Computers in Human Behavior 121 (2021) 106808

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: http://www.elsevier.com/locate/comphumbeh

Full length article

A meta-analysis of virtual reality training programs

Matt C. Howard [a,*], Melissa B. Gutworth [b], Rick R. Jacobs [c]

[a] The University of South Alabama, Mitchell College of Business, USA
[b] Montclair State University, Feliciano School of Business, USA
[c] PSI Services & Pennsylvania State University, Department of Psychology, USA

ARTICLE INFO

Keywords:
Virtual reality
Head mounted display
Computer-based training
Training and development
Learning
Meta-analysis

ABSTRACT

Virtual reality (VR) is the three-dimensional digital representation of a real or imagined space with interactive capabilities. The application of VR for organizational training purposes has been surrounded by much fanfare; however, mixed results have been provided for the effectiveness of VR training programs, and the attributes of effective VR training programs are still unknown. To address these issues, we perform a meta-analysis of controlled experimental studies that tests the effectiveness of VR training programs. We obtain an estimate of the overall effectiveness of VR training programs, and we identify features of VR training programs that systematically produce improved results. Our meta-analytic findings support that VR training programs produce better outcomes than tested alternatives. The results also show that few moderating effects were significant. The applied display hardware, input hardware, and inclusion of game attributes had non-significant moderating effects; however, task-technology fit and aspects of the research design did influence results. We suggest that task-technology fit theory is an essential paradigm for understanding VR training programs, and no set of VR technologies is "best" across all contexts. Future research should continue studying all types of VR training programs, and authors should more strongly integrate research and theory on employee training and development.

For several decades, organizations have applied virtual reality (VR) to achieve training and development goals across a variety of occupations, demonstrating that the medium can improve the knowledge, skills, abilities, and other characteristics (KSAOs) of employees (Alaraj et al., 2011; Gigante, 1993; Seymour et al., 2002). Traditionally, VR has been applied to develop KSAOs that are dangerous or costly to develop via other approaches. Early practitioners used the medium to train pilots, paratroopers, and other military personnel (Julier et al., 1999; Moshell, 1993; Satava & Jones, 1996), whereas later practitioners and researchers have overwhelmingly applied VR to train surgeons (Alaker et al., 2016; Moglia et al., 2016; Vaughan et al., 2016). Due to the recent technological developments surrounding VR, however, modern practitioners and researchers have begun to broaden the applications of VR for training purposes. Today, VR is used to develop KSAOs relevant to engineering (Doyle et al., 2015), computer science, (Su & Cheng, 2013), chemistry (Merchant et al., 2013), marketing (Cheng & Wang, 2011), and many other domains (Barata et al., 2015; Ganier et al., 2014; Gavish et al., 2015).

While most practitioners and researchers have expressed favorable reactions to VR for training purposes, many studies have found the medium to be equally or even less effective than comparable training approaches (Camp et al., 2010; Tergas et al., 2013; Våpenstad et al., 2017). For instance, Jensen et al. (2014) found that a traditional training program was more effective than VR for training thoracoscopic lobectomy skills, and Bertram et al. (2015) discovered similar results for the training of police officers. Further, relatively little is known regarding the best approaches for applying the medium for training purposes. The fidelity of VR is often assumed to be a benefit for training, and thereby VR technologies that provide greater fidelity are assumed to improve training outcomes (e.g., head-mounted displays [HMDs]) (Kruglikova et al., 2010; McMahan et al., 2012); however, this notion and similar others have not been fully supported. While individual studies have provided notable inferences and qualitative reviews have uncovered important research questions, a quantitative review has yet to provide comprehensive evidence that VR is indeed more effective than comparable alternatives. Likewise, a quantitative review has yet to produce substantial support for the characteristics of VR that produce more beneficial training outcomes across a variety of applications and

* Corresponding author. 5811 USA Drive S., Rm. 337, Mitchell College of Business, University of South Alabama, Mobile, AL, 36688, USA.
E-mail addresses: mhoward@southalabama.edu (M.C. Howard), gutworthm@montclair.edu (M.B. Gutworth), rrj@psu.edu (R.R. Jacobs).

https://doi.org/10.1016/j.chb.2021.106808
Received 23 September 2020; Received in revised form 30 March 2021; Accepted 2 April 2021
Available online 7 April 2021

contexts.

Given these concerns, the current article reports a meta-analysis of controlled experimental studies that tests the efficacy of VR to develop KSAOs.[1] We estimate the effectiveness of VR training programs relative to alternative comparisons, and we determine which attributes of VR training programs systematically produce larger effects by assessing moderating influences. These attributes include the display hardware (immersive displays vs. computer monitors), input hardware (specialized input vs. keyboard & mouse), software (game attributes), task-technology fit, and study design (participant population, type of control group). In studying hardware and software, we assess whether theories regarding fidelity and motivation are useful for understanding the benefits of VR for training success. We also test whether the effectiveness of VR training programs depends on the developed KSAO. By applying experiential learning theory (Kolb et al., 2001; Kolb & Kolb, 2009), we hypothesize that VR is more effective at developing complex KSAOs (opposed to basic) as well as skills (opposed to knowledge), and the results can identify the most ideal training applications for VR. Lastly, we provide a direct assessment of task-technology fit theory by studying task-technology fit.

Achieving these goals provides many benefits to both practice and research. Regarding practice, the current article demonstrates whether VR is an effective training medium or whether it should be reconsidered until further technological breakthroughs. We also provide guidance regarding the best applications of the technology – both the attributes of the training program as well as the context (e.g., targeted KSAOs). Regarding research, the current article identifies theoretical perspectives that may be most essential to future investigations of VR training programs. If immersive hardware, for instance, is supported to produce better training outcomes, then theories involving fidelity may be necessary for identifying further improvements to the training medium. Likewise, if VR training programs more effectively develop skills compared to other outcomes, then experiential learning theory may too be insightful in the future study of VR training program. We also identify strengths and weaknesses of the current literature by analyzing study characteristics, such as the use of specific populations, training evaluation designs, and measurement approaches. Thus, we not only provide theoretical insights but also methodological insights into the current and future study of VR training programs.

The most important contribution of the current article, however, may be the synthesis of prior VR training research with modern training perspectives. The majority of research on VR training programs is not conducted by scholars in fields that typically study organizational training programs, such as human resources, management, and industrial-organizational psychology. Instead, fields including medicine and engineering typically test the efficacy of these programs. These prior studies have provided strong contributions to our knowledge, but the findings remain generally unknown to many researchers of fields that typically study training programs, and thereby many significant theoretical perspectives of training may have yet to be applied to VR training programs. By integrating these prior findings with modern research, we introduce new avenues for future study to researchers in multiple domains. Thus, while the current article is a synthesis of prior research, it has widespread implications for future studies.

1. Theoretical background and hypothesis development

VR is defined as the three-dimensional digital representation of a real or imagined space with interactive capabilities (Cruz-Neira et al., 1993; Steuer, 1992; Zyda, 2005). Several notes should be made about this definition. First, users navigate their three-dimensional digital environments using digital representations called avatars. Users may view their avatar from a third-person perspective, or they may be completely unaware of their avatar's appearance by taking a first-person perspective (Didehbani et al., 2016; Hammick & Lee, 2014; Plante et al., 2003). Second, the interactive capabilities of VR can vary greatly. Some VR programs enable users to explore entire digital worlds, whereas others only allow users to perform certain actions with specific objects. All VR programs, however, include some extent of interactivity. Third, VR programs can be used for training, gaming, telecommunication, or any other purpose. Fourth, VR programs are immersive and often prompt feelings of presence, but immersion and presence are not defining features of VR. Both exist on a spectrum, and it would be difficult to place an exact point for either at which programs would begin to be considered VR (among other considerations; Murray et al., 2007; Shin, 2018). Fifth, VR includes many popular video games, such as Fortnight, Minecraft, World of Warcraft, and Second Life (Kohler et al., 2011; Pellas, 2014). On the other hand, it does not include other popular video games, such as Solitaire, that do not present a three-dimensional representation with interactive capabilities. The breadth of our applied VR definition reflects the wide variety of VR programs.

VR programs can be relatively simple. They may include unrealistic graphics, few interactive opportunities, and typical computer hardware (monitor,[2] keyboard, and mouse). Alternatively, VR programs can be relatively complex. They may include lifelike graphics, an interactive world, and advanced computer hardware (e.g., HMD, motion sensors). Perhaps the best approach to understand the variation in VR programs is to discuss hardware and software.

The hardware can be differentiated as display and input hardware. Display hardware is the technology used to present the VR program, and the most common is the computer monitor (Howard, 2017, 2019). While monitors are two-dimensional, they can nevertheless present representations of three-dimensional environments. For instance, the popular computer games Minecraft and Fortnight are three-dimensional representations of a real or imagined space with interactive capabilities, and both are commonly presented on monitors. VR programs are also regularly presented via immersive displays, including HMDs and cave automatic virtual environments (CAVEs; Cordeil et al., 2017; Cruz-Neira et al., 1993; Howard, 2017, 2019). HMDs are worn on the head of users, and the digital display entirely covers the user's vision. HMDs track head movements of users to align the presented environment with the user's point of view, which provides an immersive experience. Similarly, CAVEs project a digital environment around users on the walls of a room, causing users to be surrounded by their digital, immersive environments. While these technologies are more expensive than computer monitors, they are commonly believed to produce greater fidelity and better training outcomes.

Alternatively, input hardware is the technology used to receive user commands. The most common input hardware is the keyboard and mouse, but joysticks are also commonly used in VR applications (Howard, 2017, 2019). Given the importance of computers in daily life, most employees are familiar with using a keyboard, mouse, and even joystick. This allows practitioners to utilize these technologies in conjunction with VR training programs to reduce the time needed to familiarize trainees with the program itself, which may not be true with other types of specialized input hardware, including motion sensors,

[1] In the current article, controlled experimental studies refer to empirical investigations wherein a treatment group (e.g., VR training) is compared to a control group, whether a no treatment, unequal treatment, or equal treatment control group. Such a criterion excludes one-group experimental designs (i.e., pre-post assessment of treatment group alone), which can produce inaccurate results.

[2] Some authors assert that programs must use immersive displays to be considered VR. Due to its more common usage, we apply the broader conceptualization of VR that includes both programs using immersive and non-immersive displays, but we replicate our analyses restricted to immersive display studies (Supplemental Material D). We verify that our results are consistent regardless of whether the broader or narrower definition is applied.

controllers, and custom devices (Alaraj et al., 2011; Barata et al., 2015; Gavish et al., 2015; Man, 2018). For instance, many surgical training programs create custom input devices that replicate surgical tools. These custom surgical tools can measure precise movements of trainees to determine whether they are performing the appropriate behaviors. While these specialized input devices require trainees to become accustomed to novel hardware, they are believed to produce better outcomes because they can mimic the natural performance of activities and increase fidelity.

Given these various types of display and input hardware, the physical properties of a virtual reality training program can take many forms. The program may present a digital environment on a computer monitor, and participants may interact with this environment using their keyboard and mouse. Alternatively, participants may wear an HMD, and they could use motion sensors to track their bodily movements. Even yet, participants may use a combination of non-immersive and immersive hardware, such as using a computer monitor to view their environment and custom-made input devices to deliver input (e.g., surgical tools with sensors). Across all these applications, researchers have proposed that VR training programs provide more beneficial outcomes than alternative training programs (Julier et al., 1999; Moglia et al., 2016; Moshell, 1993; Vaughan et al., 2016). Early practitioners and researchers largely used the medium due to logistical necessities, but it is now applied in broader contexts due to its assumed benefits – particularly its relation to fidelity, motivation, and task-technology fit.

Fidelity refers to the extent that the training environment reproduces the transfer environment, and prior research has differentiated two types of fidelity: physical fidelity and psychological fidelity (Baldwin & Ford, 1988; Kozlowski & DeShon, 2004). Physical fidelity refers to the extent that the training environment replicates the transfer environment, whereas psychological fidelity refers to the extent that the training environment prompts similar cognitive processes. Both types of fidelity have been linked to training effectiveness (e.g., identical elements theory; Baldwin & Ford, 1988; Hochmitz & Yuviler-Gavish, 2011; Salas, Rosen, et al., 2009), and VR has been supported to produce higher levels of both types of fidelity (Gavish et al., 2015; Howard, 2017; McMahan et al., 2012). VR can completely recreate any transfer environment given proper resources (e.g., computers and coders), and the training environment can produce cues that trigger relevant psychological processes. For example, a VR training program could present a realistic cockpit with warning messages and alert noises when the trainee is in danger of crashing. While the outcome of crashing may not be the same, research has supported that such cues in VR can produce realistic levels of stress (Auerbach, 1999; Taber, 2014). Thus, VR training programs may produce better outcomes than alternatives due to increases in fidelity.

Alternatively, some authors have speculated that the novelty of VR training programs causes users to be more interested in the learning experience and increases their motivation (Freina & Ott, 2015; Pan et al., 2006). Trainees may allocate minimal efforts to learning from otherwise boring training activities. An exciting training program, however, may prompt trainees to allocate additional mental resources to learning, assuming that they do not become distracted by the exciting elements (Landers, 2014; Landers et al., 2018). Because most trainees perceive VR to be novel and exciting (Barata et al., 2015; Gavish et al., 2015; Man, 2018), it is possible that VR can improve training outcomes due to increased motivation.

Lastly, task-technology fit refers to "the degree to which a technology assists an individual in performing his or her portfolio of tasks" (Goodhue & Thompson, 1995, p. 216), and recent authors regularly propose that VR training programs provide better fit to more training tasks than alternative training approaches (Ammenwerth et al., 2006; Howard & Rose, 2019; Rosen et al., 2006). That is, alternative training programs may require users to perform tasks that are dissimilar to the trained ability, such as by learning welding abilities via a video rather than practicing with a welder. This instance would likely represent poor task-technology fit, as the technology may not provide the capabilities necessary to learn the details necessary to weld. VR, however, can provide a complete reproduction of the transfer environment that includes every detail, and therefore it may provide better task-technology fit for learning various KSAOs.

We apply these three theoretical perspectives in detailing our hypotheses. Specifically, we use fidelity to explain the beneficial effects of VR hardware, motivation to explain the beneficial effects of VR software, and task-technology fit to explain the differing effects of VR training programs on various outcomes. Given the widespread assumptions regarding the effectiveness of VR, however, we first propose the following hypothesis

Hypothesis 1. VR training programs produce more beneficial outcomes than alternative comparisons, which include equal, unequal, and no training comparisons.

Below, we propose moderating effects that may influence the extent that VR training programs produce more beneficial outcomes than alternative comparisons using the theoretical perspectives mentioned above. We highlight that VR programs with certain attributes and/or contexts may produce improved outcomes than VR programs without those attributes (e.g., hardware, software) or contexts (e.g., participant population, control group). In such instances, the former would produce larger outcomes relative to alternative comparisons than the latter, which would result in these attributes moderating the effect of VR training programs relative to alternative comparisons (i.e., Hypothesis 1). It should be noted, however, that we do not directly compare VR program attributes or contexts against each other, and we can only instead assess their merits relative to alternative comparisons. Doing so can identify attributes that produce more beneficial VR training program outcomes and provide substantial theoretical insights.

1.1. Virtual reality hardware

Many authors have proposed that immersive hardware (e.g., HMDs, specialized input) produces better training outcomes than non-immersive hardware (e.g., monitor, keyboard, mouse) due to increases in physical and psychological fidelity (Moglia et al., 2016; Vaughan et al., 2016), which is supported by identical elements theory (Baldwin & Ford, 1988; Salas, Rosen, et al., 2009). The use of immersive hardware has also been repeatedly shown to produce greater feelings of presence, which is sensations of existing in the digital environment. These authors argue that presence subsequently relates to attention, motivation, learning, and ultimately behavioral change (Dubovi et al., 2017; Tüzün & Özdinç, 2016). When using an HMD with motion sensors, for instance, trainees may feel transported to their digital environments, which reduces the need to "translate" behaviors between training and transfer environments. We propose two hypotheses regarding display and input hardware in VR training programs:

Hypothesis 2. The effect of VR training programs relative to alternative comparisons is moderated by the applied display hardware, such that the effect is stronger when using immersive display hardware (e.g., HMDs, CAVEs) than non-immersive display hardware (e.g., computer monitor) in the VR training program.

Hypothesis 3. The effect of VR training programs relative to alternative comparisons is moderated by the applied input hardware, such that the effect is stronger when using specialized input hardware (e.g., motion sensors, custom devices) than non-specialized input hardware (e.g., keyboard, mouse, joystick) in the VR training program.

1.2. Virtual reality software

Few – if any – features of VR software have been discussed more than whether the program is gamified or non-gamified (Ferro et al., 2013;

Landers, 2014; Li, Chan, & Skitmore, 2012). A VR program is considered gamified if it includes multiple game attributes that are not required of VR, whereas game attributes are individual aspects that can be used to define and differentiate games. Game attributes are often considered the core characteristics that cause a game to be enjoyable, and the presence of game attributes often improves user attention and motivation. Subsequently, improvements to attention and motivation can improve learning and behaviors. Trainees may allocate minimal cognitive efforts to learning KSAOs in boring training environments and thereby experience reduced outcomes, but a similar training program with added game attributes may cause trainees to allocate more efforts and better develop their KSAOs. Recently, researchers have even begun applying modern motivational theory to understand these effects of gamification (e.g., self-determination theory, expectancy-value models; Di Serio et al., 2013; Ferro et al., 2013; Howard, 2017; Landers, 2014; Li, Chan, & Skitmore, 2012).

Several authors have developed game attribute taxonomies (Bedwell et al., 2011; Garris et al., 2002; Wilson et al., 2009) that can be used to identify and analyze the effects of individual game attributes, but they can also be used to determine whether a VR program is gamified. Currently, most researchers of VR training programs make dichotomous comparisons between gamified and non-gamified programs by applying these prior taxonomies (Freina & Ott, 2015; Howard, 2019; Howard & Gutworth, 2020; Seaborn & Fels, 2015; Theng et al., 2015; Xu et al., 2017). A program is considered gamified if it includes multiple elements that are not required by the medium (e.g., conflict, surprise), whereas it is considered non-gamified if it includes only elements required by the medium (e.g., environment) or very few other game attributes. We make a similar distinction in the current article, and we propose the following:

Hypothesis 4. The effect of VR training programs relative to alternative comparisons is moderated by the inclusion of game attributes, such that the effect is stronger when using gamified VR training programs than non-gamified VR training programs.

1.3. Virtual reality training context

To identify the outcomes for which VR training programs may be most effective in developing, it is beneficial to integrate experiential learning theory. Experiential learning theory proposes that people can effectively learn when they perform, grasp, and transform experiences (Kolb et al., 2001; Kolb & Kolb, 2009). In other words, people can effectively learn by doing. The naturalistic feedback provided in experiential learning applications also enables participants to understand which behaviors can lead to mistakes and the detrimental outcomes of these mistakes. Programs that actively encourage trainees to make mistakes often produce better learning outcomes than those that do not (Carter & Beier, 2010; Keith & Frese, 2008). VR provides the unique opportunity for trainees to perform any task in any environment, which may be otherwise unfeasible due to risks or financial costs. Trainees can create new conceptual connections and make mistakes in VR training programs. For instance, trainees in a VR pilot training can learn exactly which behaviors would cause a plane crash, providing a clearer frame-of-reference for detrimental behaviors and their catastrophic outcomes. Many authors have suggested that the greatest benefit of VR is its ability to teach skills by doing (Cheng & Wang, 2011; Doyle et al., 2015; Su & Cheng, 2013), and VR training programs are believed to produce the largest improvements relative to alternatives when developing skills.

Some authors, however, have used VR to develop declarative knowledge and general mental abilities of trainees (Lee et al., 2017; Ren et al., 2015), wherein declarative knowledge refers to factual information and general mental abilities refers to quickness of thought, spatial reasoning, and other abilities associated with fluid intelligence (Cohen et al., 1985; Schmidt & Hunter, 1993). In these applications, researchers cite the allure of VR for any improvements to training outcomes. That is, trainees tend to naturally enjoy VR, and they may allocate more efforts to training tasks and learning processes.

However, most alternative training approaches can develop declarative knowledge and general mental abilities in a similar manner to VR training programs. For instance, Merchant et al. (2013) developed a VR training program for spatial abilities and chemistry knowledge, in which participants were asked to digitally rotate objects and construct molecular structures. They compared this program to a traditional chemistry course using two-dimensional images of rotated objects and molecular structures. While the study had sufficient statistical power, the authors did not see significant differences in outcome improvement between the two groups. In this instance, rotating objects and drawing molecular structures were very similar in the VR and non-VR training programs, and we suggest that similar dynamics may occur for broader VR training programs aimed at developing declarative knowledge and general mental abilities.

Results such as Merchant et al. (2013) reinforce that the greatest potential benefit of VR may indeed be its ability to teach skills by doing, and VR training programs may produce the greatest benefits when applied to develop skills. While we propose that VR training programs are more effective than alternatives across all contexts, we also suggest that the medium is more effective for developing skills than it is for developing declarative knowledge and general mental abilities. We further suggest that VR is more effective at developing declarative knowledge than general mental abilities. Prior research has supported that general mental abilities are relatively difficult to substantially improve via training approaches, especially compared to skills and declarative knowledge (Au et al., 2015; Colom et al., 2013; Harrison et al., 2013). Because any training approach may be limited in their ability to improve general mental abilities, it is less likely that VR training programs produce substantially larger outcomes than other approaches. That is, it is most likely that all approaches produce small effects in developing general mental abilities, causing VR training programs to have only a slightly larger effect. Thus, we propose:

Hypothesis 5. The effect of VR training programs relative to alternative comparisons is moderated by the type of outcome, such that the effect's strength varies in the following order: general mental abilities, declarative knowledge, and skills (weakest to strongest).

Training outcomes can further be categorized by their complexity. Typically, declarative knowledge and general mental abilities are considered low to moderate complexity, and developing these outcomes requires straightforward practice and repetition. On the other hand, skills can represent the entire range of complexity, ranging from low-, to moderate-, to high-complexity. For example, VR has been used to train street-crossing skills, which is considered a low-complexity outcome (Josman et al., 2008); it has been used to train motor vehicle driving skills, which is considered a moderate-complexity outcome (Haeger et al., 2018); and it has been used to train surgical skills, which is considered a high-complexity outcome (Gurusamy et al., 2009). In general, higher-complexity outcomes are believed to be more difficult to train, and trainees may need to experience unpredictable breakthrough moments before they can advance beyond certain skill plateaus (Brunner et al., 2004; Grantcharov et al., 2003; Hogle et al., 2007; Selvander & Åsman, 2012). Alternatively, lower-complexity outcomes are believed to be easier to train, and trainees may only need to perform repetitions before they can master the skill.

By applying experiential learning theory (Kolb et al., 2001; Kolb & Kolb, 2009), we suggest that VR is more effective at developing complex training outcomes relative to alternatives. Experiential learning theory proposes that trainees undergo certain mental processes when performing a task wherein they draw connections between the task and their prior knowledge. In doing so, they reflect upon the most appropriate method to perform a task, and they identify approaches to perfect these methods (Kolb et al., 2001; Kolb & Kolb, 2009). Engaging in these mental processes is difficult without performing the task, but VR can

provide this much-needed opportunity where alternative approaches cannot. Thus, VR may be most effective in developing complex training outcomes relative to alternatives, as these alternative training programs may not provide a suitable practice space for these complex skills.

Alternatively, less-complex outcomes may not require as much reflection and critical thinking to develop abilities, and thereby these outcomes may not necessitate the performance of these skills in VR. Instead, these less-complex outcomes may be sufficient to develop via mental rehearsal and other similar approaches. Therefore, we propose the following hypothesis:

Hypothesis 6. The effect of VR training programs relative to alternative comparisons is moderated by the complexity of outcome, such that the effect's strength varies in the following order: less, moderately, and more complex (weakest to strongest).

1.4. Task-technology fit

In proposing many of the hypotheses above, we stressed that VR training programs may produce more beneficial outcomes in contexts suited for their applications, which closely aligns with the proposals of task-technology fit theory. Prior studies have shown that technologies perform better when they are matched to the context, resulting in a positive relation between task-technology fit and performance (Goodhue & Thompson, 2015; Howard & Rose, 2019). For this reason, we suggest that VR training programs with greater task-technology fit produce stronger effects relative to alternative comparisons than VR training programs with less task-technology fit, and task-technology fit is therefore a moderating effect. In the current meta-analysis, we code the task-technology fit of training (task) by VR programs (technology) to determine whether VR training programs with greater task-technology fit indeed produce stronger effects. For example, a VR training program for surgical skills using realistic tools as input, a complete visual representation of a patient, and integrated instruction would have high task-technology fit, whereas a VR training program for surgical skills using a keyboard, mouse, poor representation of a patient, and no instruction would have low task-technology fit.

Hypothesis 7. The effect of VR training programs relative to alternative comparisons is moderated by task-technology fit, such that the effect's strength varies in the following order: less fit, moderate fit, more fit (weakest to strongest).

1.5. Study design

Study designs should be considered for two reasons. First, certain confounding factors of research design may cause VR training programs to appear superior to alternatives, whereas this may not reflect their true impact. Second, problematic research designs may be repeatedly used in VR training research, and identifying the popularity of these designs could be the first step towards reducing their usage. Therefore, we discuss three aspects of the research design.

Kirkpatrick (1975, 1979) distinguished four levels that training outcomes can be measured by: reactions, learning, behaviors, and results. Reactions refers to the trainees' perceptions of the training program; learning refers to the trainees' acquisition of knowledge and skills; behaviors refers to behavioral change of participants in natural settings after the training; and results refers to the extent to which the training influenced the ultimate desired outcomes, such as organizational performance. Research has repeatedly supported that training outcomes are increasingly difficult to develop when ascending the levels, with reactions being the easiest and results being the most difficult (Salas & Cannon-Bowers, 2001; Salas, Rosen, et al., 2009). This is, in part, because the outcomes become more temporally distant from the training when ascending levels, leaving the opportunity for other influences to likewise impact these outcomes and attenuate the effect of the training program. It is tempting for researchers to measure reactions as the sole training outcome in order to force their training program to appear effective, but higher-level outcomes often reflect the true desired effects on the trainee and organization. Researchers should strive to measure these higher-level outcomes in evaluation studies, and we propose the following hypothesis to understand (a) whether this phenomenon holds in VR training research and (b) the extent to which researchers measure these preferred higher-level outcomes.

Hypothesis 8. The effect of VR training programs relative to alternative comparisons is moderated by the measurement approach, such that the effect's strength varies in the following order: results, behaviors, learning, and reactions (weakest to strongest).

It is similarly preferred for researchers to test their developed training programs against comparable training programs that are matched for duration and/or intensity (Salas & Cannon-Bowers, 2001; Salas, Rosen, et al., 2009). This is not always the case, however. Some researchers compare a VR training program against an alternative training program that is shorter and/or includes fewer training activities (Bhagat et al., 2016; Seixas-Mikelus et al., 2010), and other researchers compare a VR training program against a true control group that does not receive an intervention whatsoever (Cheng & Wang, 2011; Haeger et al., 2018; Häll et al., 2011). While these research designs produce larger effects, they do not reflect the impact of VR training programs relative to alternative approaches that are currently used for learning purposes. Thus, we test whether the type of control group influences the magnitude of observed results.

Hypothesis 9. The effect of VR training programs relative to alternative comparisons is moderated by the type of control group, such that the effect's strength varies in the following order: equal, unequal, and no training control (weakest to strongest).

VR has been used with a variety of populations, including employees, college students, children, and the elderly.[3] Some authors have speculated that VR may be most effective for students and children, due to their natural excitement regarding the medium (Freina & Ott, 2015; Pan et al., 2006). It is unclear whether this notion is true, and we propose the following:

Research Question 1. Is the effect of VR training programs relative to alternative comparisons moderated by the participant population?

1.6. Source characteristics

We test two aspects of the source characteristics. Practitioners and researchers have continuously developed new VR technologies to improve outcomes. It is expected that these developments are effective and more recent studies produce larger effects than older studies.

Hypothesis 10. The effect of VR training programs relative to alternative comparisons is moderated by the publication year, such that the effect become stronger with year.

When performing a meta-analysis, researchers are recommended to discover unpublished sources, as these may be more likely to include non-significant results than published sources (Rosenberg, 2005). We test whether this notion is also true for VR training programs.

Hypothesis 11. The effect of VR training programs relative to alternative comparisons is moderated by the publication source, such that the effect is larger in published sources (e.g., journal articles) than unpublished sources (e.g., dissertations).

[3] We meta-analyze all types of VR training programs, including both educational and organizational programs, but we also provide all results restricted to organizational training programs alone (Supplemental Material F).

2. Method

We proposed that VR training programs are more effective than alternative training approaches, and we identified many possible moderating characteristics. To best test these proposals, we conduct a meta-analysis of controlled experimental investigations on VR training programs to determine if these training programs produce greater effects, overall, than alternative training programs. We code and analyze attributes of the sources to determine whether specific characteristics alter observed effects. For instance, we code whether the study used immersive or non-immersive display hardware, and we calculate estimates separately for each. Comparing the results can determine whether immersive displays indeed produce more beneficial effects than non-immersive displays, and we provide similar analyses for all other hypotheses and research questions. We followed suggestions of prior authors, placing a particular focus on the preferred reporting for systematic reviews and meta-analyses (PRISMA) (Borenstein et al., 2011; Cheung, 2015; Cheung & Cheung, 2016; Hunter & Schmidt, 2004; Jak, 2015; Kepes et al., 2013; Liberati et al., 2009; Lipsey & Wilson, 2001; Moher et al., 2009, 2015).

2.1. Identifying sources

We used typical approaches to discover all relevant published and unpublished sources. Searches were conducted using EBSCO and Google Scholar in January of 2020. EBSCO searches return results for several other databases including Academic Source Complete, ERIC, and PsycINFO. All EBSCO results were included in our initial coding database. Alternatively, Google Scholar is more comprehensive than other academic databases, as it catalogues more conference materials and journals, but Google Scholar is also more likely to return only tangentially related search results. For instance, the search "'Virtual Reality' 'Training'" returned 2472 results in EBSCO but 130,000 results in Google Scholar. We only included the first 1000 results of our Google Scholar searches in our coding database. Results after the first several hundred were tangentially related (e.g., terms not in primary text but appeared once in references). Thus, the decision to only record the first 1000 results was supported.

Using these databases, each search included two parts. The first part included the phrase "Virtual Reality", "Computer Simulation", or "Digital Simulation", and the second part included the word "Training", "Intervention", "Therapy", "Enhance", "Promote", "Support" (quotation marks included) or several other terms relevant to a simultaneous but separate meta-analysis on VR for the development of social skills (Howard & Gutworth, 2020). Further, we also conducted forwards and backwards searches of prominent articles of virtual reality training programs, which included both identifying relevant sources cited by the article as well as relevant sources that cited the article (using Google Scholar). These searches returned 16,684 sources to be initially coded.

It should also be noted that all searches were restricted to the year 2010 to the present. Practitioners and researchers have greatly advanced the sophistication of VR technology in the past decade (Alaker et al., 2016; Moglia et al., 2016). Prior to 2010, the development of HMDs, CAVES, and other VR technologies was mainly restricted to small, specialized companies. With the rise of Oculus in 2012, however, many other organizations began to develop their own VR research and development efforts. Today, large corporations, such as Google and Facebook, produce their own VR technologies that are much more advanced than the technologies of only a decade prior. For this reason, we believed that restricting our searches from 2010 to the present would provide a more accurate depiction of the current effects of VR training programs, as it would only include studies slightly before and during the current growth of VR technologies. This search restriction has also been used in prior meta-analyses of VR (Howard, 2017, 2019).

2.2. Inclusion criteria

For each coding phase of the current meta-analysis, two coders first jointly constructed coding guidelines, developed a coding rulebook, reviewed the material, and trained each other on the coding rules. Then, the two trained coders initially coded the same sources until a sufficient level of inter-rater agreement was reached (Cohen $\kappa > 0.80$). These two coders then coded each source independently after meeting sufficient agreement, which is common in meta-analyses (Hunter & Schmidt, 2004; Kepes et al., 2013; Liberati et al., 2009; Lipsey & Wilson, 2001).

In the first phase, the sources were coded for the following criteria: (a) written in the English language, (b) studied human participants, (c) included more than nine participants, (d) provided quantitative results, (e) used VR, (f) tested a training program, and (g) included a control/comparison group that did not undergo a VR training program. Each of these criteria were chosen due to prior precedence in meta-analyses of computer-based training programs and VR (Alaker et al., 2016; Blume et al., 2010). This reduced the list of 16,684 sources to 802.

In the second phase, the two coders coded whether the training developed skills, knowledge, and/or general mental abilities. We specifically excluded sources that developed social skills, as prior research has shown that development of social skills involves very different processes than the development of procedural skills, knowledge, and general mental abilities (Didehbani et al., 2016; Howard & Gutworth, 2020). We did not want our results to be confounded by the inclusion of other types of training. Also, we did not include studies that provided a physical outcome training program (e.g., treadmill training) and measured cognitive skills, knowledge, and/or general mental abilities. Again, these programs invoke different processes than more typical cognitive training programs, and we did not want their influence confounding our results. This coding phase reduced the list of 802 sources to 238.

In the third phase, both coders coded each of the final 238 sources for their effect sizes and the moderating characteristics described below. We contacted all corresponding authors for unreported results and unpublished datasets. Most authors did not reply, resulting in 74 sources being excluded from analyses because the original source did not provide useable statistics and/or the source could not be obtained. Lastly, we coded individual sources as separate studies if they (a) included multiple entirely separate studies or (b) included multiple groups in a single study that underwent a VR training program along with one or more groups that did not undergo a VR training program. We did not consider a source to include separate studies if it reported a single group undergoing a VR training program along with multiple groups that did not undergo a VR training program. These phases produced a final dataset of 164 sources and 184 separate studies.

2.3. Source and study characteristic coding

Coding results for each source and effect are provided in Supplemental Material A.

Display hardware. We coded whether the VR training program used a monitor or an immersive display (e.g., HMDs, CAVEs). If the authors did not directly state the display hardware, it was assumed that a monitor was used.

Input hardware. We coded whether the VR training program used a keyboard, mouse, and/or joystick or a specialized input technology (e.g., motion sensors). If the authors did not state the input hardware, it was assumed that a keyboard, mouse, and/or joystick was used.

Gamified. We coded the VR intervention as gamified if (a) the authors stated it was a game or game-like training or (b) it appeared to include multiple game attributes. We coded the VR intervention as not gamified if (a) the authors did not label it as a game or game-like training or (b) it included very few game attributes beyond those required for VR (e.g., environment).

Type of outcome. We coded the VR training outcome as either a

skill, declarative knowledge, or general mental abilities. A skill was a trained activity that was typically measured with a practice test of performing the activity. Declarative knowledge was a memorized set of information that was typically measured with a written test. General mental abilities was a trained set of aptitudes that included quickness of thought, spatial reasoning, working memory, and other similar constructs. These were typically measured with intelligence tests.

Complexity of outcome. We coded the VR training outcome as either less complex, moderately complex, or more complex. This coding decision was primarily determined by the amount of information required for the training outcome, with the assumption that skills require many different, inter-related pieces of information to effectively perform. While the type of outcome spanned all three levels of complexity, declarative knowledge and general mental abilities were most often coded as low- or moderate-complexity, whereas skills were most often coded as moderate or high-complexity (although not exclusively). An example low-complexity outcome is teaching medical residents basic parts of anatomy (Seixas-Mikelus et al., 2010); an example moderate-complexity outcome is teaching construction workers how to dismantle a crane (Li, Chan, & Skitmore, 2012); and an example high-complexity outcome is teaching surgical trainees how to complete a laparoscopic cholecystectomy (Kowalewski et al., 2018).

Task-technology fit. We coded the VR training program as producing either less, moderate, or more task-technology fit. While fidelity played a role in determining the task-technology fit of the VR training program, it was not the only contributing factor. Instead, we also considered the extent that the training program was catered to the specific context and outcome of interest as well as the amount of instruction it provided. For instance, a realistic VR training program on welding may produce moderate task-technology fit, but a realistic VR training program on welding a specific underwater gas tank that is the exact transfer task would produce more task-technology fit. Likewise, both of these VR training programs would produce poor task-technology fit if they were applied to train driving skills (or any other irrelevant task).

Outcome measure. We coded the outcome using Kirkpatrick's model, which includes four categories: reactions, learning, behavior, and results. These categories are well defined by Kirkpatrick (1975, 1979), which readers can reference for more information regarding the categories. An example reaction measure is a survey asking trainees about their perceptions of the training; an example learning measure is a graded test; an example behavior measure is observations of whether trainees performed specific trained behaviors in their workplace; and an example result measure is organizational revenue or employee career outcomes.

Control group. We coded whether the included control group was an equal training, unequal training, or no training group. An equal training group received a training matched for the duration and intensity of the VR training. An unequal training control group received an intervention that was not matched for the duration or intensity of the VR training. A no training control group did not receive any training whatsoever.

Participant population. We coded whether participants were children/teenagers, college, adult, or specialized samples. Child/teenage samples included participants under the age of 18 and not enrolled in college. College samples included participants enrolled in college. Adult samples included participants that were the age of 18 or older and not currently enrolled in college. All other populations were specialized populations, which included the elderly and those with physical (e.g., stroke survivors) or mental disabilities (e.g., brain injury survivors).

Organizational training. We coded whether the training was used for organizational purposes, which included training for military, medicine, and beyond. We did not consider training programs for college courses to be organizational trainings. We considered training programs to be for organizational purposes if medical residents were currently performing the KSAOs in their rotations, but we did not consider these programs to be for organizational purposes if the medical students had not yet begun using those skills in their rotations.

Publication year. We coded the year the source was published, presented, or completed.

Source. We coded whether the source was an unpublished source (e.g., research report), thesis or dissertation, conference paper, or journal article.

2.4. Analyses

We first calculated several indices of publication bias, including fail-safe k, Egger's test, trim-and-fill method, and weight-function model analysis. More information regarding each of these is provided in Supplemental Material B; the results for the first three are provided in Table 1, whereas the weight-function model analysis results are provided in Supplemental Material B. We also analyzed our dataset for outliers and influential cases with a focus on studentized deleted residuals, Cook's distance, and covariance ratios (Viechtbauer & Cheung, 2010). Reporting of these values is provided in Supplemental Material B and summarized below.

To calculate the primary effects, we used a random-effects model in Comprehensive Meta-Analysis V3, which weighted each effect size by its associated sample size (discussed further below). Random-effects meta-analyses are more resilient to outlier studies than fixed-effects approaches, and random-effects approaches provide more accurate results across most circumstances (Hedges & Vevea, 1998; Hunter & Schmidt, 2000). We report all results as Cohen's *d*, which is an estimate of the mean difference between two groups (e.g., VR training vs. control training) relative to their standard deviations. For example, a Cohen's *d* of one indicates that the means of the two groups were one standard deviation apart. We did not include effect sizes that represent the relationship of more than two variables (e.g., ANCOVA) due to statistical concerns (Anderson et al., 2010; Boxer et al., 2015; Rothstein & Bushman, 2015); however, we did include effect sizes that represent comparisons of pre- and post-training changes between two groups (e.g., pre- and post-training means and standard deviations for two groups). No corrections were made for unreliability, as most sources did not report internal consistency estimates for outcomes. We also averaged multiple effects from the same study together before calculating our analyses to avoid over-weighting individual studies; however, we also utilized other approaches for accounting for this dependency in the sensitivity analyses discussed below.

To test the effects of sources and study characteristics, we calculated meta-analytic estimates separated by each level of the source/study characteristic and compared confidence intervals. We also calculated separate meta-regressions for each characteristic. If the associated coefficient is statistically significant, then the characteristic significantly influences VR training program efficacy (Thompson & Higgins, 2002; Van Houwelingen et al., 2002).

Lastly, we performed sensitivity analyses, which is the replication of meta-analyses using varied approaches to ensure that results are not driven by analysis decisions. Supplemental Material B includes reanalyses using a three-level meta-analytic approach, which identifies sources of dependency within and across studies and addresses multiple effect sizes from a single study without averaging them together (Cheung, 2015; Jak, 2015). Supplemental Material C includes reanalyses without outliers and influential cases; Supplemental Material D includes reanalyses with only adult samples; Supplemental Material E includes reanalyses with only organizational training programs; Supplemental Material F includes reanalyses with only studies that applied immersive display hardware. Supplemental Material G includes reanalyses with differing assumptions for the coding of input and output hardware (described below). None of our inferences were altered from these analyses, supporting the robustness of our results.

Table 1
Publication bias results.

	I^2	k	Fail Safe k	Egger's test β_0	*Egger's Test t*	Implied Missing	
						Left of Mean	Right of Mean
1.) Overall Effectiveness of VR Training vs. Control Training	72.975	184	21,604	1.219	3.588*	0	12
Type of Display Hardware							
2a.) Computer Monitor	73.145	133	2338	.957	2.499*	0	0
2 b.) Immersive Display	70.982	54	1328	2.411	3.451*	0	7
Type of Input Hardware							
3a.) Basic Input	76.275	65	4195	1.383	2.474*	0	12
3 b.) Specialized Input	70.592	120	6781	1.547	3.335*	25	0
Game Attributes							
4a.) Excluded	69.335	153	3719	1.025	2.775*	27	0
4 b.) Included	83.438	31	859	2.069	2.208*	0	6
Type of Outcome							
5a.) Mental Abilities	79.179	31	572	1.281	1.332	0	9
5 b.) Declarative Know.	78.495	53	2040	1.256	1.732	0	11
5c.) Skills	67.186	112	7132	1.353	3.191*	24	0
Complexity of Outcome							
6a.) Basic	80.457	78	3842	1.206	1.888	0	14
6 b.) Moderate	37.042	29	364	.528	.824	0	1
6c.) Complex	66.449	85	4985	1.172	2.360*	13	0
Task-Technology Fit							
7a.) Less	57.680	43	176	-.102	.134	0	0
7 b.) Moderate	59.327	57	1886	1.380	2.813*	0	2
7c.) More	77.115	84	7791	1.146	2.191*	0	0
Type of Outcome Measure							
8a.) Reactions	87.115	28	590	-.222	.153	0	6
8 b.) Learning	74.774	154	2668	1.315	3.492*	0	14
8c.) Behaviors	44.827	26	573	1.266	1.622	4	0
8 d.) Outcomes	0	2	–	–	–	–	–
Type of Control Group							
9a.) No Treatment	73.167	33	1197	2.582	3.994*	11	0
9 b.) Non-Equal Treatment	75.521	41	1639	1.819	2.116*	0	11
9c.) Equal Treatment	77.114	130	7775	.682	1.456	0	24
Type of Participant							
10a.) Children & Teenagers	84.556	20	423	.317	.192	0	5
10 b.) College Students	64.111	54	1781	1.348	2.301*	0	2
10c.) Adults	74.370	89	4194	1.970	3.536*	0	0
10 d.) Special Populations	61.182	23	325	−1.312	.992	0	1
Organizational Training							
11a.) Not Org. Training	71.940	120	9864	1.106	2.544*	0	0
11 b.) Org. Training	75.019	64	2211	1.627	2.596*	0	9
Type of Source							
12a.) Unpublished	72.590	16	1	.347	.251	1	0
12 b.) Thesis & Dissertation	0	5	0	-.643	.262	0	0
12c.) Conference Paper	0	1	–	–	–	–	–
12 d.) Article	73.057	162	9418	1.312	3.699*	0	10

*$p < .05$ (two-tailed).

3. Results

3.1. Publication bias and outlier analyses

Table 1 includes indices of publication bias. The overall fail-safe k was 21,604, and all but two fail-safe k were larger than 300 when restricted to specific significant study/source characteristics. According to prior guidelines (Orwin, 1983; Rothstein et al., 2005), these fail-safe k values are sufficiently large and the meta-analytic effects can be considered robust.

Egger's test, trim-and-fill method, and weight-function analyses signified publication biases in the overall and many subgroup analyses. These instances of publication bias primarily arose in two manners. First, many studies (>5) were implied to be missing to the left of the mean for five subgroup analyses, which is the expected direction of publication biases. In these instances, some studies with positive and large effects had small sample sizes. Because an equal number of small-sample-size studies with negative and large effects are not also included in the dataset, these results suggest that publication bias may indeed be present. Second, many studies were unexpectedly implied to be missing to the right of the mean for the overall analysis and thirteen subgroup analyses. The analysis identified several effects missing above the mean because two positive outliers shifted the mean upwards (discussed below), causing more effects to fall below the mean. In these cases, the two possible outliers were the apparent causes of potential biases, rather than substantial effects of publication biases.

To further explore possible biases in the dataset, we calculated relevant statistics to test for influential cases and outliers (Supplemental Material B). Each of these indicated that two outliers may be present in the overall analyses. We included these two sources in analyses because random-effects approaches are resilient to outliers and none of these sources altered our interpretations of results; however, Supplemental Material C includes the current results recalculated without these two sources, which readers can reference to ensure that the current results

are indeed consistent when removing these influential cases and/or outliers.

Lastly, the I^2 values were large (~75) for the overall analysis and most subgroup analyses, suggesting that moderator variables are yet to be discovered and additional future directions can be pursued beyond the current meta-analysis.

3.2. Primary analysis

Table 2 presents our primary meta-analytic results. VR training programs were overall more effective than comparison groups with a mean that was, on average, 0.541 of a standard deviation larger than the comparison group ($d = 0.541$, *95% C·I.* [0.450, 0.631], $z = 11.648$, $p < .001$). Hypothesis 1 was supported.

Hypothesis 2 proposed that the effect of VR training programs relative to alternative comparisons is moderated by the applied display hardware, such that the effect is stronger when using immersive display hardware (e.g., HMDs, CAVEs) than non-immersive display hardware (e.g., computer monitor) in the VR training program. Traditional display hardware produced a moderate effect ($d = 0.568$, *95% C·I.* [0.461, 0.676], $z = 10.333$, $p < .001$), whereas immersive display hardware produced a similar moderate effect ($d = 0.462$; *95% C·I.* [0.299, 0.625], $z = 5.557$, $p < .001$). A meta-regression using display hardware as the sole predictor was not statistically significant ($\beta = 0.097$, *S.E.* = 0.103, *95% C·I.* [-0.105, 0.300], $z = 0.939$, $p = .348$, $R^2 = 0.00$), and Hypothesis 2 was not supported.

Hypothesis 3 proposed that the effect of VR training programs relative to alternative comparisons is moderated by the applied input hardware, such that the effect is stronger when using specialized input hardware (e.g., motion sensors, custom devices) than non-specialized input hardware (e.g., keyboard, mouse, joystick) in the VR training program. Traditional input hardware ($d = 0.605$, *95% C·I.* [0.462, 0.749], $z = 8.266$, $p < .001$) and specialized input hardware ($d = 0.503$, *95% C·I.* [0.385, 0.620], $z = 8.373$, $p < .001$) both produced moderate effects, and the associated meta-regression was not statistically significant ($\beta = 0.112$, *S.E.* = 0.097, *95% C·I.* [-0.073, 0.306], $z = 1.205$, $p = .228$, $R^2 = 0.00$). Hypothesis 3 was not supported.

Hypothesis 4 proposed that the effect of VR training programs relative to alternative comparisons is moderated by the inclusion of game attributes, such that the effect is stronger when using gamified than non-gamified VR training programs. Programs without game attributes ($d = 0.525$, *95% C·I.* [0.428, 0.622], $z = 10.604$, $p < .001$) as well as with game attributes ($d = 0.609$, *95% C·I.* [0.362, 0.855], $z = 4.836$, $p < .001$) produced moderate effects. A meta-regression using only game attributes as the predictor was not statistically significant ($\beta = 0.067$, *S.E.* = 0.122, *95% C·I.* [-0.171, 0.306], $z = 0.553$, $p = .580$, $R^2 = 0.00$). Hypothesis 4 was not supported.

Hypothesis 5 proposed that the effect of VR training programs relative to alternative comparisons is moderated by the type of outcome, such that the effect's strength varies in the following order: general mental abilities, declarative knowledge, and skills (weakest to strongest). VR training programs had a moderate effect on general mental abilities ($d = 0.501$, *95% C·I.* [0.269, 0.773], $z = 4.236$, $p < .001$), declarative knowledge ($d = 0.464$, *95% C·I.* [0.314, 0.615], $z = 6.044$, $p < .001$), and skills ($d = 0.576$, *95% C·I.* [0.454, 0.698], $z = 9.271$, $p < .001$). Dummy-coded meta-regressions were conducted to determine whether these group differences were significant using each possible combination of dummy-codes, but none of the effects were statistically significant (all $p > .05$, all $R^2 = 0.00$). Hypothesis 5 was not supported.

Hypothesis 6 proposed that the effect of VR training programs relative to alternative comparisons is moderated by the complexity of outcome, such that the effect's strength varies in the following order: less complex, moderately complex, more complex (weakest to strongest). VR training programs had a moderate effect on basic ($d = 0.480$, *95% C·I.* [0.339, 0.622], $z = 6.640$, $p < .001$), moderate ($d = 0.423$, *95% C·I.* [0.276, 0.570], $z = 5.642$, $p < .001$), and complex outcomes ($d = 0.626$, *95% C·I.* [0.484, 0.768], $z = 8.614$, $p < .001$). In the dummy-coded meta-regressions, no predictor was statistically significant (all $p > .05$, all $R^2 = 0.00$). Hypothesis 6 was not supported.

Hypothesis 7 proposed that the effect of VR training programs relative to alternative comparisons is moderated by task-technology fit, such that the effect's strength varies in the following order: less fit, moderate fit, more fit (weakest to strongest). Programs with less fit produced a small effect ($d = 0.207$, *95% C·I.* [0.063, 0.352], $z = 2.816$, $p = .005$), programs with moderate fit produced moderate effects ($d = 0.450$, *95% C·I.* [0.331, 0.570], $z = 7.384$, $p < .001$), and programs with more fit produced strong effects ($d = 0.820$, *95% C·I.* [0.655, 0.984], $z = 9.738$, $p < .001$). Dummy-coded meta-regressions to compare these conditions produced significant effects for the comparison of less fit and moderate fit ($\beta = 0.257$, *S.E.* = 0.117, *95% C·I.* [-0.028, 0.485], $z = 2.196$, $p = .028$, $R^2 = 0.13$), less fit and more fit ($\beta = 0.582$, *S.E.* = 0.112, *95% C·I.* [0.362, 0.801], $z = 5.200$, $p < .001$), as well as moderate fit and more fit ($\beta = 0.325$, *S.E.* = 0.103, *95% C·I.* [0.124, 0.527], $z = 3.162$, $p = .002$). Hypothesis 7 was supported.

Hypothesis 8 proposed that the effect of VR training programs relative to alternatives is moderated by the measurement approach, such that the effect's strength varies in the following order: results, behaviors, learning, and reactions (weakest to strongest). VR training programs had a strong effect on behaviors ($d = 0.725$, *95% C·I.* [0.516, 0.935], $z = 6.781$, $p < .001$), whereas they had moderate effects on reactions ($d = 0.502$, *95% C·I.* [0.223, 0.782], $z = 3.521$, $p < .001$), learning ($d = 0.491$, *95% C·I.* [0.390, 0.591], $z = 9.572$, $p < .001$), and outcomes ($d = 0.446$, *95% C·I.* [-0.112, 1.004], $z = 1.568$, $p = .117$). In the multiple dummy-coded meta-regressions, the dummy code representing the difference between learning and behaviors was statistically significant ($\beta = -0.330$, *S.E.* = 0.149, *95% C·I.* [-0.622, −0.038], $z = -2.216$, $p = .027$, $R^2 = 0.02$), whereas no other predictor was statistically significant (all $p > .05$). Hypothesis 8 was partially supported.

Hypothesis 9 proposed that the effect of VR training programs relative to alternative comparisons is moderated by the type of control group, such that the effect's strength varies in the following order: equal-training control, unequal training control, and no training control (weakest to strongest). VR training programs had a moderate effect on equal treatment control groups ($d = .427$, *95% C·I.* [0.317, 0.536], $z = 7.650$, $p < .001$), and they had very strong effects on no treatment ($d = 0.867$, *95% C·I.* [0.616, 1.119], $z = 6.752$, $p < .001$) and unequal treatment control groups ($d = 0.803$, *95% C·I.* [0.565, 1.041], $z = 6.616$, $p < .001$). In the multiple dummy-coded meta-regressions, the dummy code representing the difference between no treatment control groups and equal-training control groups was statistically significant ($\beta = -0.349$, *S.E.* = 0.153, *95% C·I.* [-0.649, −0.049], $z = -2.281$, $p = .023$, $R^2 = 0.00$), whereas no other dummy-coded predictors were statistically significant (all $p > .05$). Hypothesis 9 was partially supported.

Research Question 1 suggested that the effect of VR training programs relative to alternative comparisons may be moderated by the participant population. VR training programs had moderate effects on outcomes for children/teenagers ($d = 0.527$, *95% C·I.* [0.257, 0.798], $z = 3.825$, $p < .001$), college students ($d = 0.432$, *95% C·I.* [0.310, 0.553], $z = 6.941$, $p < .001$), special populations ($d = 0.614$, *95% C·I.* [0.354, 0.874], $z = 4.629$, $p < .001$), and adults ($d = 0.614$, *95% C·I.* [0.451, 0.776], $z = 7.407$, $p < .001$). In the dummy-coded meta-regressions, no predictor was statistically significant (all $p > .05$, all $R^2 = 0.00$). Because the participant populations were comparable in their results, our decision to analyze them together was supported.

We also tested whether the effect sizes reported in organizational VR training programs differed from effect sizes in non-organizational VR training programs. Both organizational ($d = 0.625$, *95% C·I.* [0.430, 0.821], $z = 6.267$, $p < .001$) and non-organizational VR training programs ($d = 0.512$, *95% C·I.* [0.411, 0.613], $z = 9.907$, $p < .001$) produced moderate effects, and a meta-regression did not find a significant difference between the two groups ($\beta = 0.076$, *S.E.* = 0.101, *95% C·I.* [-0.122, 0.274], $z = 0.750$, $p = .453$, $R^2 = 0.00$). The two groups of

Table 2
Primary meta-analysis results.

	# of Sources	k	N	d	95% C·I.	z-value	Sig
1.) Overall Effectiveness of VR Training vs. Control Training	164	184	9007	.541	.450, .631	11.648	**<.001**
Type of Display Hardware							
2a.) Computer Monitor	122	133	6518	.568	.461, .676	10.333	**<.001**
2 b.) Immersive Display	48	54	2564	.462	.299, .625	5.557	**<.001**
Type of Input Hardware							
3a.) Basic Input	61	65	4231	.605	.462, .749	8.266	**<.001**
3 b.) Specialized Input	106	120	4810	.503	.385, .620	8.373	**<.001**
Game Attributes							
4a.) Excluded	136	153	6979	.525	.428, .622	10.604	**<.001**
4 b.) Included	28	31	2028	.609	.362, .855	4.836	**<.001**
Type of Outcome							
5a.) Mental Abilities	23	31	1752	.501	.269, .773	4.236	**<.001**
5 b.) Declarative Know.	45	53	3966	.464	.314, .615	6.044	**<.001**
5c.) Skills	107	112	4242	.576	.454, .698	9.271	**<.001**
Complexity of Outcome							
6a.) Basic	63	78	4986	.480	.339, .622	6.640	**<.001**
6 b.) Moderate	26	29	1418	.423	.276, .570	5.642	**<.001**
6c.) Complex	83	85	3028	.626	.484, .768	8.614	**<.001**
Task-Technology Fit							
7a.) Less	34	43	2206	.207	.063, .352	2.816	**.005**
7 b.) Moderate	51	57	3377	.450	.331, .570	7.384	**<.001**
7c.) More	79	84	3425	.820	.655, .984	9.738	**<.001**
Type of Outcome Measure							
8a.) Reactions	26	28	1965	.502	.223, .782	3.521	**<.001**
8 b.) Learning	134	154	7819	.491	.390, .591	9.572	**<.001**
8c.) Behaviors	26	26	845	.725	.516, .935	6.781	**<.001**
8 d.) Outcomes	2	2	90	.446	-.112, 1.004	1.568	.117
Type of Control Group							
9a.) No Treatment	28	33	1263	.867	.616, 1.119	6.752	**<.001**
9 b.) Non-Equal Treatment	39	41	1507	.803	.565, 1.041	6.616	**<.001**
9c.) Equal Treatment	114	130	7163	.427	.317, .536	7.650	**<.001**
Type of Participant							
10a.) Children & Teenagers	15	20	1713	.527	.257, .798	3.825	**<.001**
10 b.) College Students	47	54	1767	.432	.310, .553	6.941	**<.001**
10c.) Adults	84	89	3032	.614	.451, .776	7.407	**<.001**
10 d.) Special Populations	21	23	730	.614	.354, .874	4.629	**<.001**
Organizational Training							
11a.) Not Org. Training	102	120	6851	.512	.411, .613	9.907	**<.001**
11 b.) Org. Training	63	64	2156	.625	.430, .821	6.267	**<.001**
Type of Source							
12a.) Unpublished	14	16	849	.161	-.128, .449	1.092	.275
12 b.) Thesis & Dissertation	3	5	151	.323	.000, .646	1.960	**.050**
12c.) Conference Paper	1	1	126	.502	.147, .857	2.769	**.006**
12 d.) Article	146	162	7881	.589	.491, .687	11.773	**<.001**

Notes: k = Number of Studies; N = Total Sample Size; d = Standard Difference of Means; 95% C.I. = 95% Confidence Interval; Sig = p-value.

studies were comparable in their results, likewise supporting the decision to analyze them together in the meta-analysis.

As a further supplemental analysis, we assessed whether outcomes differed by occupation in the adult samples.[4] To obtain sufficiently sized groups, we categorized these occupations into medical professionals, military/police, and all other adult participants. VR training programs had no effect on outcomes for military/police ($n = 6$, $k = 6$, $d = 0.030$, *95% C·I.*[-0.650, 0.709], $z = 0.086$, $p = .932$), a moderate effect on outcomes for medical professionals ($n = 70$, $k = 72$, $d = 0.618$, *95% C·I.*[0.459, 0.777], $z = 7.615$, $p < .001$), and a large effect on outcomes for all other adult participants ($n = 9$, $k = 11$, $d = 0.921$, *95% C·I.*[0.159, 1.683], $z = 2.368$, $p = .018$). No dummy-code was statistically significant in our meta-regressions, but the comparison of military/police with medical professionals ($\beta = 0.608$, *S.E.* = 0.324, *95% C·I.*[-0.027, 1.243], $z = 1.877$, $p = .061$, $R^2 = 0.00$) as well as the comparison of military/police with all other adult participants ($\beta = 0.763$, *S.E.* = 0.390, *95% C·I.*[-0.001, 1.527], $z = 1.958$, $p = .050$) closely approached statistical significance. Thus, some variation was seen in outcomes based on these categories.

Hypothesis 10 predicted that the effect of VR training programs relative to alternative comparisons is moderated by the publication year, such that the effect becomes stronger with year. A meta-regression was conducted to determine the effect of year on observed effects of VR training programs, wherein year was treated as a continuous predictor. The effect of year was statistically significant ($\beta = -0.037$, *S.E.* = 0.015, *95% C·I.*[-0.067, −0.007], $z = -2.408$, $p = .016$, $R^2 = 0.02$), but the effect was in a negative direction. Hypothesis 10 was not supported.

Hypothesis 11 predicted that the effect of VR training programs relative to alternative comparisons is moderated by the publication source, such that the effect is larger in published sources (e.g., journal articles) than unpublished sources (e.g., dissertations and conference presentations). VR training programs had small effects in unpublished reports ($d = 0.161$, *95% C·I.*[-0.128, 0.449], $z = 1.092$, $p = .275$) and theses/dissertations ($d = 0.323$, *95% C·I.*[0.000, 0.646], $z = 1.960$, $p = .050$), whereas they had moderate effects in conference papers ($d =$

[4] We thank the reviewers for suggesting this analysis.

0.502, *95% C·I.*[-0.147, 0.857], $z = 2.769$, $p = .006$) and articles ($d = 0.589$, *95% C·I.*[0.491, 0.687], $z = 11.773$, $p < .001$). In the dummy-coded meta-regressions, the dummy code representing the difference of articles and unpublished reports was statistically significant ($\beta = 0.426$, *S.E.* = 0.160, *95% C·I.*[0.112, 0.740], $z = 2.657$, $p = .008$, $R^2 = 0.01$). Hypothesis 11 was partially supported.

We also conducted a final meta-regression that included all predictors described above. Most predictors that were statistically significant in the individual meta-regressions were still significant in this final meta-regression, which included the dummy-codes representing the comparison of less fit and moderate fit ($\beta = 0.528$, *S.E.* = 0.178, *95% C·I.*[0.180, 0.876], $z = 2.973$, $p = .003$, $R^2 = 0.13$), the comparison of less fit and more fit ($\beta = 0.660$, *S.E.* = 0.194, *95% C·I.*[0.280, 1.040], $z = 3.403$, $p < .001$), the comparison of learning and behaviors ($\beta = 0.433$, *S.E.* = 0.200, *95% C·I.*[0.042, 0.825], $z = 2.168$, $p = .030$), the comparison of no treatment control groups and equal-training control groups ($\beta = 0.400$, *S.E.* = 0.180, *95% C·I.*[0.047, 0.752], $z = 2.223$, $p = .026$), and the comparison of articles and unpublished reports ($\beta = -0.477$, *S.E.* = 0.218, *95% C·I.*[-0.905, −0.050], $z = -2.187$, $p = .029$). The effect of year was no longer significant ($\beta = -0.037$, *S.E.* = 0.024, *95% C·I.*[-0.084, 0.011], $z = -1.500$, $p = .134$), as was the dummy-code representing the comparison of moderate fit and more fit ($\beta = 0.132$, *S.E.* = 0.159, *95% C·I.*[-0.179, 0.443], $z = 0.833$, $p = .405$). The dummy code representing the comparison of traditional input hardware and specialized input hardware became statistically significant ($\beta = 0.423$, *S.E.* = 0.172, *95% C·I.*[0.086, 0.761], $z = 2.459$, $p = .014$). These results overall support the robustness of our findings, but our limitations section notes some possible concerns regarding this final analysis.

Lastly, we conducted a series of sensitivity analyses by recalculating all effects using a three-level meta-analytic approach (Supplemental Material B), with influential cases removed (Supplemental Material C), only including adult participants (Supplemental Material D), only including organizational VR training programs (Supplemental Material E), only including VR training programs using immersive hardware (Supplemental Material F), and treating sources that did not specify their output or input hardware as missing observations (Supplemental Material G). None of our inferences differed, suggesting that the current results are robust.

4. Discussion

Our results supported that VR training programs perform better than relative comparisons, producing results that are, on average, half a standard deviation better. Our results also showed, however, that only some tested effects significantly influenced the observed results. Display hardware, input hardware, and game attributes did not significantly influence results. The trained outcome did not significantly influence results, whether differentiated by type or complexity. Likewise, the participant population did not influence results, and the difference of organizational and non-organizational training programs was not significant. These results support our aggregation of effects from all types of studies. Alternatively, task-technology fit significantly influenced results, wherein VR training programs with better fit produced improved outcomes relative to those with worse fit. The measurement approach also influenced results, with behaviors producing a surprisingly larger result than learning. Similarly, publication year had a significant relationship with the magnitude of the observed effects, but it was in the opposite direction than expected – the effectiveness of VR training programs is decreasing over time. The type of comparison group also influenced observed effects, with no treatment control group studies producing larger effects than equal-treatment control group studies. Lastly, the source type influenced results, with articles reporting larger effects than unpublished sources.

Given these results, it should be considered why VR training programs produce desirable effects that did not notably differ across many applications in the current meta-analysis. We suggest that researchers considered the task-technology fit of their applications when developing their VR training programs. VR training programs come in many forms, and they are not "one size fits all". For instance, many VR training programs for surgery skills include custom input devices that mimic surgical tools, and a virtual environment is presented on a monitor (Alaker et al., 2016; Moglia et al., 2016; Vaughan et al., 2016). A computer monitor is used because many surgeries now require cameras to be inserted inside the patient, and the live feed of the camera is presented on a monitor. Input devices that mimic surgical tools along with monitors therefore replicate the transfer environment better than traditional input devices and even more immersive display devices (e.g., CAVE, HMD). Alternatively, many military training programs use immersive displays to develop soldier combat skills, such as target identification and shooting accuracy (Bhagat et al., 2016; Champney et al., 2017; Moshell, 1993; Satava & Jones, 1996). These displays can provide a more realistic experience along with relevant distracting elements (e.g., surrounding visual stimuli), which produces greater fidelity to the transfer environment.

While these applications of VR training programs are notably different, they were applied in situations to maximize task-technology fit. If an HMD was applied in the surgery training, the results may have not been as ideal; if a monitor was applied in the military example, the results again may have not been as ideal. This argument is supported by the findings regarding task-technology fit. The dummy-codes representing differences in fit were, by far, the strongest predictor of the relative effectiveness of VR training programs compared to alternatives, indicating that fit was the most important studied attribute of VR training programs. For these reasons, no combination of VR technologies should be seen as "the best". Instead, they should be seen as complementary approaches that can be utilized appropriately to maximize task-technology fit, as researchers have done in prior studies and practical applications.

4.1. Implications and future directions

The current results strongly encourage the further application of VR training programs in organizational settings. Their effectiveness suggests that the importance of VR training programs to organizations will continue to increase, supporting that these programs should be further researched. The current results also identify the frequency that certain technologies are studied, which highlights opportunities for new research. Despite widespread appeal, immersive display hardware and gamified programs were among the least studied VR training program attributes, drawing clear attention to a present research need.

In the current literature, researchers almost always study VR training programs by comparing their performance to an alternative comparison that does not include VR, which is why we studied the attributes of the VR training program and context as moderating effects. Now that VR has been supported as an apt training medium with merits beyond alternative comparisons, future researchers should conduct more studies that directly compare VR training programs against each other to determine the most effective aspects for a given context. Researchers could obtain a more direct assessment of effective aspects, but they could also provide more robust assessments of mediating and moderating effects (discussed below).

Researchers and practitioners should abandon the notion that certain VR training technologies universally produce improved learning and transfer outcomes, and they should incorporate task-technology fit into their research models and theoretical frameworks. While certain novel technologies, such as HMDs, are surrounded by much fanfare, they need to be applied in contexts that call for their application. More specifically, a significant effect was not seen for the influence of display hardware, input hardware, and game attributes; however, these null results indicate that future research should continue investigating a wide range of VR technologies. That is, because no one technology was most effective in each of these categories, the meta-analytic results suggest that each

technology in these categories may be effective in relevant circumstances that produce higher levels of task-technology fit. We provide suggestions for future research regarding display hardware, input hardware, and game attributes below.

Regarding display and input hardware, researchers in human-computer interaction, human factors, and other similar fields have long studied the user reactions and outcomes of precise changes in these hardware, such as differences in screen resolution or sensitivity of controllers (Maniar et al., 2008; Ullrich et al., 2010). Taxonomies have likewise been created to conceptualize the differences in certain display and input hardware (Basaeed et al., 2007; Wiley, 2000), but they have not been widely applied in fields that typically study training programs. We suggest that researchers should take a more granular view of these technological aspects when studying computer-based training and VR training programs. While significant results were not observed with omnibus comparisons between HMD and monitors in the current meta-analysis, future research may discover significant effects when comparing aspects of HMDs for example. In doing so, relevant psychological theory could be applied to explain observed differences in hardware and answer calls for such theory to be applied in these technology-oriented fields.

Further, immersive display hardware and specialized input hardware were no more effective than their traditional counterparts, overall, but they did have improved outcomes in some studies. Researchers and practitioners should identify the most appropriate applications for these technologies, and fidelity should be tested as a mediating effect between the technologies and outcomes. While we argued that fidelity explains the effect of VR training hardware on outcomes, fidelity has rarely been tested as a mediator of these effects. Measuring fidelity can be an integration of a traditional training topic that is needed in the study of VR training programs (Barata et al., 2015; Ganier et al., 2014; Gavish et al., 2015; Man, 2018).

Regarding game attributes, researchers have already developed sophisticated game attribute taxonomies (Bedwell et al., 2011; Garris et al., 2002; Wilson et al., 2009), and a steady stream of research has begun to study the effects of these individual elements within human resources, management, applied psychology, and other fields (Freina & Ott, 2015; Howard, 2019; Howard & Gutworth, 2020; Seaborn & Fels, 2015; Theng et al., 2015; Xu et al., 2017). While VR training programs are often dichotomized as either gamified or non-gamified, akin to the current meta-analysis, we call for future research to adopt the approach of broader game attribute research and test the influence of individual game attributes in these programs. In doing so, game attributes that are specifically effective for VR training programs could be identified and applied more broadly, and relevant theory could be identified to explain the effectiveness of these elements. Specifically, researchers should continue the trend of applying modern motivational theories, as prior findings may be generalized to modern research on game attributes. Integrating such theory could also address concerns that much of training, computer-based training, and VR training literature is atheoretical (Salas et al., 2009, 2012).

Relatedly, trainees often have positive reactions to training programs utilizing advanced technologies (Kulik, 1994; Sitzmann, 2011; Theng et al., 2015); however, VR training programs had comparable effects on reactions relative to learning and results (Kirkpatrick, 1975, 1979). This finding suggests that, while employees may have enjoyed the VR training program, they may not have necessarily perceived these programs to be effective – perhaps even when these programs were indeed effective as evidenced by their larger effect on behaviors. Future research should investigate why this may be the case. Employees may be skeptical of VR environments as being "too artificial" for training purposes, although such experiences may be exciting or interesting, and they may inherently prefer training tasks performed in the real world. Future research should test whether trainee reactions can be a suitable indicator of training effectiveness when studying VR training programs, and these studies should also investigate whether trainees' preconceptions regarding VR negatively influence their motivation and subsequent performance in such training programs. If biases in these perceptions can be identified and addressed, future research may be able to incorporate reactions as an avenue to improve training success.

In this vein, future research should investigate the wider outcomes of reactions, which is rarely seen in research on VR training programs (but called in broader research of computer-based training; Salas, Rosen, et al., 2009; Salas et al., 2012). Integrating advanced technologies into organizational processes may serve as a signaling effect that the organization is forward-thinking and dedicated to those processes. VR training programs may signal to employees that the development of their skills is valued, thereby improving employee engagement and commitment. Likewise, these programs may improve the organization's public image, and shareholders may have improved perceptions of the organization. Thus, VR training programs may improve employee learning and transfer, but they may also influence broader organizational dynamics.

We did not observe significantly varying effects of VR training programs on different types or complexity of outcomes. This finding was surprising, as it was assumed that VR training programs may be less effective for training straightforward information and more effective for training complex information than other training approaches. These results suggest that VR training programs may be effective across a wider range of applications than typically realized, and the limits of VR training programs should be further tested. We discussed experiential learning theory and identical elements theory in suggesting the outcomes for which VR training programs may be most effective (Kolb et al., 2001; Kolb & Kolb, 2009). While many of our analyses were not significant, focused investigations could identify which propositions of these theories are appropriate for understanding VR training programs. The entire scope of these theories may not be relevant, but specific aspects may be beneficial to study VR training programs. Future researchers of VR training programs should also apply theories relevant to modern technologies but rarely seen in the study of training. For instance, social presence theory has been used to detail the dynamics of technological immersion and presence, suggesting that presence causes individuals to be more psychologically engaged (Short et al., 1976); however, such theories have not been applied to understand training and development.

On the other hand, task-technology fit is a very strong influence on VR training success, demanding future applications of task-technology fit theory. To our knowledge, no author has applied the entire scope of task-technology fit theory to understand VR training programs, which should be immediately investigated. Recent authors have also expanded this theory, though, which likewise merits attention. Howard and Rose (2019) recently coined the construct of task-technology misfit to describe the degree in which a technology does *not* assist in performing a portfolio of tasks. Task-technology misfit exists in two forms: when the technology does not provide enough features to complete the task at hand (Too Little) and when the technology provides too many features to complete the task at hand (Too Much). In general, Too Little produces worse outcomes than Too Much, but both produce worse outcomes than task-technology fit (Howard & Rose, 2019). These authors suggested that researchers may be wary to study task-technology misfit, as it requires acknowledging that the studied application may not produce an ideal pairing of tasks with technologies; however, exploring the limits of VR training programs provides an ideal context for the integration of task-technology misfit, which could benefit research on both VR training programs as well as task-technology misfit. Researchers could identify the mechanisms that may cause VR training programs to be inappropriate for certain contexts (e.g., Too Little, Too Much) as well as the antecedents and outcomes of task-technology misfit that may be difficult to study in other applications. Thus, integrating VR training programs into the study of task-technology misfit may benefit both research domains.

Similar considerations may also explain the observed differences in

VR effectiveness by occupation. VR training programs were less effective in police and military applications, which may be due to the misfit and/or objectives of these applications. Many of these applications trained skills that are difficult to replicate even in a VR environment. For instance, Butavicius et al. (2012) found that their VR parachute training did not produce significantly better live jump attempts than classroom-based instruction, likely due to the difficulty in including all relevant stimuli in a VR training program. In other cases, government entities have placed great resources in creating realistic real-world training programs that would be difficult to improve upon due to their satisfactory fit, but VR training programs can provide cheaper alternatives with similar fit and comparable outcomes. For example, Bertram et al. (2015) showed that their VR training program to train police officers to interpret and respond to helicopter crew commands was more effective than no training at all, but it produced similar results to the traditional training with real-world actors. In this case, the VR and traditional training programs may have produced similar fit, but the VR training could provide greater cost-to-benefit utility in future uses because it would not require actors. Researchers should recognize the various objectives of training programs, such as cost-to-benefit utility, when assessing their relative merits.

Alternatively, VR training programs for medical professionals were significantly more effective than relative alternatives, but they were less effective than VR training programs applied for all other adult samples. This finding may be due to the familiarity of high-tech applications for medical professionals. That is, medical professionals may be accustomed to VR training programs and other high-tech training context, and they may be less enthusiastic to participate in such opportunities (Alaker et al., 2016; Moglia et al., 2016; Vaughan et al., 2016). On the other hand, general adult participants may still be excited by VR training programs, and they may be more motivated to perform the training tasks. We again urge researchers to study both task-technology fit and excitement in the contexts of VR training programs.

Additionally, the effect of publication year was statistically significant, but the direction of this relationship suggests that VR training programs are becoming less effective over time. We assert two possibilities for this finding. First, trainees may be more accustomed to VR. Whereas the technology was much more novel in initial investigations of VR training programs, it is becoming more commonplace in recent years (Alaker et al., 2016; Moglia et al., 2016; Vaughan et al., 2016). Trainees may be less excited to participate in VR training programs, resulting in poorer trainee motivation. If true, then researchers should continuously assess the efficacy of VR training programs over time, but also develop more in-depth investigations into the relation of trainee excitement and learning outcomes. Excitement may be an understudied key predictor of VR training outcomes. Second, researchers may be expanding the contexts that VR training programs are applied. In early VR studies, authors restricted applications of the technology to contexts that would clearly benefit from the technology, such as aviation training. In more recent studies, researchers are testing whether the technology can benefit applications with less-obvious needs. For instance, several authors have designed VR training programs to train street-crossing abilities (Josman et al., 2008; McComas et al., 2002; Schwebel & McClure, 2010). We urge authors to continue testing the boundaries of VR training programs.

Lastly, some aspects of the methodological design had a significant effect, although most aspects of the studied tasks and technologies did not. The choice of control group influenced the magnitude of observed results. Researchers have increasingly recognized that effect size interpretations should be dependent on the context being studied (Bosco et al., 2015; Gignac & Szodorai, 2016), and therefore researchers using no treatment control groups should interpret their effects differently than those using equal treatment control groups. Likewise, the effect sizes reported in articles were much larger than those reported in unpublished reports. While recent authors have begun to cast doubt over the file-drawer problem (Dalton et al., 2012), our results suggest that null or negative results are less likely to be seen in published VR training studies. Researchers should keep this in mind when interpreting research on VR training.

4.2. Limitations

Meta-analyses are bound by the researchers' methodological and analytical decisions. For instance, we considered all articles that did not report their output hardware as utilizing a monitor, whereas we considered all articles that did not report their input hardware as utilizing a keyboard and mouse. We did so because most studies of immersive hardware and/or specialized input devices use the technologies as the focal point of the articles, but articles typically only mention briefly in their methods if they utilized a monitor, keyboard, and mouse. We believed that authors would be more likely to not report their utilized technologies in this latter instance, but such assumptions cannot be guaranteed. For this reason, we conducted a reanalysis of our results when treating these studies as missing observations in Supplemental Material G, and we performed additional analyses for other potential concerns in Supplemental Material B, C, D, E, and F. Each analysis replicated those reported in the primary text, supporting the robustness of our results. Nevertheless, unrecognized decisions could sway our results, and we encourage future authors to reanalyze our results utilizing our database (Supplemental Material A).

Most of our hypotheses were tested via meta-regression. We placed a larger focus on our meta-regressions with a single predictor because prior research has supported the statistical power of similar meta-regressions (Sánchez-Meca & Marín-Martínez, 1998Sánchez-Meca & Marín-Martínez, 1998; Schmidt, 2017; Tipton et al., 2019). López-López et al. (2014) supported the accuracy of meta-regression when the number of included studies was 80, which the current article greatly exceeded ($k = 184$). The authors also showed that the accuracy of meta-regressions increased with average sample sizes even beyond 50, which did exceed our average sample size ($\overline{N} = 49$); however, the influence of sample size appeared to be largely negligible when the number of included studies was 80, again supporting the present meta-regressions. While we reported a meta-regression that included all predictors for the sake of completeness, we were uncertain about the statistical power and accuracy of including a relatively large number of predictors in a single meta-regression. Fewer prior studies have investigated this possible concern with meta-regressions, and some authors have argued that the number of required studies to perform a sufficiently powered meta-regression is almost impossible to obtain with even a relatively modest number of predictors (e.g., 10) (Schmidt, 2017; Tipton et al., 2019). For this reason, we placed less of a focus on this final meta-regression, although it largely replicated the results of the individual meta-regressions.

5. Conclusion

We supported that VR training programs are, on average, more effective than alternative training programs, but very few boundary conditions significantly influenced these results. We suggest that these repeated null results provide strong support for the overall effectiveness of VR training programs, and they also indicate that no one type of VR training program is superior. Instead, any attribute of VR training programs could be beneficial in the appropriate circumstance. We call for many future directions to study these individual elements, and we also call for the deeper integration of task-technology fit and misfit in VR training research.

Credit author statement

All authors agree to their inclusion and order of the authorship as listed on the title page. All authors were involved in the writing and/or revisions of the current work, which merits their authorship. The first

two authors were involved with the methodology and statistical analyses.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2021.106808.

References

Alaker, M., Wynn, G., & Arulampalam, T. (2016). Virtual reality training in laparoscopic surgery: Systematic review & meta-analysis. *International Journal of Surgery, 29*, 85–94.

Alaraj, A., Lemole, M. G., Finkle, J. H., Yudkowsky, R., Wallace, A., Luciano, C., & Charbel, F. T. (2011). Virtual reality training in neurosurgery: Review of current status and future applications. *Surgical Neurology International, 2*, 52.

Ammenwerth, E., Iller, C., & Mahler, C. (2006). IT-adoption and the interaction of task, technology and individuals: A fit framework and a case study. *BMC Medical Informatics and Decision Making, 6*(1), 3.

Auerbach, A. (1999). Making decisions under stress: Implications for individual and team training. *Personnel Psychology, 52*(4), 1050.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin & Review, 22*(2), 366–377.

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*(1), 63–105.

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

Basaeed, E., Berri, J., Zemerly, M. J., & Benlamri, R. (2007). Learner-centric context-aware mobile learning. *IEEE Multidisciplinary Engineering Education Magazine, 2*(2), 30–33.

Bedwell, W., Pavlas, D., Heyne, K., Lazzara, E., & Salas, E. (2012). Toward a taxonomy linking game attributes to learning: An empirical study. *Simulation & Gaming, 43*(6), 729–760.

Bertram, J., Moskaliuk, J., & Cress, U. (2015). Virtual training: Making reality work? *Computers in Human Behavior, 43*, 284–292.

Bhagat, K. K., Liou, W. K., & Chang, C. Y. (2016). A cost-effective interactive 3D virtual reality system applied to military live firing training. *Virtual Reality, 20*(2), 127–140.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*(4), 1065–1105.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*(2), 431–449.

Boxer, P., Groves, C. L., & Docherty, M. (2015). Video games do indeed influence children and adolescents' aggression, prosocial behavior, and academic performance: A clearer reading of Ferguson (2015). *Perspectives on Psychological Science, 10*(5), 671–673.

Brunner, W. C., Korndorffer, J. R., Jr., Sierra, R., Massarweh, N. N., Dunne, J. B., Yau, C. L., & Scott, D. J. (2004). Laparoscopic virtual reality training: Are 30 repetitions enough? 1. *Journal of Surgical Research, 122*(2), 150–156.

Butavicius, M., Vozzo, A., Braithwaite, H., & Galanis, G. (2012). Evaluation of a virtual reality parachute training simulator: Assessing learning in an off-course augmented feedback training schedule. *The International Journal of Aviation Psychology, 22*(3), 282–298.

Camp, C. L., Krych, A. J., Stuart, M. J., Regnier, T. D., Mills, K. M., & Turner, N. S. (2016). Improving resident performance in knee arthroscopy: A prospective value assessment of simulators and cadaveric skills laboratories. *JBJS, 98*(3), 220–225.

Carter, M., & Beier, M. E. (2010). The effectiveness of error management training with working-aged adults. *Personnel Psychology, 63*(3), 641–675.

Champney, R. K., Stanney, K. M., Milham, L., Carroll, M. B., & Cohn, J. V. (2017). An examination of virtual environment training fidelity on training effectiveness. *International Journal of Learning Technology, 12*(1), 42–65.

Cheng, Y., & Wang, S. H. (2011). Applying a 3D virtual learning environment to facilitate student's application ability–The case of marketing. *Computers in Human Behavior, 27*(1), 576–584.

Cheung, M. (2015). *Meta-analysis: A structural equation modeling approach*. John Wiley & Sons.

Cheung, M., & Cheung, S. (2016). Random-effects models for meta-analytic structural equation modeling: Review, issues, and illustrations. *Research Synthesis Methods, 7*(2), 140–155.

Cohen, N. J., Eichenbaum, H., Deacedo, B. S., & Corkin, S. (1985). Different memory systems underlying acquisition of procedural and declarative knowledge. *Annals of the New York Academy of Sciences, 444*(1), 54–71.

Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., & Karama, S. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence, 41*(5), 712–727.

Cordeil, M., Dwyer, T., Klein, K., Laha, B., Marriott, K., & Thomas, B. H. (2017). Immersive collaborative analysis of network connectivity: Cave-style or head-mounted display? *IEEE Transactions on Visualization and Computer Graphics, 23*(1), 441–450.

Cruz-Neira, C., Sandin, D., & DeFanti, T. (1993). Surround-screen projection-based virtual reality: The design and implementation of the CAVE. *Proceedings of the 20th annual conference on computer graphics and interactive techniques*.

Dalton, D. R., Aguinis, H., Dalton, C. M., Bosco, F. A., & Pierce, C. A. (2012). Revisiting the file drawer problem in meta-analysis: An assessment of published and nonpublished correlation matrices. *Personnel Psychology, 65*(2), 221–249.

Di Serio, Á., Ibáñez, M. B., & Kloos, C. D. (2013). Impact of an augmented reality system on students' motivation for a visual art course. *Computers & Education, 68*, 586–596.

Didehbani, N., Allen, T., Kandalaft, M., Krawczyk, D., & Chapman, S. (2016). Virtual reality social cognition training for children with high functioning autism. *Computers in Human Behavior, 62*, 703–711.

Doyle, T. E., Booth, J. M. J., Musson, D. M., & McMaster University. (2015). *Closing the design loop in first-year engineering: Modelling and simulation for iterative design*. Higher Education Quality Council of Ontario.

Dubovi, I., Levy, S. T., & Dagan, E. (2017). Now I know how! the learning process of medication administration among nursing students with non-immersive desktop virtual reality simulation. *Computers & Education, 113*, 16–27.

Ferro, L. S., Walz, S. P., & Greuter, S. (2013). Towards personalised, gamified systems: An investigation into game design, personality and player typologies. In *Proceedings of the 9th australasian conference on interactive entertainment: Matters of life and death* (Vol. 7).

Freina, L., & Ott, M. (2015). *A literature review on immersive virtual reality in education: State of the art and perspectives. 1*. eLearning & Software for Education.

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.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming, 33*(4), 441–467.

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*(6), 778–798.

Gigante, M. A. (1993). Virtual reality: Definitions, history and applications. *Virtual reality systems*. Academic Press.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74–78.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly*, 213–236.

Grantcharov, T. P., Bardram, L., Funch-Jensen, P., & Rosenberg, J. (2003). Learning curves and impact of previous operative experience on performance on a virtual reality simulator to test laparoscopic surgical skills. *The American Journal of Surgery, 185*(2), 146–149.

Gurusamy, K., Aggarwal, R., Palanivelu, L., & Davidson, B. (2009). Virtual reality training for surgical trainees in laparoscopic surgery. *Cochrane Database of Systematic Reviews, 1*.

Haeger, M., Bock, O., Memmert, D., & Hüttermann, S. (2018). Can driving-simulator training enhance visual attention, cognition, and physical functioning in older adults? *Journal of Aging Research*, 2018.

Häll, L., Söderström, T., Ahlqvist, J., & Nilsson, T. (2011). Collaborative learning with screen-based simulation in health care education: An empirical study of collaborative patterns and proficiency development. *Journal of Computer Assisted Learning, 27*(5), 448–461.

Hammick, J., & Lee, M. (2014). Do shy people feel less communication apprehension online? The effects of virtual reality on the relationship between personality characteristics and communication outcomes. *Computers in Human Behavior, 33*, 302–310.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science, 24*(12), 2409–2419.

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

Hochmitz, I., & Yuviler-Gavish, N. (2011). Physical fidelity versus cognitive fidelity training in procedural skills acquisition. *Human Factors, 53*(5), 489–501.

Hogle, N. J., Briggs, W. M., & Fowler, D. L. (2007). Documenting a learning curve and test-retest reliability of two tasks on a virtual reality training simulator in laparoscopic surgery. *Journal of Surgical Education, 64*(6), 424–430.

van Houwelingen, H., Arends, L., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine, 21*(4), 589–624.

Howard, M. C. (2017). A meta-analysis and systematic literature review of virtual reality rehabilitation programs. *Computers in Human Behavior, 70*, 317–327.

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

Howard, M. C., & Gutworth, M. B. (2020). A meta-analysis of virtual reality training programs for social skill development. *Computers & Education, 144*, 103707.

Howard, M. C., & Rose, J. C. (2019). *Refining and extending task–technology fit theory: Creation of two task–technology fit scales and empirical clarification of the construct*. Information & Management.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8*, 275–292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Jak, S. (2015). *Meta-analytic structural equation modelling*. Dordrecht, Neth: Springer.

Jensen, K., Ringsted, C., Hansen, H., Petersen, R., & Konge, L. (2014). Simulation-based training for thoracoscopic lobectomy: A randomized controlled trial. *Surgical Endoscopy, 28*(6), 1821–1829.

Josman, N., Ben-Chaim, H. M., Friedrich, S., & Weiss, P. L. (2008). Effectiveness of virtual reality for teaching street-crossing skills to children and adolescents with autism. *International Journal on Disability and Human Development, 7*(1), 49–56.

Julier, S., King, R., Colbert, B., Durbin, J., & Rosenblum, L. (1999). The software architecture of a real-time battlefield visualization virtual environment. *Proceedings IEEE Virtual Reality*, 29–36.

Keith, N., & Frese, M. (2008). Effectiveness of error management training: A meta-analysis. *Journal of Applied Psychology, 93*(1), 59.

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. (2013). Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the Meta-Analytic Reporting Standards). *Journal of Business and Psychology, 28*(2), 123–143.

Kirkpatrick, D. L. (1975). *Evaluating training programs*. McGraw-Hill Education.

Kirkpatrick, D. L. (1979). Techniques for evaluating training programs. *Training & Development Journal, 33*(6), 78–92.

Kohler, T., Fueller, J., Stieger, D., & Matzler, K. (2011). Avatar-based innovation: Consequences of the virtual co-creation experience. *Computers in Human Behavior, 27*(1), 160–168.

Kolb, D. A., Boyatzis, R. E., & Mainemelis, C. (2001). Experiential learning theory: Previous research and new directions. *Perspectives on Thinking, Learning, and Cognitive Styles, 1*(8), 227–247.

Kolb, A. Y., & Kolb, D. A. (2009). *Experiential learning theory: A dynamic, holistic approach to management learning, education and development*. The SAGE Handbook of Management Learning, Education and Development.

Kowalewski, K., Garrow, C., Proctor, T., Preukschas, A., Friedrich, M., Müller, P., & Nickel, F. (2018). LapTrain: Multi-modality training curriculum for laparoscopic cholecystectomy—results of a randomized controlled trial. *Surgical Endoscopy, 32*(9).

Kozlowski, S. W., & DeShon, R. P. (2004). A psychological fidelity approach to simulation-based training: Theory, research and principles. *Scaled worlds: Development, validation, and applications*, 75–99.

Kruglikova, I., Grantcharov, T. P., Drewes, A. M., & Funch-Jensen, P. (2010). The impact of constructive feedback on training in gastrointestinal endoscopy using high-fidelity virtual-reality simulation: A randomised controlled trial. *Gut, 59*(2), 181–185.

Kulik, J. A. (1994). Meta-analytic studies of findings on computer-based instruction. *Technology Assessment in Education and Training, 1*, 9–34.

Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & Gaming, 45*(6), 752–768.

Landers, R., Auer, E., Collmus, A., & Armstrong, M. (2018). Gamification science, its history and future: Definitions and a research agenda. *Simulation & Gaming, 49*(3), 315–337.

Lee, S. H., Sergueeva, K., Catangui, M., & Kandaurova, M. (2017). Assessing Google Cardboard virtual reality as a content delivery system in business classrooms. *The Journal of Education for Business, 92*(4), 153–160.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine, 6*(7), Article e1000100.

Li, H., Chan, G., & Skitmore, M. (2012). Multiuser virtual safety training system for tower crane dismantlement. *Journal of Computing in Civil Engineering, 26*(5), 638–647.

Li, W., Grossman, T., & Fitzmaurice, G. (2012). October). GamiCAD: A gamified tutorial system for first time autocad users. *Proceedings of the 25th annual ACM symposium on User interface software and technology* (pp. 103–112).

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE publications, Inc.

López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology, 67*(1), 30–48.

Man, D. W. (2018). Virtual reality-based cognitive training for drug abusers: A randomised controlled trial. *Neuropsychological Rehabilitation*, 1–18.

Maniar, N., Bennett, E., Hand, S., & Allan, G. (2008). The effect of mobile phone screen size on video based learning. *Journal of Software, 3*(4), 51–61.

McComas, J., MacKay, M., & Pivik, J. (2002). Effectiveness of virtual reality for teaching pedestrian safety. *CyberPsychology and Behavior, 5*(3), 185–190.

McMahan, R. P., Bowman, D. A., Zielinski, D. J., & Brady, R. B. (2012). Evaluating display fidelity and interaction fidelity in a virtual reality game. *IEEE Transactions on Visualization and Computer Graphics, 18*(4), 626–633.

Merchant, Z., Goetz, E. T., Keeney-Kennicutt, W., Cifuentes, L., Kwok, O. M., & Davis, T. J. (2013). Exploring 3-D virtual reality technology for spatial ability and chemistry achievement. *Journal of Computer Assisted Learning, 29*(6), 579–590.

Moglia, A., Ferrari, V., Morelli, L., Ferrari, M., Mosca, F., & Cuschieri, A. (2016). A systematic review of virtual reality simulators for robot-assisted surgery. *European Urology, 69*(6), 1065–1080.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group.. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), Article e1000097.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1), 1.

Moshell, M. (1993). Three views of virtual reality: Virtual environments in the US military. *Computer, 26*(2), 81–82.

Murray, C. D., Fox, J., & Pettifer, S. (2007). Absorption, dissociation, locus of control and presence in virtual reality. *Computers in Human Behavior, 23*(3), 1347–1354.

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

Pan, Z., Cheok, A. D., Yang, H., Zhu, J., & Shi, J. (2006). Virtual reality and mixed reality for virtual learning environments. *Computers & Graphics, 30*(1), 20–28.

Pellas, N. (2014). The influence of computer self-efficacy, metacognitive self-regulation and self-esteem on student engagement in online learning programs: Evidence from the virtual world of second life. *Computers in Human Behavior, 35*, 157–170.

Plante, T. G., Aldridge, A., Bogden, R., & Hanelin, C. (2003). Might virtual reality promote the mood benefits of exercise? *Computers in Human Behavior, 19*(4), 495–509.

Ren, S., McKenzie, F. D., Chaturvedi, S. K., Prabhakaran, R., Yoon, J., Katsioloudis, P. J., & Garcia, H. (2015). Design and comparison of immersive interactive learning and instructional techniques for 3D virtual laboratories. *Presence, 24*(2), 93–112.

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution, 59*(2), 464–468.

Rosen, B., Furst, S., & Blackburn, R. (2006). Training for virtual teams: An investigation of current practices and future needs. *Human Resource Management, 45*(2), 229–247.

Rothstein, H. R., & Bushman, B. J. (2015). Methodological and reporting errors in meta-analytic reviews make other meta-analysts angry: A commentary on ferguson (2015). *Perspectives on Psychological Science, 10*(5), 677–679.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication bias in meta-analysis*. Publication Bias in Meta-Analysis.

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52*(1), 471–499.

Salas, E., Rosen, M., Held, J., & Weissmuller, J. (2009). Performance measurement in simulation-based training: A review and best practices. *Simulation & Gaming, 40*(3).

Salas, E., Tannenbaum, S. I., Kraiger, K., & Smith-Jentsch, K. A. (2012). The science of training and development in organizations: What matters in practice. *Psychological Science in the Public Interest, 13*(2), 74–101.

Salas, E., Wildman, J. L., & Piccolo, R. F. (2009). Using simulation-based training to enhance management education. *The Academy of Management Learning and Education, 8*(4), 559–573.

Sánchez-Meca, J., & Marín-Martínez, F. (1998). Testing continuous moderators in meta-analysis: A comparison of procedures. *British Journal of Mathematical and Statistical Psychology, 51*(2), 311–326.

Satava, R. M., & Jones, S. B. (1996). An integrated medical virtual reality program. The military application. *IEEE Engineering in Medicine and Biology Magazine, 15*(2), 94–97.

Schmidt, F. L. (2017). *Statistical and measurement pitfalls in the use of meta-regression in meta-analysis*. Career Development International.

Schmidt, F. L., & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. *Current Directions in Psychological Science, 2*(1), 8–9.

Schwebel, D. C., & McClure, L. A. (2010). Using virtual reality to train children in safe street-crossing skills. *Injury Prevention, 16*(1), e1.

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies, 74*, 14–31.

Seixas-Mikelus, S. A., Adal, A., Kesavadas, T., Baheti, A., Srimathveeravalli, G., Hussain, A., & Guru, K. A. (2010). Can image-based virtual reality help teach anatomy? *Journal of Endourology, 24*(4), 629–634.

Selvander, M., & Åsman, P. (2012). Virtual reality cataract surgery training: Learning curves and concurrent validity. *Acta Ophthalmologica, 90*(5), 412–417.

Seymour, N. E., Gallagher, A. G., Roman, S. A., O'brien, M. K., Bansal, V. K., Andersen, D. K., & Satava, R. M. (2002). Virtual reality training improves operating room performance: Results of a randomized, double-blinded study. *Annals of Surgery, 236*(4), 458.

Shin, D. (2018). Empathy and embodied experience in virtual environment: To what extent can virtual reality stimulate empathy and embodied experience? *Computers in Human Behavior, 78*, 64–73.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. London, England: Wiley.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology, 64*(2), 489–528.

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication, 42*(4), 73–93.

Su, C., & Cheng, C. (2013). 3D game-based learning system for improving learning achievement in software engineering curriculum. *Turkish Online Journal of Educational Technology, 12*(2), 1–12.

Taber, M. J. (2014). Simulation fidelity and contextual interference in helicopter underwater egress training: An analysis of training and retention of egress skills. *Safety Science, 62*, 271–278.

Tergas, A. I., Sheth, S. B., Green, I. C., & Giuntoli, R. L. (2013). A pilot study of surgical training using a virtual robotic surgery simulator. *Journal of the Society of Laparoendoscopic Surgeons: Journal of the Society of Laparoendoscopic Surgeons, 17*(2), 219.

Theng, Y. L., Lee, J. W., Patinadan, P. V., & Foo, S. S. (2015). The use of videogames, gamification, and virtual environments in the self-management of diabetes: A systematic review of evidence. *Games for Health Journal, 4*(5), 352–361.

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*, 1559–1573.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods, 10*(2), 161–179.

Tüzün, H., & Özdinç, F. (2016). The effects of 3D multi-user virtual environments on freshmen university students' conceptual and spatial learning and presence in departmental orientation. *Computers & Education, 94*, 228–240.

Ullrich, C., Shen, R., Tong, R., & Tan, X. (2010). A mobile live video learning system for large-scale learning—system design and evaluation. *IEEE Transactions on Learning Technologies, 3*(1), 6–17.

Våpenstad, C., Hofstad, E. F., Bø, L. E., Kuhry, E., Johnsen, G., Mårvik, R., & Hernes, T. N. (2017). Lack of transfer of skills after virtual reality simulator training with haptic feedback. *Minimally Invasive Therapy & Allied Technologies, 26*(6), 346–354.

Vaughan, N., Dubey, V., Wainwright, T., & Middleton, R. (2016). A review of virtual reality based training simulators for orthopaedic surgery. *Medical Engineering & Physics, 38*(2), 59–71.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125.

Wiley, D. A. (2000). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. *The Instructional Use of Learning Objects, 2830*(435), 1–35.

Wilson, K. A., Bedwell, W. L., Lazzara, E. H., Salas, E., Burke, C. S., Estock, J. L., & Conkey, C. (2009). Relationships between game attributes and learning outcomes: Review and research proposals. *Simulation & Gaming, 40*(2), 217–266.

Xu, F., Buhalis, D., & Weber, J. (2017). Serious games and the gamification of tourism. *Tourism Management, 60*, 244–256.

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer, 38*(9), 25–32.