were calculated to determine the effects that XR-based factors, trainee-based factors, and task-based factors had on performance measures after XR-based training.

Results: Results showed that training in XR does not express a different outcome than training in a nonsimulated, control environment. It is equally effective at enhancing performance.

Conclusion: Across numerous studies in multiple fields, extended realities are as effective of a training mechanism as the commonly accepted methods. The value of XR then lies in providing training in circumstances, which exclude traditional methods, such as situations when danger or cost may make traditional training impossible.

Address correspondence to Alexandra D. Kaplan, Department of Psychology, University of Central Florida, 4111 Pictor Lane, Orlando, FL 32816, USA; e-mail: adkaplan@knights.ucf.edu

HUMAN FACTORS
Vol. 63, No. 4, June 2021, pp. 706–726
DOI:10.1177/0018720820904229

Keywords: virtual environments, transfer of training, immersive environments, meta-analysis

THE PROMISE OF AUGMENTED REALITY, VIRTUAL REALITY, AND MIXED REALITY FOR TRAINING

Training is required in order for humans to develop necessary performance skills (see Holding, 1989). Learning protocols can be both expensive and time consuming. Thus, any advancement in technology or methodology that might reduce the cost, either in financial or in temporal terms, will be of great relevance to many individuals, organizations, and industries. For this reason, simulation-based training has shown promise and is gaining acceptance as a means to increase training efficiency (Hancock, 2009). Although simulation-based training can be delivered via portable tablets or conventional flat panel displays, we are witnessing an increasing use of augmented reality (AR) and virtual reality (VR) displays, and new mixed reality (MR) displays, which include both AR and VR. VR technology generally uses a headset, blocking out visual stimulus from the real world. AR allows users to see the real world, but overlays virtual elements. MR combines the two, including aspects of both the real and virtual world. Other definitions of the technology employed in those categories, as well as sources, are listed in Appendix A. Extended reality (XR) is the umbrella term that refers to these three different types of simulations. These technologies promise to reduce some of the costs associated with expensive training, especially where

spatial information is important, such as in the field of aviation (Salas et al., 1998). XR training can also eliminate some of the risks inherent to high-level training by placing individuals in a simulation rather than a real-world dangerous situation.

Training in XR promises to have benefits beyond simply supplementing traditional training protocols. Whether it is a time-saving, cost-saving measure or not, the benefits that exist may still outweigh potential drawbacks. Even low-fidelity VR contains aspects of the physical world that cannot be replicated in the traditional classroom settings (Kozak et al., 1993). One can argue that a simulated battlefield has more in common with a real battlefield than does a classroom.

The potential also exists that XR might be used to help people prepare for situations that do not yet exist, or are not yet safe for humans, and thus cannot be adequately prepared for in situ, for example, prospective missions to Mars (Hancock, 2017). XR allows for training in locations and for events where there are no safe and realistic parallels. Additionally, such simulations can be rapidly and efficiently updated as new information becomes available, unlike other full-fidelity built environments, which are much less malleable.

However, training in XR does not solve all of the problems that plague current training methods. While simulation may be the solution to some issues, it comes with its own set of caveats and concerns that have to be balanced against those of more traditional methods. One such caveat is the rate of technological innovation, which far exceeds the speed of designing, implementing, and testing a training regimen (Hancock & Hoffman, 2015). Therefore, by the time a simulator's efficacy as a training tool is fully tested, it is already out of date. The variability between the technology in use makes it difficult (if not impossible) to empirically replicate an earlier investigation of XR-based training effectiveness.

In light of these concerns and the numerous benefits of XR-based training, it is important to assess the applicability of the training (in XR) to execution of the task (in the real world). The principle of "encoding specificity" indicates that when the learning environment is sufficiently different from the environment in which learning is subsequently measured, performance tends to suffer (Tulving & Thomson, 1973). This principle was further explored experimentally by Godden and Baddeley (1975), in which they found that scuba divers who memorized lists of words on dry land recalled those lists better above, rather than below, the surface of the ocean. This calls into question whether learning can fully transfer from practice to performance when the performance occurs in a different environment from training, such as is the case when XR is used. In the Godden and Baddeley (1975) example, what if rather than memorizing words, an individual was learning how to safely operate an underwater air tank? In such a situation, training on land for subsequent performance underwater could prove disastrous if the training did not transfer effectively. This same concern can be potentially extended to training in XR. The situations in which simulation-based training has the most benefit (i.e., risky, expensive, and/or unsafe conditions) also have the highest cost of failure when training proves inadequate.

Outcomes of Extended Reality and Simulation-Based Training

Simulation-based training has already proven advantageous for the military. It has been shown that pilots who first trained in simulators required less in-flight training time before reaching an acceptable level of competence (Rantanen & Talleur, 2005). Simulators, as surrogates for many of the expensive and limited resources or dangerous situations encountered by the military, free up equipment (such as runways) that might be unavailable due to other operational demands and allow the training of dangerous operations (such as flight and air traffic control) in a safe manner. Additionally, training in simulation environments offers the possibility of immediate feedback (Haque & Srinivasan, 2006). Such immediacy promotes faster and more accurate training by letting the learner self-correct mistakes before the result of the error is propagated.

Training in XR appears to hold similar promise as a solution for many of the problems that

currently make traditional training difficult and ineffective. Of course, the question of whether or not XR is a suitable medium for training is the subject of some debate. Applicability of XR as a training platform lacks some of the haptic feedback that the real world offers. Additionally, the variability in visual quality of different XR products, lag and tracking problems, and the potential for simulator sickness are all sources of limitation that may diminish training efficacy. To that end, the success of XR-based training must be evaluated empirically across differing applied fields. To accomplish this is not a straightforward task. Learner capacities vary, and inherent individual differences have been shown to affect the transfer of training, whether from real or simulated sources (Blume et al., 2010). Additionally, the modes of delivery of virtual training vary in fidelity and quality. Both of these factors have significant effects on later performance measures.

The Transfer Effectiveness Ratio

One of the most beneficial aspects of XR assessment is that transfer of training from simulation to the target environment can be directly measured. The transfer effectiveness ratio (TER) determines the value of time spent training in a simulator by calculating the efficacy of the (virtual) training session (and see Roscoe, 1971). The equation is as follows:

$$TER = \frac{Yc - Yx}{Yc} \times 100$$

where Y_c indicates the amount of time or number of trials it takes to train an individual on a specific task, and Y_x indicates the time it takes to train someone who has already trained on a simulator, to complete that same task to the same level of competence. Thus, a TER value of 0.5 indicates that training on a simulator can reduce the in-person training time by one-half. Using this formula, it is possible to specify numerically the time saved by training using simulation in general or a particular XR technology. However, not all training success factors can be measured in terms of time saved. Further, not all domains have the resources or ability to experiment in order to specify the efficacy of each particular set of simulation content and delivery mechanism that might be considered. How, then, can the efficacy of simulator-based (and specifically XR-based) training be explored? To answer this question we conducted a meta-analysis of the current empirical literature on the topic.

The Present Meta-Analysis

VR has previously been examined in meta-analysis. However, XR is, in general, so broad a topic, and training so important an area, that not all aspects of training in XR have been addressed in research, and let alone in meta-analysis. One previous meta-analysis examined only the efficacy of surgical simulators (see Haque & Srinivasan, 2006), a vital but small area. Fletcher et al. (2017) examined a broader scope, analyzing the effectiveness of VR in training. However, their selection criteria were less stringent than for the meta-analysis we employ. Fletcher's analysis allowed articles where psychological flow and enjoyment during virtual training represented an outcome variable; additionally, articles were included in their assessment where performance was measured during the time in the virtual environment or with the help of virtual aids. In the present meta-analysis we employed stringent selection criteria in order to fill an important need for a tightly controlled, methodologically sound, and comprehensive meta-analysis. We only included articles if performance measurement took place after virtual training, but entirely in the non-virtual world, to demonstrate training transfer. Our focus is narrower, but no less important; we look to determine the direct effect that training using XR has on real-world performance. These findings will serve to inform the design and application of training regimens using XR.

METHOD

Searching the Literature

A literature search was conducted in order to identify all published, peer-reviewed articles on the topic of training transfer from XR-based training. Search terms consisted of a primary phrase describing forms of XR, combined with a secondary group of phrases pertaining to training. All possible combinations of the

TABLE 1: Tabulation and Combination of Search Terms

Primary Term	Secondary Term
Virtual reality	Training
VR	Learning
Augmented reality	Encoding specificity
Mixed reality	
Simulation	

search terms were used, and the terms are listed in Table 1.

The 15 search strings were each entered into a series of search engines (ProQuest, EbscoHost, and Google Scholar). All results were briefly examined to determine whether they met inclusion criteria. The search took place in February 2019 and included all articles published prior to that time. Additionally, prominent scholars in the area of XR were contacted and asked whether they had any relevant research approaching fruition, which might fit the criteria. Identified relevant articles (n = 130) were then examined more closely and rejected (n = 105) or included (n = 24) in the meta-analysis. One article was identified that was published after the initial search and was included in the analysis (Whitmer et al., 2019). This process is illustrated in Figure 1.

Inclusion Criteria

Articles met inclusion criteria if at least one of the reported outcome variables measured performance that took place after training in XR. Articles were also required to be published in a peer-reviewed journal, the proceedings of a conference, part of a dissertation or thesis, or a peer-reviewed technical report. Articles were not included if the population was under 18, such as elementary school-age students. Articles were also rejected for inclusion if the reported statistics did not provide sufficient information so as to determine an effect size. Suitable statistics in this analysis were r, d, F, t, or means and standard deviations. Finally, it was required that all articles involved training in MR, AR, or VR. Articles were *not* included if the outcome variable measured something other than performance after training, such as level of enjoyment or engagement. Additionally, the performance being measured had to take place in the real world. Experimental results were rejected if the outcome variable was performance with the aid of XR or performance in a simulation. Articles were required to include original empirical data. If a dissertation included a sample, and that same data were then later used in a referred publication or conference proceedings paper, the sample was only included once in the final analysis. Determination of inclusion and subsequent coding of the statistical data in the articles were completed by two individuals.

If a study examined the appropriate variables but did not include sufficient statistical information to determine an effect size, the authors were contacted directly and asked to supply such information. If the authors did not respond with or could not supply the needed statistics, the article was not included. While there were a variety of different forms of XR used for training, all fell into one of the aforementioned three categories (AR, VR, or MR).

Variables

The outcome variable, in all included studies, was some dimension of performance taken after training in XR had occurred. Predictor variables fell into three general categories related to (a) the simulation, such as immersiveness, (b) the trainee, such as age, or (c) the task, such as task difficulty.

Immersiveness. Of the XR-related variables, one often-explored concept involved comparisons across differing virtual environments. For example, VR using a headset was considered more virtually immersive than desktop VR or AR. Each of these differing levels of immersion may have been compared to an entirely nonsimulated control condition (e.g., real-world training) or to a less immersive training tool, such as an interactive video or a simple instruction manual. Here, we call this variable "immersiveness," and use the word to refer to any comparison between environments where one is more virtually immersive than the other. Despite the

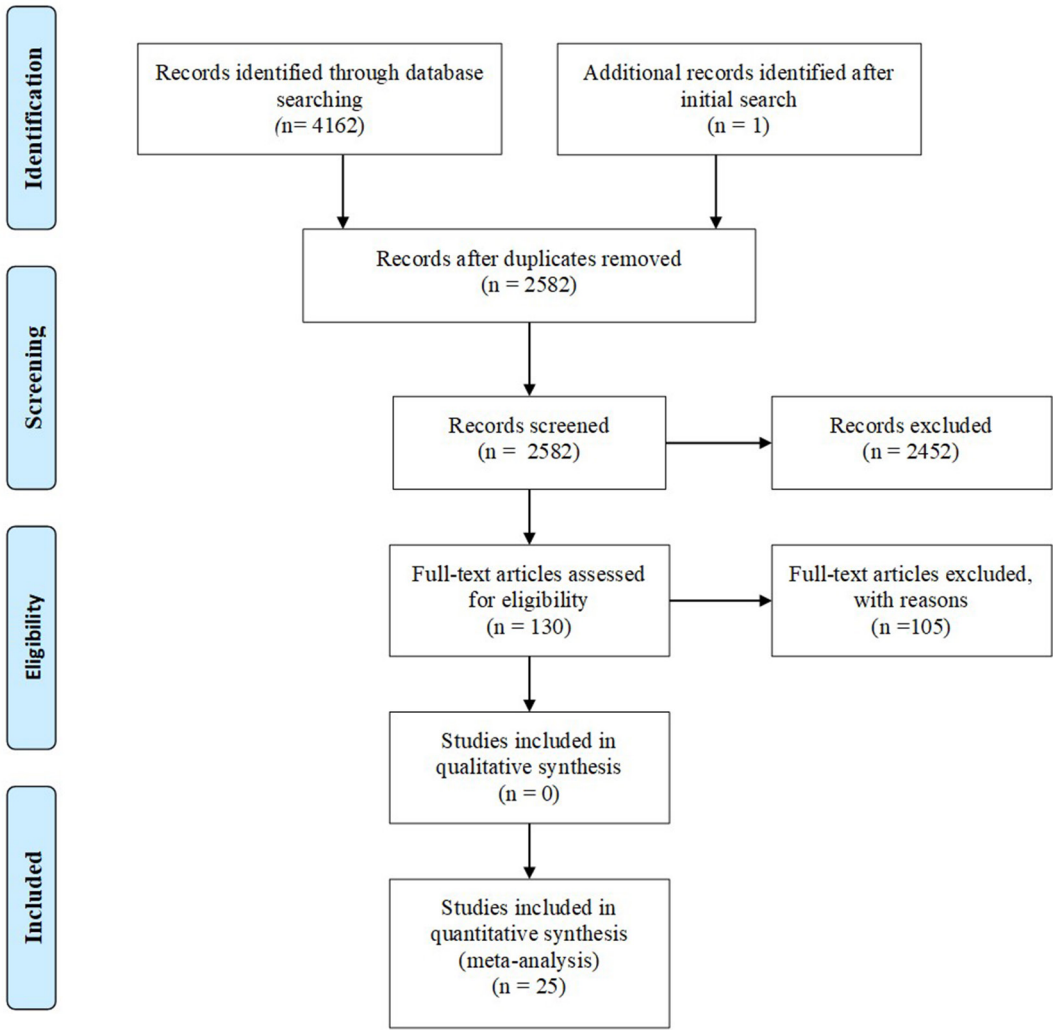

Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram (Moher et al., 2009).

differences in level of immersiveness of the training environment, all studies included in this analysis measured final performance in the real world.

VR vs. control. A subset of the studies where immersiveness was a factor compared training in a fully immersive VR setting to training in a nonvirtually immersive control. Such studies were included both in the overall effect size analysis of immersiveness and in their own specific subanalysis.

Pre/post training. The variable "pre/post" includes any comparison between an individual or a group's performance before XR-based training intervention, with performance after that same intervention. This variable examines the post-training improvement (or lack thereof). Regardless of whether or not performance improvement was the hypothesis of the original article, if performance was measured and sufficient statistical information was supplied, then the prescore was compared to the post-training

score *only* for groups where the training was virtual.

Task Type

The data were examined to determine the direct effect on performance of each variable described above. The data were also examined with task type as a moderator. The three types of task categories were as follows.

Cognitive tasks. Cognitive tasks included situations in which participants learned information that later had to be either remembered directly, such as in a test of recall, or utilized in a subsequent applied setting.

Physical tasks. Physical tasks involved some sort of bodily training, such as balance or aerobic activities. The predominance here was on psychomotor skill assimilation.

Mixed tasks. Some tasks included combinations of both physical and cognitive requirements, such as a maintenance task that required participants to use learned physical skills while simultaneously recalling applicable procedural information (see Marras & Hancock, 2014).

Included Articles

Twenty-five articles met the above-stated criteria and so were included in the analysis. Twenty-three articles examined the XR-related factor of immersiveness. The majority of these included at least one pairwise comparison between a VR training condition and a control setting ($k = 21$). A number also examined AR compared to a control setting ($k = 5$). One examined training results after training in VR as compared to AR ($k = 1$), and two studies looked at different levels of AR ($k = 2$).

Twelve studies were included in the pre/post comparison, one of which focused on AR and the rest on VR. Only one article that met our inclusion criteria examined predictor variables other than XR (Bier et al., 2018). This work included the effects of both age and task difficulty on performance after training in VR.

Data Analytic Strategy

Many of these articles examined both VR and AR and most reported more than one pairwise relationship between variables of interest. Thus, multiple effect sizes were taken from each. Some articles were included in the number of studies (k) for multiple predictor variables if that article reported enough data to determine an effect size of two different variables (e.g., if a study reported enough statistical information to calculate an effect size for both immersiveness and task difficulty, then the two separate effect sizes were calculated). For variables where $k = 1$, only one article reported results in a method suitable for inclusion in the meta-analysis. The calculated Cohen's d is provided here, but, as data only come from one source, any confidence intervals surrounding d would not be meaningful and thus were not included. Such information is included to illustrate what is covered by the current research. Individual effect sizes are listed in Appendix B.

A total of 176 effect sizes were included, which were each converted to Cohen's d and weighted, based on the number of participants included. Effects between similar pairs within the same study were combined. Therefore, even if any one particular study had several effect sizes measuring the same variable, results were aggregated in order that each study only had one overall effect size for each specific predictor. If an article had two separate studies, using two *different* samples, then two effect sizes were calculated. This was done so that the results from one sample would not disproportionately influence the outcome and to maintain independence.

Although all dependent variables represented a performance outcome, the scales used to measure performance varied widely. In addition, the concept of "performance" itself varied between articles. Therefore, it was not possible to compare directly between studies. As a result we used a random-effects model when calculating the meta-analytic results. For each study a weighted value d was determined as the effect of the predictor variable on performance, *in that particular study*. These weighted effect sizes were then used to determine the overall effect

TABLE 2: Overall Effect Sizes of the Associated Variables

Predictor	Number of Studies (*K*)	Cohen's *d*	95% Confidence Interval Lower Limit	Upper Limit
Immersiveness	23	−.07	−0.22	+0.07
VR compared to control	21	−.13	−0.27	+0.02
Pre/post training	12	.09	−1.05	+1.23
Age	1	−.08		
Task difficulty	1	−.15		

VR, virtual reality.

TABLE 3: Effect Sizes by Task Type

Task Type	Number of Studies (*K*)	Cohen's *d*	95% Confidence Interval Lower Limit	Upper Limit
Cognitive tasks	9	.01	−0.24	+0.27
Mixed tasks	12	−.07	−0.31	+0.17
Physical tasks	8	.36*	+0.01	+0.70

* indicates significant effect beyond $p < .05$ level.

size and the associated 95% confidence intervals. SPSS was used to compute the effect sizes.

The effect sizes determined from this analysis were not intended to determine whether XR training was effective. Rather, results addressed the question of whether it was different from the other methods of training to which it was being compared. If the effect size of immersiveness is both significant and positive, it represents improvement on traditional training. If the effect is both significant and negative, then the opposite is true. If a zero effect size falls within the 95% confidence interval, it would indicate here that all levels of immersiveness exert an equivalent (or similar) effect on performance outcome.

RESULTS

The effect sizes are reported in Table 2. The table also indicates the number of separate studies investigating each respective predictor (*k*). Some articles included more than one study. Weighted overall levels of *d* are included, as are 95% confidence intervals for each relationship. Analysis of the associated variable of immersiveness, as well as the subset analysis of VR compared to control, showed no significant difference between levels of performance post training, regardless of the virtual immersiveness. While the negative effect sizes ($d = -.07$ and $d = -.13$, respectively) indicate a slight decrement in training effectiveness when a virtual environment was used, the fact that the confidence interval included zero indicates that whether one trains in a virtual or a real setting, the results are essentially equivalent. In essence, these findings indicate that XR experiences are as effective as traditional training approaches.

Table 3 shows the effect size based on task type. Results show that XR is a more suitable medium for training on physical tasks ($d = .36$), but otherwise the type of task learned in simulation does not have an effect on the performance outcome.

The overall pattern of effect sizes is compared in Figure 2. Additionally, in Figure 2, potentially moderating influencers are divided by task type. While these "interaction" effects are interesting, we have to caution against relying excessively on these results at present.

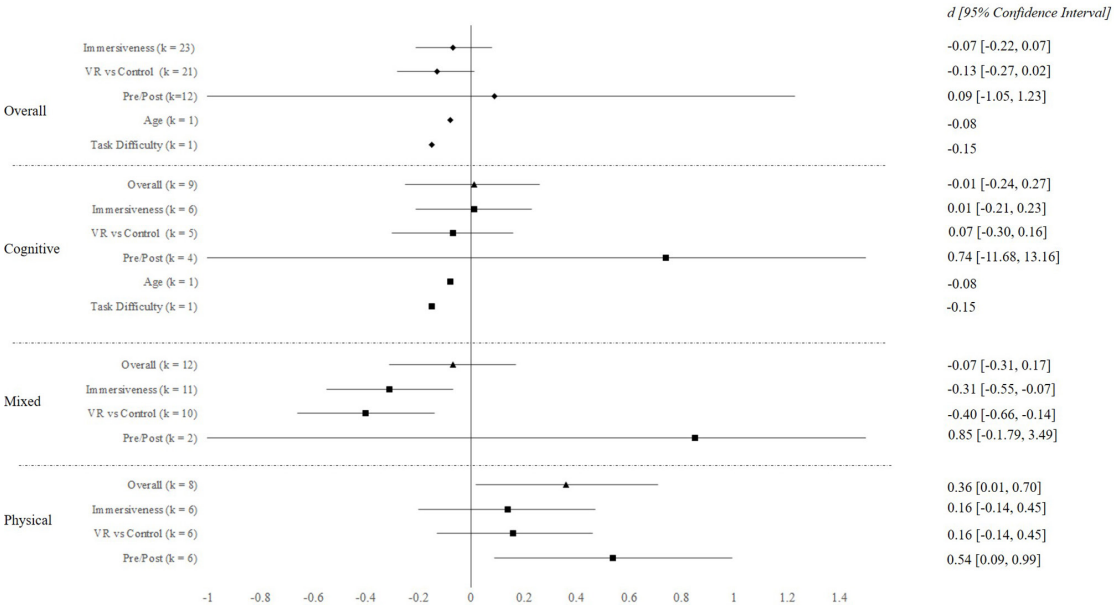

Figure 2. Forest plot of effect sizes by associated variable; 95% confidence intervals shown.

This is because, with the addition of each moderating factor, the number of applicable studies is smaller. Thus, there is less between-study variation in the calculations for the smaller number of studies. For forest plots showing the effects of individual studies by task type/predictor clusters, see Figures 4 through 6 in Appendix C.

Additional Analysis and Overlap Between Conditions

Within each study, performance tended to be quite similar between immersiveness conditions. To that end, it was important to determine the similarity between the conditions beyond simply noting that for immersiveness variables confidence intervals included zero. For this reason, scores were compared in order to determine overlap. The data could not be compared directly, as each study used different scales to measure performance. So, the mean performance of each training condition within a study was converted to a *z*-score. The average *z*-score for each condition, as well as a 95% confidence interval, is shown below in Figure 3. As these *z*-scores came from the two conditions present in each evaluation, they are mirror images of each other. Only their size and magnitude are meaningful; the value of the average *z* itself has no real-world meaning except in indicating the difference between scores of participants in each condition. The fact that average *z*-scores were so small in value serves to highlight the similarity between conditions. The mean *z*-score for the more immersive condition was lower than the less immersive and control conditions, yet the overlap between confidence intervals was large. These findings indicate that, though a more virtually immersive training condition results in slightly worse performance than a real training setting, the majority of individuals will show similar results after training, regardless of the level of virtual immersiveness.

DISCUSSION

The fact that the zero could not be excluded from the confidence intervals relating to any of the present overall predictors indicates an apparent equivalency between XR training and traditional instructional techniques employed in situations such as a classroom. If we take

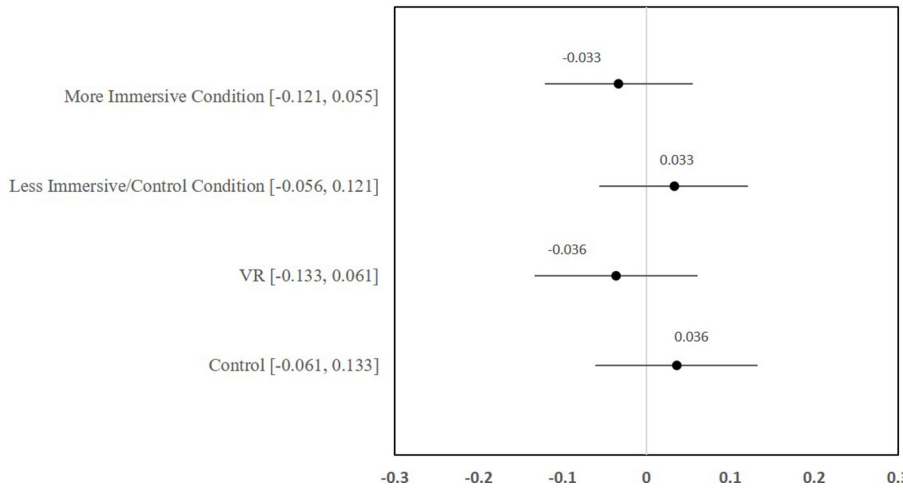

Figure 3. Immersiveness variables and confidence intervals by condition.

an optimistic perspective, these results confirm that the use of VR, MR, and AR training procedures provides at least an equivalent performance result to that which is normally experienced in traditional instruction methods. If this is the case, and performance outcomes after XR training are not significantly different than outcomes after traditional training, then the previously enumerated benefits of XR training (such as safety, cost, and ease of implementing changes) make it, on the whole, a more valuable investment of time than traditional training methods. After all, if the performance outcome is essentially the same, the other benefits of XR training make it a superior option.

However, it must be acknowledged that XR has often been held out to offer superior training capacities (especially in popular press and by various vendors). The results of the present meta-analysis indicate that the case for this proposition is at best "*not proven*." Though one study did find that training in VR improved speed of a maintenance task, compared to a no-training control group (Ganier et al., 2014), the comparison of interest is not between XR and no training, but XR and traditional training. Overall performance following XR-based training is neither better nor worse than performance following traditional training.

Of course, the benefit of training can be measured in more than just performance after-the-fact. Though it is beyond the quantitative scope of this meta-analysis, the use of VR in training has been shown to affect presence and immersion, as well as the psychological dimension of flow (see Lackey et al., 2016). All of these are important factors to consider in training beyond evaluating performance outcome alone.

In addition to the mean level of the effect sizes noted in Figure 2, there proved to be unusually large confidence intervals, particularly concerning the pre/post variable. These ranges of variability mean that there were an approximately equal number of effects reporting strong transfer, as there were effects indicating negative transfer. This range may be viewed as disturbing. On the principle of "*do no harm*," it is important to know that an imposed training regimen will not actually cause the trained individual to be less proficient than they would have been with a traditional training approach. At present, because of the associated degree of variability we cannot ensure that this is always so. It may be that XR provides significant performance benefits but equally we cannot rule out that such a manipulation may inhibit learning in some cases. Some studies found large positive effects of XR training, but a few did find negative effects (see Appendix C). The sources were varied enough that it was not immediately clear whether there were any commonalities between

those finding negative effects. To investigate further would require a larger body of research from which to draw conclusions. The fact that XR-based training had the same level of success as traditional training indicates that "encoding specificity" (see Godden & Baddeley, 1975) does not pose a problem for XR training. That is, the virtual environments employed in XR are clearly similar enough to the real environment that transfer can occur effectively. It can therefore be accepted that any negative transfer or otherwise poor performance after XR-based training is not a result of XR itself being an unsuitable medium, but a result of some other factors such as fidelity or individual differences.

There are several possibilities as to why this high degree of variability occurs. First, the actual training tasks represented in the summation here were highly heterogeneous. While each study examined a separate form of task, the two main categories of tasks were physical (where participants were required to practice or learn some spatial, procedural task) and cognitive (where participants acquired new information, but did not need to use it in a physical sense). Yet, even within these categories, there proved to be large variations. For example, cognitive tasks ranged from rote memorization of facts about planets to conducting simulated medical dissections. Physical tasks involved balancing skills, as well as performing a maintenance task similar to that which a factory worker might do on the job.

While overall XR training was more successful on physical tasks than cognitive tasks, this finding was not consistent. One study included in the analysis of the physical tasks used XR in order to teach the recovery of balance to stroke patients (Lee et al., 2015). Results of this particular study found that a large number of participants performed worse in the follow-up assessment. Of course it is possible that the stroke patients were deteriorating in capacity over time, that is, a declining baseline. On the other hand, a cognitive task study where participants learned mathematics showed that scores were consistently higher after the VR training intervention (Bier et al., 2018). However, due to the variability in the literature, this question needs further study.

Many studies examining physical tasks did show actual performance improvement after XR training. Several of these studies included in the analysis involved special populations such as stroke victims, with the virtual training a method to improve their physical abilities and retrain them in lost skills. Studies on healthy populations have occasionally shown that even procedural tasks such as way-finding can benefit more from virtual training than from standard training (Goldiez et al., 2007). Such findings may have been washed out by the variability of the populations examined in the included studies. The literature does not yet support a more thorough examination of task or population differential as a subpredictor. However, these shortfalls can be rectified with future research. Table 4 shows each study by task type and population examined. Here, a typical population refers to any population where participants were not selected for any specific expertise or illness, but in nearly every case was an undergraduate or university sample.

In the analysis of cognitive tasks, one study involved adults with autism, and six involved samples from the general population (one of which having age restrictions; Bier et al., 2018). The mixed tasks involved four populations from medical school, two groups of experienced technicians, and five typical populations. In the examination of physical tasks, three studies involved typical populations, although one had an age restriction (Prasertsakul et al., 2018), and four medical samples with specific ailments. The disparity in populations examined in each task type is thus fairly clear.

One other caveat with respect to the findings of the present meta-analysis is the technology used in each study. While some of the identified studies included information about the specific model of VR or AR technology used, not all did. Of the studies that did provide information, many utilized different levels of the XR platform (e.g., interactive video games, full-motion simulators). This differentiation might help explain some of the performance differences. While the examined variable of immersiveness addressed some of the differences between the degrees of virtuality, no

TABLE 4: Task Types and Populations

Citation	Task Type	Population
Andersen et al. (2016)	Mixed	Otorhinolaryngology residents
Andersen et al. (2018)	Mixed	Otorhinolaryngology residents
Bailey et al. (2017)	Mixed	Normal
Bier et al. (2018)	Cognitive	27 older and 30 younger adults
Buttussi and Chittaro (2018)	Cognitive	Normal
Chan et al. (2011)	Physical	Normal; dancers
Ganier et al. (2014)	Mixed	Normal
Gavish et al. (2015)	Mixed	Experienced technicians
Gerson and Van Dam (2003)	Mixed	Medical residents
Gonzalez-Franco et al. (2016)	Mixed	Normal
Hamblin (2005)	Mixed	Normal
Kober et al. (2013)	Physical	Population: spatial disorientation
Lee et al. (2015)	Physical	Stroke population
Ma et al. (2011)	Physical	Parkinson's population
Macchiarella (2004)	Cognitive	Normal
Madden et al. (2018)	Cognitive	Normal
Martín-Gutiérrez et al. (2010)	Mixed	Normal
Prasertsakul et al. (2018)	Physical	Adults age 40–60
Rose et al. (2000)	Physical	Normal
Smith et al. (2014)	Cognitive	Autistic adults
Valimont et al. (2007)	Cognitive	Normal
Wang et al. (2014)	Mixed	Medical students
Webel et al. (2013)	Mixed	Experienced technicians
Whitmer et al. (2019)	Cognitive	Normal
Yang et al. (2008)	Physical	Stroke population

such distinctions can be made in the case of difference in quality. Not all XR technologies are created equal, and to compare two studies using different XR systems may even be inappropriate to some degree. Disparate technology makes it difficult to determine direct effects of each training intervention with so few studies being suitable for analysis (see Hancock & Hoffman, 2015).

Finally, it is important to reiterate that results of the present meta-analysis, as are results from all such analyses, are constrained by the limits and extent of the existing body of literature. There were insufficient studies to examine many of the factors related to either the task or the learner. Nor was there enough information to fully examine the subject of training transfer from XR, in sufficient depth so that reliable conclusions can be reached. Further, there were insufficient numbers of studies on AR to analyze its affects as distinct from those of VR. The "count of studies" columns in Tables 2 and 3 reveal the surprising paucity of research in this vital area. This then is not simply a case of "more research is needed," but a case in which more *diverse* research is needed. This may well be an issue involved with the impetus and constituencies to fund such research. Many organizations "sell" training but frequently do not provide robust quantitative evidence of the value of that

training in their promotional literature. The goal here is not to simply point out shortcomings in the existing field of research, but to identify those points where future research should be conducted in order to best examine quantifiable antecedents of training efficacy in XR.

LIMITATIONS AND CONCLUSIONS

Although the current literature is surprisingly sparse, posing some limitations for the present meta-analysis, our present results are not inconclusive. However, due to this present paucity, certain analyses cannot be effectively performed. For example, it might be useful and insightful to consider the ways in which the variables associated specifically with training per se are nested within those particularly focused upon the state of the technology in each of AR, VR, and XR. To date, insufficient information has been collected upon these combinations such that we may be confident of the outcome. As with this and other current shortfalls, we have highlighted important gaps in the literature that need to be addressed if this important field is to move forward. For any meaningful effects to be determined from future comprehensive studies (meta or otherwise), the questions raised in our present work must be addressed. One of the most pressing areas needing more research is individual differences; soldiers are a very different population from elderly stroke victims. Studies are needed that enable performance comparisons *between* populations by holding variables such as simulation platform, task, and performance measures constant and studying performance by different population groups. This is essential for understanding which differences in results can be attributed directly to the effect of population factors such as experience or comfort with XR technology.

The second critical, but varying influence is the technology in use. At present there is a wide range of VR headsets and simulated environments used in studies. These are potentially of very different quality, although quality was rarely reported in the methods of each study. "Fidelity" is a word which was used with some regularity, but in the absence of consideration of how the affordances of a virtual environment or of a simulation used met the needs of those being trained. In this endeavor, a useful set of dimensions have already been defined in Extent of World Knowledge, Reproduction Fidelity, and Extent of Presence Metaphor (see Milgram et al., 1995). Researchers would do well to complete any simulation studies multiple times with different display technologies, especially when the difference in quality is already quantified (Hancock et al., 2015).

The third area which requires more specification is the task designation. It may be that training sessions were entirely different as they were meant to train unique tasks. Indeed, the present body of evidence suggests that not all task types benefit equally and the physical/cognitive division may not be the most critical one. What makes tasks amenable to consistently efficacious XR training is presently not well understood. Indeed, all three of these factors, and any interactions between them, make it difficult to determine the effect of virtual training on later performance. What the literature does support, currently, is the fact that XR training has similar performance outcomes to traditional training. In the absence of any significant differences between XR and traditional training, there is a bright future in considering the many benefits that XR training promises.

APPENDIX A: Definitions of AR, VR, and MR

Type of Simulated Reality	Definition	Source
Augmented reality	Any system that has the following three characteristics: 1. Combines real and virtual 2. Is interactive in real time 3. Is registered in three dimensions	Azuma (1997)
	All cases in which the display of an otherwise real environment is augmented by means of virtual (computer graphic) objects	Milgram and Kishino (1994)
	Augmenting natural feedback to the operator with simulated cues	Milgram et al. (1995)
	The enhancement of the real world by a virtual world, which subsequently provides additional information	Feiner et al. (1993)
	AR displays are those in which the image is of a primarily real environment, which is enhanced, or augmented, with computer-generated imagery	Drascic and Milgram (1996)
Virtual reality	VR can be defined as a three-dimensional computer-generated environment, updating in real time, and allowing human interaction through various input/output devices	Boud et al. (1999)
	Strictly the term virtual reality describes something that is "real in effect although not in fact" [virtual] and which "can be considered capable of being considered fact for some purposes" [reality]. A virtual environment, put simply, is an environment other than the one in which the participant is actually present; more usefully it is a computer-generated model, where a participant can interact intuitively in real time with the environment	Wilson (1997)
	A "virtual reality" is defined as a real or simulated environment in which a perceiver experiences telepresence	Steuer (1992)
	Virtual reality is an alternate world filled with computer-generated images that respond to human movements. These simulated environments are usually visited with the aid of an expensive data suit which features stereophonic video goggles and fiber-optic data gloves	Greenbaum (1992)
	It is a new emergent mode of reality in its own right, that comes together with actual reality to construct an extended world of human experience	Yoh (2001)
	Virtual reality is a technology that convinces the participant that he or she is actually in another place by substituting the primary sensory input with data produced by a computer	Heim (1998)
	A computer-generated display that allows or compels the user (or users) to have a sense of being present in an environment other than the one they are actually in, and to interact with that environment	Schroeder (1996)
Mixed reality	Mixed reality refers to the class of all displays in which there is some combination of a real environment and virtual reality	Drascic and Milgram (1996)
	Mixed reality environment is one in which real-world and virtual world objects are presented together within a single display	Milgram et al. (1995)

APPENDIX B: Effect Sizes by Study

Source	*N*	Task Type	Associated Variable	Effect Size
Andersen et al. (2018)	37	Mixed	Immersiveness	−0.54
Andersen et al. (2018)	37	Mixed	Immersiveness	−0.55
Andersen et al. (2016)	40	Mixed	Immersiveness	−1.40
Andersen et al. (2016)	40	Mixed	Immersiveness	−1.12
Andersen et al. (2016)	40	Mixed	Immersiveness	−0.92
Andersen et al. (2016)	20	Mixed	Pre/post	0.54
Andersen et al. (2016)	20	Mixed	Pre/post	0.07
Andersen et al. (2016)	20	Mixed	Pre/Ppost	0.47
Bailey et al. (2017)	83	Mixed	Immersiveness	0.27
Bailey et al. (2017)	83	Mixed	Immersiveness	0.06
Bier et al. (2018)	27	Cognitive	Task difficulty	−0.92
Bier et al. (2018)	30	Cognitive	Task difficulty	−1.28
Bier et al. (2018)	27	Cognitive	Task difficulty	0.28
Bier et al. (2018)	30	Cognitive	Task difficulty	0.19
Bier et al. (2018)	27	Cognitive	Task difficulty	−0.53
Bier et al. (2018)	30	Cognitive	Task difficulty	0.88
Bier et al. (2018)	27	Cognitive	Task difficulty	−0.05
Bier et al. (2018)	30	Cognitive	Task difficulty	0.26
Bier et al. (2018)	57	Cognitive	Age	3.83
Bier et al. (2018)	57	Cognitive	Age	0.61
Bier et al. (2018)	57	Cognitive	Age	−0.08
Bier et al. (2018)	57	Cognitive	Age	−0.23
Bier et al. (2018)	57	Cognitive	Age	−2.07
Bier et al. (2018)	57	Cognitive	Age	−0.77
Bier et al. (2018)	57	Cognitive	Age	−1.10
Bier et al. (2018)	57	Cognitive	Age	−0.87
Bier et al. (2018)	14	Cognitive	Pre/post	0.98
Bier et al. (2018)	13	Cognitive	Pre/post	0.06
Bier et al. (2018)	15	Cognitive	Pre/post	1.42
Bier et al. (2018)	15	Cognitive	Pre/post	0.85
Bier et al. (2018)	14	Cognitive	Pre/post	0.58
Bier et al. (2018)	13	Cognitive	Pre/post	0.19
Bier et al. (2018)	15	Cognitive	Pre/Post	−0.51
Bier et al. (2018)	15	Cognitive	Pre/post	−0.54
Bier et al. (2018)	14	Cognitive	Pre/post	−1.18
Bier et al. (2018)	13	Cognitive	Pre/post	−1.69
Bier et al. (2018)	15	Cognitive	Pre/Post	−1.77
Bier et al. (2018)	15	Cognitive	Pre/Post	−1.31
Bier et al. (2018)	14	Cognitive	Pre/post	−0.55

(continued)

Source	*N*	Task Type	Associated Variable	Effect Size
Bier et al. (2018)	13	Cognitive	Pre/Post	−0.18
Bier et al. (2018)	15	Cognitive	Pre/post	−0.39
Bier et al. (2018)	15	Cognitive	Pre/post	−0.08
Buttussi and Chittaro (2018)	96	Cognitive	Immersiveness	0.12
Buttussi and Chittaro (2018)	96	Cognitive	Immersiveness	1.00
Buttussi and Chittaro (2018)	96	Cognitive	Immersiveness	−0.88
Buttussi and Chittaro (2018)	96	Cognitive	Pre/post	6.26
Buttussi and Chittaro (2018)	96	Cognitive	Pre/post	5.62
Buttussi and Chittaro (2018)	96	Cognitive	Pre/Post	6.50
Chan et al. (2011)	8	Physical	Immersiveness	1.65
Chan et al. (2011)	4	Physical	Pre/post	−2.07
Ganier et al. (2014)	42	Mixed	Immersiveness	1.17
Ganier et al. (2014)	42	Mixed	Immersiveness	−1.14
Gavish et al. (2015)	20	Mixed	Immersiveness	0.28
Gavish et al. (2015)	20	Mixed	Immersiveness	0.28
Gavish et al. (2015)	20	Mixed	Immersiveness	−0.21
Gavish et al. (2015)	20	Mixed	Immersiveness	0.00
Gerson and Van Dam (2003)	16	Mixed	Immersiveness	−1.12
Gonzalez-Franco et al. (2016)	24	Mixed	Immersiveness	−0.12
Gonzalez-Franco et al. (2016)	24	Mixed	Immersiveness	−0.58
Hamblin (2005)	18	Mixed	Immersiveness	0.06
Hamblin (2005)	18	Mixed	Immersiveness	−1.67
Hamblin (2005)	18	Mixed	Immersiveness	−2.30
Hamblin (2005)	18	Mixed	Immersiveness	−0.34
Hamblin (2005)	18	Mixed	Immersiveness	−3.70
Hamblin (2005)	18	Mixed	Immersiveness	−2.13
Martín-Gutiérrez et al. (2010)	49	Mixed	Immersiveness	0.63
Martín-Gutiérrez et al. (2010)	49	Mixed	Immersiveness	0.51
Martín-Gutiérrez et al. (2010)	25	Mixed	Pre/post	1.02
Martín-Gutiérrez et al. (2010)	25	Mixed	Pre/post	1.27
Kober et al. (2013)	11	Physical	Pre/post	0.21
Kober et al. (2013)	11	Physical	Pre/post	2.92
Lee et al. (2015)	24	Physical	Immersiveness	0.07
Lee et al. (2015)	24	Physical	Immersiveness	0.07
Lee et al. (2015)	24	Physical	Immersiveness	0.04
Lee et al. (2015)	24	Physical	Immersiveness	0.04
Lee et al. (2015)	24	Physical	Immersiveness	0.25
Lee et al. (2015)	24	Physical	Immersiveness	0.26
Lee et al. (2015)	24	Physical	Immersiveness	0.03
Lee et al. (2015)	24	Physical	Immersiveness	0.03

(continued)

Source	*N*	Task Type	Associated Variable	Effect Size
Lee et al. (2015)	24	Physical	Immersiveness	0.49
Lee et al. (2015)	12	Physical	Pre/post	−0.38
Lee et al. (2015)	12	Physical	Pre/post	−0.38
Lee et al. (2015)	12	Physical	Pre/post	−0.42
Lee et al. (2015)	12	Physical	Pre/post	−0.42
Lee et al. (2015)	12	Physical	Pre/post	−0.49
Lee et al. (2015)	12	Physical	Pre/post	−0.49
Lee et al. (2015)	12	Physical	Pre/post	−0.51
Lee et al. (2015)	12	Physical	Pre/pst	−0.51
Lee et al. (2015)	12	Physical	Pre/post	1.41
Ma et al. (2011)	33	Physical	Immersiveness	−0.73
Ma et al. (2011)	33	Physical	Immersiveness	0.45
Ma et al. (2011)	33	Physical	Immersiveness	−0.24
Ma et al. (2011)	33	Physical	Immersiveness	0.00
Ma et al. (2011)	33	Physical	Immersiveness	−0.28
Ma et al. (2011)	33	Physical	Immersiveness	−0.53
Ma et al. (2011)	33	Physical	Immersiveness	−0.4
Ma et al. (2011)	33	Physical	Immersiveness	0.46
Ma et al. (2011)	33	Physical	Immersiveness	0.10
Ma et al. (2011)	33	Physical	Immersiveness	−0.76
Ma et al. (2011)	33	Physical	Immersiveness	−0.16
Ma et al. (2011)	33	Physical	Immersiveness	−0.16
Ma et al. (2011)	33	Physical	Immersiveness	−0.02
Ma et al. (2011)	33	Physical	Immersiveness	0.26
Ma et al. (2011)	33	Physical	Immersiveness	−0.10
Ma et al. (2011)	33	Physical	Immersiveness	−0.08
Ma et al. (2011)	33	Physical	Immersiveness	−0.24
Ma et al. (2011)	33	Physical	Immersiveness	−0.73
Ma et al. (2011)	33	Physical	Immersiveness	−0.42
Ma et al. (2011)	33	Physical	Pre/post	−0.71
Ma et al. (2011)	33	Physical	Pre/post	0.61
Ma et al. (2011)	33	Physical	Pre/post	−0.3
Ma et al. (2011)	33	Physical	Pre/ppost	0.38
Ma et al. (2011)	33	Physical	Pre/post	0.10
Ma et al. (2011)	33	Physical	Pre/post	−0.25
Ma et al. (2011)	33	Physical	Pre/post	0.18
Ma et al. (2011)	33	Physical	Pre/post	−0.10
Ma et al. (2011)	33	Physical	Pre/post	−0.44
Ma et al. (2011)	33	Physical	Pre/post	−0.13
Ma et al. (2011)	33	Physical	Pre/post	−0.33

(continued)

Source	*N*	Task Type	Associated Variable	Effect Size
Ma et al. (2011)	33	Physical	Pre/post	0.12
Ma et al. (2011)	33	Physical	Pre/post	0.04
Ma et al. (2011)	33	Physical	Pre/post	0.03
Ma et al. (2011)	33	Physical	Pre/post	0.03
Ma et al. (2011)	33	Physical	Pre/post	−0.18
Ma et al. (2011)	33	Physical	Pre/post	0.00
Ma et al. (2011)	33	Physical	Pre/post	−0.45
Ma et al. (2011)	33	Physical	Pre/post	−0.33
Macchiarella (2004)	96	Cognitive	Immersiveness	−0.05
Macchiarella (2004)	96	Cognitive	Immersiveness	−0.44
Macchiarella (2004)	96	Cognitive	Immersiveness	−0.82
Macchiarella (2004)	96	Cognitive	Immersiveness	−0.39
Macchiarella (2004)	96	Cognitive	Immersiveness	−0.76
Madden et al. (2018)	172	Cognitive	Immersiveness	0.11
Madden et al. (2018)	172	Cognitive	Immersiveness	0.24
Madden et al. (2018)	56	Cognitive	Pre/post	1.48
Prasertsakul et al. (2018)	8	Physical	Immersiveness	0.61
Prasertsakul et al. (2018)	8	Physical	Immersiveness	0.23
Prasertsakul et al. (2018)	8	Physical	Immersiveness	0.90
Prasertsakul et al. (2018)	8	Physical	Immersiveness	0.23
Prasertsakul et al. (2018)	8	Physical	Immersiveness	−0.01
Prasertsakul et al. (2018)	8	Physical	Immersiveness	−0.05
Prasertsakul et al. (2018)	8	Physical	Immersiveness	0.07
Prasertsakul et al. (2018)	8	Physical	Immersiveness	−0.02
Prasertsakul et al. (2018)	8	Physical	Immersiveness	0.93
Prasertsakul et al. (2018)	8	Physical	Immersiveness	1.37
Prasertsakul et al. (2018)	4	Physical	Pre/post	0.11
Prasertsakul et al. (2018)	4	Physical	Pre/post	0.42
Prasertsakul et al. (2018)	4	Physical	Pre/post	−0.57
Prasertsakul et al. (2018)	4	Physical	Pre/post	−0.32
Prasertsakul et al. (2018)	4	Physical	Pre/post	−0.26
Prasertsakul et al. (2018)	4	Physical	Pre/post	0.11
Prasertsakul et al. (2018)	4	Physical	Pre/Post	−0.15
Prasertsakul et al. (2018)	4	Physical	Pre/post	−0.03
Prasertsakul et al. (2018)	4	Physical	Pre/post	−0.22
Prasertsakul et al. (2018)	4	Physical	Pre/post	0.32
Rose et al. (2000)	100	Physical	Immersiveness	0.17
Smith et al. (2014)	26	Cognitive	Immersiveness	0.72
Smith et al. (2014)	16	Cognitive	Pre/post	0.56
Valimont et al. (2007)	32	Cognitive	Immersiveness	0.45

(continued)

Source	*N*	Task Type	Associated Variable	Effect Size
Valimont et al. (2007)	32	Cognitive	Immersiveness	0.69
Valimont et al. (2007)	32	Cognitive	Immersiveness	1.01
Valimont et al. (2007)	32	Cognitive	Immersiveness	0.51
Valimont et al. (2007)	32	Cognitive	Immersiveness	0.59
Valimont et al. (2007)	32	Cognitive	Immersiveness	0.67
Wang et al. (2014)	16	Mixed	Immersiveness	0.35
Wang et al. (2014)	16	Mixed	Immersiveness	−1.51
Webel et al. (2013)	20	Mixed	Immersiveness	−1.51
Whitmer et al. (2019)	41	Cognitive	Immersiveness	−0.89
Yang et al. (2008)	20	Physical	Immersiveness	1.08
Yang et al. (2008)	20	Physical	Immersiveness	0.67
Yang et al. (2008)	20	Physical	Immersiveness	−0.99
Yang et al. (2008)	20	Physical	Immersiveness	−0.89
Yang et al. (2008)	20	Physical	Immersiveness	0.61
Yang et al. (2008)	20	Physical	Immersiveness	1.00
Yang et al. (2008)	20	Physical	Immersiveness	0.47
Yang et al. (2008)	20	Physical	Immersiveness	0.13

APPENDIX C: Forest plot of individual studies by task type/predictor clusters

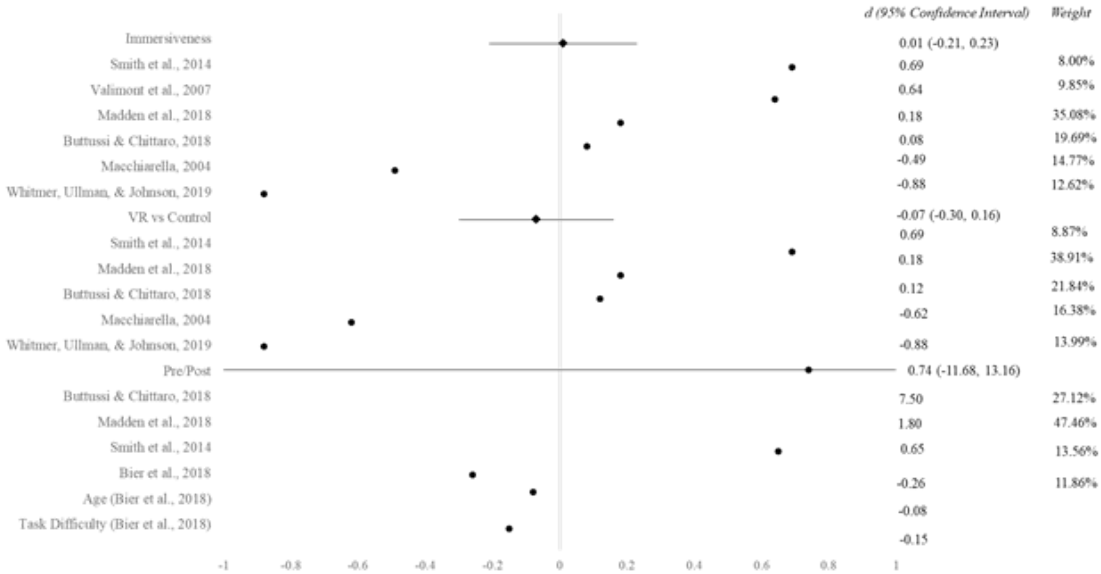

Figure 4. Studies involving cognitive tasks.

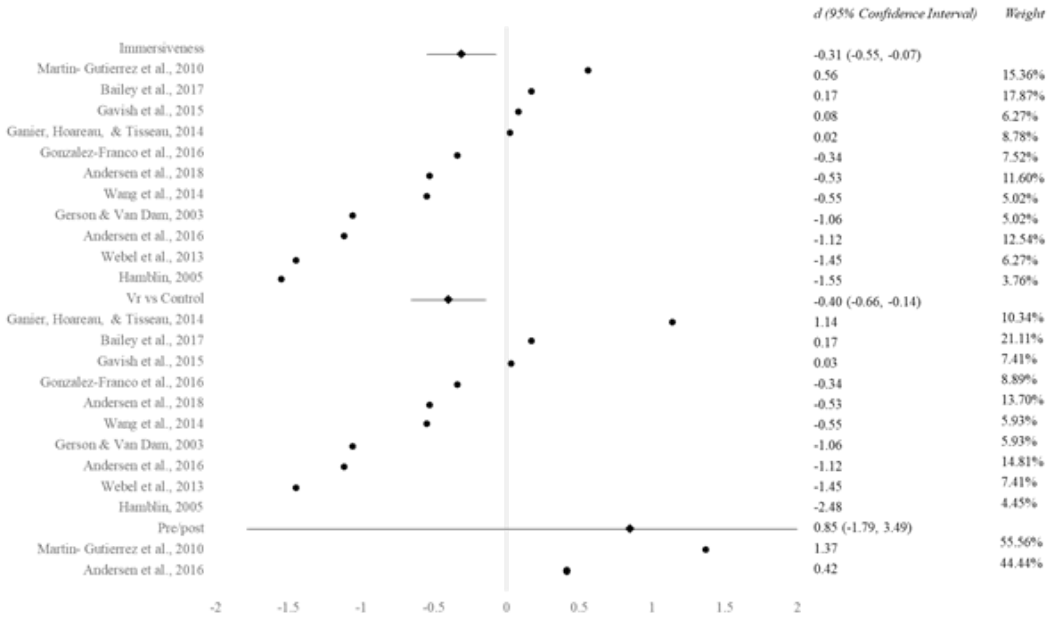

Figure 5. Studies involving mixed tasks.

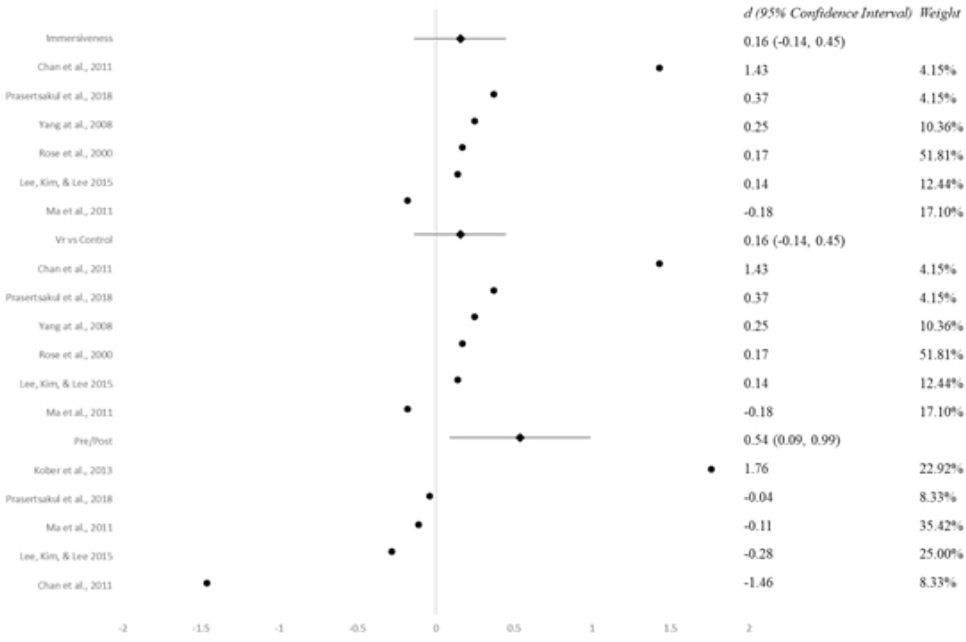

Figure 6. Studies involving physical tasks.

KEY POINTS

- Performance after training in VR/AR is generally comparable to performance after training in a traditional setting.
- The population being trained, and task being trained upon, can affect whether VR/AR is an effective medium for training.
- The field of research is too disparate to determine precisely which factors contribute to better training transfer from VR/AR.

ORCID iD

Alexandra D. Kaplan https://orcid.org/0000-0003-0051-0150
Mica Endsley https://orcid.org/0000-0002-2359-947X
P. A. Hancock https://orcid.org/0000-0002-4936-066X

REFERENCES

References marked with an asterisk indicate studies included in the meta-analysis.

*Andersen, S. A. W., Konge, L., & Sørensen, M. S. (2018). The effect of distributed virtual reality simulation training on cognitive load during subsequent dissection training. *Medical Teacher*, *40*, 684–689.

*Andersen, S. A. W., Mikkelsen, P. T., Konge, L., Cayé-Thomasen, P., & Sørensen, M. S. (2016). Cognitive load in mastoidectomy skills training: Virtual reality simulation and traditional dissection compared. *Journal of Surgical Education*, *73*, 45–50.

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, *6*, 355–385.

*Bailey, SKT., Johnson, C. I., Schroeder, BL., & Marraffino, MD. (2017). *Using virtual reality for training maintenance procedures* [Conference session]. Proceedings of the Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), Orlando, FL.

*Bier, B., Ouellet, Émilie., & Belleville, S. (2018). Computerized attentional training and transfer with virtual reality: Effect of age and training type. *Neuropsychology*, *32*, 597–614.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*, 1065–1105.

Boud, AC., Haniff, DJ., Baber, C., & Steiner, SJ. (1999). *Virtual reality and augmented reality as a training tool for assembly tasks* [Conference session]. IEEE International Conference on Information Visualization (Cat. No. PR00210), London, England, 32–36.

*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*, 1063–1076.

*Chan, J. C. P., Leung, H., Tang, J. K. T., & Komura, T. (2011). A virtual reality dance training system using motion capture technology. *IEEE Transactions on Learning Technologies*, *4*, 187–195.

Drascic, D., & Milgram, P. (1996, April). Perceptual issues in augmented reality. *Stereoscopic Displays and Virtual Reality Systems III*, *2653*, 123–135.

Feiner, S., Macintyre, B., & Seligmann, D. (1993). Knowledge-based augmented reality. *Communications of the ACM*, *36*, 53–62.

Fletcher, JD., Belanich, J., Moses, F., Fehr, A., & Moss, J. (2017). *Effectiveness of augmented reality & augmented virtuality* [Conference session]. Proceedings of the MODSIMWORLD Conference, Virgina Beach, VA.

*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*, 828–843.

*Gavish, N., Gutiérrez, T., Webel, S., Rodríguez, J., Peveri, M., Bockholt, U., & Tecchia, F. (2015). Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interactive Learning Environments*, *23*, 778–798.

*Gerson, L. B., & Van Dam, J. (2003). A prospective randomized trial comparing a virtual reality simulator to bedside teaching for training in sigmoidoscopy. *Endoscopy*, *35*, 569–575.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*, 325–331.

Goldiez, B. F., Ahmad, A. M., & Hancock, P. A. (2007). Effects of augmented reality display settings on human wayfinding performance. *IEEE Transactions on Systems, Man and Cybernetics, Part C*, *37*, 839–845.

*Gonzalez-Franco, M., Cermeron, J., Li, K., Pizarro, R., Thorn, J., Hutabarat, W., Hannah, P., Tiwari, A., & Bermell-Garcia, P. (2016). Immersive augmented reality training for complex manufacturing scenarios. *arXiv:1602.01944*.

Greenbaum, P. (1992, March). The lawnmower man. *Film and Video*, *9*, 58–62.

*Hamblin, C. J. (2005) *Learning transfer in virtual environments* [Unpublished doctoral dissertation]. Witchita State University

Hancock, P. A. (2009). The future of simulation. In D. Vincenzi, J. Wise, M. Mouloua, & P. A. Hancock (Eds.), *Human Factors in Simulation and Training* (pp. 169–186). CRC Press.

Hancock, P. A. (2017). On bored to Mars. *Journal of Astro-Sociology*, *2*, 103–120.

Hancock, P. A., & Hoffman, R. R. (2015). Keeping up with intelligent technology. *IEEE Intelligent Systems*, *30*, 62–65.

Hancock, P. A., Sawyer, B. D., & Stafford, S. (2015). The effects of display size on performance. *Ergonomics*, *58*, 337–354.

Haque, S., & Srinivasan, S. (2006). A meta-analysis of the training effectiveness of virtual reality surgical simulators. *IEEE Transactions on Information Technology in Biomedicine*, *10*, 51–58.

Heim, M. (1998). *Virtual Realism*. Oxford.

Holding, D. H. (Ed.). (1989). *Human Skills*. John Wiley & Sons Incorporated.

*Kober, S., Wood, G., Hofer, D., Kreuzig, W., Kiefer, M., & Neuper, C. (2013). Virtual reality in neurologic rehabilitation of spatial disorientation. *Journal of Neuroengineering and Rehabilitation*, *10*, 17–30.

Kozak, J. J., Hancock, P. A., Arthur, E. J., & Chrysler, S. T. (1993). Transfer of training from virtual reality. *Ergonomics*, *36*, 777–784.

Lackey, S. J., Salcedo, J. N., Szalma, J. L., & Hancock, P. A. (2016). The stress and workload of virtual reality training: The effects of presence, immersion and flow. *Ergonomics*, *59*, 1060–1072.

*Lee, H. Y., Kim, Y. L., & Lee, S. M. (2015). Effects of virtual reality-based training and task-oriented training on balance performance in stroke patients. *Journal of Physical Therapy Science*, *27*, 1883–1888.

*Ma, H.-I., Hwang, W.-J., Fang, J.-J., Kuo, J.-K., Wang, C.-Y., Leong, I.-F., & Wang, T.-Y. (2011). Effects of virtual reality training on functional reaching movements in people with Parkinson's disease: a randomized controlled pilot trial. *Clinical Rehabilitation*, *25*, 892–902.

*Macchiarella, N. D. (2004) *Effectiveness of video-based augmented reality as a learning paradigm for aerospace maintenance training* [Unpublished doctoral dissertation]. Nova Southeastern University.

*Madden, JH., Won, AS., Schuldt, JP., Kim, B., Pandita, S., Sun, Y., Stone, TJ., & Holmes, NG. (2018). *Virtual reality as a teaching tool for moon phases and beyond* [Conference session].

Proceedings of the Physics Education Research Conference, Washington, D.C.
Marras, W. S., & Hancock, P. A. (2014). Putting mind and body back together: A human-systems approach to the integration of the physical and cognitive dimensions of task design and operations. *Applied Ergonomics*, *45*, 55–60.
*Martín-Gutiérrez, J., Luís Saorín, J., Contero, M., Alcañiz, M., Pérez-López, D. C., & Ortega, M. (2010). Design and validation of an augmented book for spatial abilities development in engineering students. *Computers & Graphics*, *34*, 77–91.
Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *Institute of Electronics, Information, and Communication Engineers Transactions on Information and Systems*, *77*, 1321–1329.
Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1995, December). Augmented reality: A class of displays on the reality-virtuality continuum. In *Telemanipulator and telepresence technologies* (Vol. 2351, pp. 282–293).
Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, *151*, 264–269.
*Prasertsakul, T., Kaimuk, P., Chinjenpradit, W., Limroongreungrat, W., & Charoensuk, W. (2018). The effect of virtual reality-based balance training on motor learning and postural control in healthy adults: A randomized preliminary study. *Biomedical Engineering Online*, *17*, 124–141.
Rantanen, E. M., & Talleur, D. A. (2005, September). Incremental transfer and cost effectiveness of groundbased flight trainers in University aviation programs. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*, 764–768.
Roscoe, S. N. (1971). Incremental transfer effectiveness. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *13*, 561–567.
*Rose, F. D., Attree, E. A., Brooks, B. M., Parslow, D. M., & Penn, P. R. (2000). Training in virtual environments: Transfer to real world tasks and equivalence to real task training. *Ergonomics*, *43*, 494–511.
Salas, E., Bowers, C. A., & Rhodenizer, L. (1998). It is not how much you have but how you use it: Toward a rational use of simulation to support aviation training. *The International Journal of Aviation Psychology*, *8*, 197–208.
Schroeder, R. (1996). *Possible worlds: The social dynamic of virtual reality technologies*. Westview Press.
*Smith, M. J., Ginger, E. J., Wright, K., Wright, M. A., Taylor, J. L., Humm, L. B., Olsen, D. E., Bell, M. D., & Fleming, M. F. (2014). Virtual reality job interview training in adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *44*, 2450–2463.
Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, *42*, 73–93.
Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352–373.
*Valimont, R. B., Gangadharan, S., Vincenzi, D., & Majoros, A. (2007). The effectiveness of augmented reality as a facilitator of information acquisition in aviation maintenance applications. *Journal of Aviation/Aerospace Education & Research*, *16*, 35–43.
*Wang, Z., Ni, Y., Zhang, Y., Jin, X., Xia, Q., & Wang, H. (2014). Laparoscopic varicocelectomy: Virtual reality training and learning curve. *JSLS : Journal of the Society of Laparoendoscopic Surgeons*, *18*, e2014.00258–7.
*Webel, S., Bockholt, U., Engelke, T., Gavish, N., Olbrich, M., & Preusche, C. (2013). An augmented reality training platform for assembly and maintenance skills. *Robotics and Autonomous Systems*, *61*, 398–403.
*Whitmer, D. E., Ullman, D., & Johnson, C. I. (2019). Virtual reality training improves real-world performance on a speeded task. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 63, No. 1, pp. 1218–1222). SAGE Publications.
Wilson, J. R. (1997). Virtual environments and ergonomics: Needs and opportunities. *Ergonomics*, *40*, 1057–1077.
*Yang, Y.-R., Tsai, M.-P., Chuang, T.-Y., Sung, W.-H., & Wang, R.-Y. (2008). Virtual reality-based training improves community ambulation in individuals with stroke: A randomized controlled trial. *Gait & Posture*, *28*, 201–206.
Yoh, MS. (2001). *The reality of virtual reality* [Conference session]. Proceedings Seventh International Conference on Virtual Systems and Multimedia, Berkeley, CA, IEEE. 666–674.

Alexandra D. Kaplan is currently a graduate student at the University of Central Florida. She obtained a master's in applied experimental and human factors psychology in 2019 from the University of Central Florida.

Jessica Cruit is currently a research scholar at the University of Central Florida. She received her PhD in human factors from Embry-Riddle Aeronautical University in 2016.

Mica Endsley is the president of SA Technologies, Inc., and Chair, Air Force Scientific Advisory Board Study Panel on 21st Century Training and Education Technologies. She received her PhD in industrial and systems engineering from the University of Southern California in 1990.

Suzanne M. Beers is the OSD Test and Evaluation Portfolio Manager at the MITRE Corporation and a member of the Air Force Scientific Advisory Board. She received her PhD in electrical engineering from the Georgia Institute of Technology in 1996.

Ben D. Sawyer is currently an assistant professor and lab director at the University of Central Florida. He received a PhD in applied experimental and human factors psychology at the University of Central Florida in 2015.

P. A. Hancock is currently a Provost Distinguished Research Professor at the University of Central Florida, and member of the Air Force Scientific Advisory Board. He received his PhD in human performance from the University of Illinois in 1983 and a DSc in human–machine systems in 2001 from the Loughborough University in Loughborough, England.

Date received: August 25, 2019
Date accepted: January 7, 2020