Policy Research Working Paper 9587

Meta-Analysis Assessing the Effects of Virtual Reality Training on Student Learning and Skills Development

Diego F. Angel-Urdinola
Catalina Castillo-Castro
Angela Hoyos

WORLD BANK GROUP
Education Global Practice
March 2021

Policy Research Working Paper 9587

Abstract

Training using virtual reality has been applied in many fields of education, but primarily in the fields of health and safety, engineering and technical education, and general education. Numerous studies assessing the use of immersive training in education have yielded promising results in educational outcomes, but there is not yet in the literature a systematic analysis of the effects of virtual reality training on student learning. This paper presents a meta-analysis of the results of available studies that assess virtual reality training's impact on student learning and skills development, and which rely on robust evaluation methods. The study's primary purpose is to identify the extent to which immersive training can successfully develop students' skills across different fields of education and the size of the effects encountered. The analysis presented here relies on 31 primary studies and more than 90 experiments. The results indicate that, on average, virtual reality training is more effective than traditional training in developing technical, practical, and socio-emotional skills. The results are particularly promising in fields related to health and safety, engineering, and technical education. The results also indicate that students who are exposed to virtual reality training are more efficient in using inputs and time and/or avoiding performance errors than students receiving traditional training.

This paper is a product of the Education Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at dangelurdinola@worldbank.org.

Meta-Analysis Assessing the Effects of Virtual Reality Training on Student Learning and Skills Development

Diego F. Angel-Urdinola, Catalina Castillo-Castro, Angela Hoyos

The World Bank [1]

JEL Classification: I20, I24

Keywords: Virtual reality, Education, Learning, Skills Development.

Corresponding Author: Diego F. Angel-Urdinola (dangelurdinola@worldbank.org)

[1] This paper is a product of the Education Global Practice. The paper was prepared as a background paper for the *Pilot Program to Test Virtual Reality Training Programs for Technological and Technical courses in Higher Education (TF0A8313)*, supported by the Korea World Bank Partnership Facility (KWPF). The authors acknowledge useful comments and support from Christine H. Joo, Robert Hawkins, and Michael Trucano.

I. Introduction

Recent events, such as the health pandemic introduced by the COVID-19 virus, have contributed to speed up alternative mechanisms to offer digital instruction to substitute and complement in-class instruction. The expansion of digital and computer assisted learning is becoming a global trend, making it of extreme importance to identify technology tools that work, while being scalable and cost-effective (Escueta, Quan, Nickow, & Oreopoulos, 2017). Even before the pandemic, it had become particularly challenging for education systems to supply digital learning opportunities that provide students the hands-on pedagogical experiences necessary to develop practical skills, especially for programs that require the use of laboratories.

Virtual reality (VR) training is often known as the process of learning in a simulated or artificial environment. VR training has existed in the realm of education for over half a century but has dramatically expanded over the past 15 years as VR simulators are becoming less expensive to develop and increasingly realistic. The term VR applies to computer-simulated environments that can imitate physical presence in places in the real world, as well as in imaginary worlds (Lorenzo, Pomares, & Lledó, 2013), and simulate the illusion of participation in a synthetic environment with an external observation of such surroundings (Gigante, 1993). VR simulations can be constructed employing 3D graphics using a desktop computer (non-immersive) or using a head-mounted display (immersive) (Makransky, Terkildsen, & Mayer, 2019). In non-immersive VR, the simulated environment is displayed on a conventional computer with sound and graphics coming through the computer's speaker and monitor, and the interaction is controlled through a regular computer mouse. Immersive VR uses a head-mounted display in which a high graphical fidelity screen is mounted in front of the user's eyes with separate lenses for each eye and with sound delivered through earphones. The interactions in the context of high-immersion VR are controlled through head-motion tracking in conjunction with a computer system that allows users to look around a simulated 360-degree environment.

In some educational fields, the development of adequate cognitive, technical, and socio-emotional skills remains a challenge for trainees and their tutors, partly because of the limited availability of hands-on training or access to proper content and learning situations. As a response, educators are starting to rely on VR simulations to develop learning experiences that would otherwise not be easily accessible to students. VR simulations can provide students practical

training opportunities without pressure, danger, and allowing repeated interventions. Also, VR simulations can provide students access to situations and learning environments (such as traveling within a cell, simulated scenarios for public speaking, among others) that would otherwise be very difficult or impossible to access. Such opportunities can accelerate students' learning curve in a simulated environment, reproducing real-life conditions and situations without time or space limitations and much fewer risks than real environments. In addition, VR simulations offer the great advantage of providing students and teachers a standardized, reproducible environment for repeated and optimized training (Apostolellis, Bowman, & Chmiel, 2018; Cheung, Fong, Fong, & Wang, 2013; Ferracani, Pezzatini, & Del Bimbo, 2014; Huang, Rauch, & Liaw, 2010; Sharma, Agada, & Ruffin, 2013).

Another advantage of using VR simulations is that gamification, performance metrics, and collaborative features (using avatars) can be embedded in the software, enabling continuous peer interaction, active learning, enjoyment, and performance feedback – all elements that enhance proficiency-based training. Indeed, constructivism is often cited as a theoretical framework that supports the implementation of learning in virtual environments. Constructivism suggests that students learn by constructing knowledge and incorporating it into their existing knowledge structure. Thus, constructivist learning environments can increase active learning, motivation, interactivity, and personalized learning (Madathil et al., 2017).

Proponents of VR simulations claim that higher motivation and presence are the main two channels through which VR training simulations can influence student learning (Mikropoulos, & Natsis, 2011). As a result, VR simulations have been regarded as a pedagogical method with the potential to increase student learning, as they increase self-motivation to learn and allow embedding to the educational experience constructivist pedagogy, collaboration, and gamification (Kavanagh, Luxton-Reilly, Wuensche, & Plimmer, 2017).

The impact of media on student learning outcomes has been highly debated among educational technologists where much of the prior literature has shown no significant difference between technology-based and traditionally delivered instruction and media. However, the counterargument contends that using the correct media could impact students' cognitive skills and that the media itself is a critical component of instructional design (Madathil et al., 2017).

Given the rising importance of identifying digital education platforms that work, this paper conducts a meta-analysis of the results of available experiments that assess the impact of VR training on learning and skills development. The study's primary purpose is to identify the extent to which VR training is conducive to learning and skills development. A secondary objective is to assess, to the extent possible, if VR training is also an efficient mechanism to deliver training. The analysis presented here relies on a total of 31 primary studies and over 90 different experiments.

There has not been a systematic assessment of the effects of VR training on learning, other than in the field of surgical education in the early 2000s (Haque, & Srinivasan, 2006). This study was conducted with a limited number of studies and focused on assessing the extent to which VR training could help students perform surgical procedures faster (i.e. improve their time-in-task). To bridge the knowledge gap, this study focuses on a more recent time period (2005-2020), during which VR technology has significantly evolved, and covers other fields such as engineering, science, technical education, and general education. Moreover, this study analyzes the effects of VR more holistically as a mechanism to develop cognitive, technical, and socio-emotional skills. As such, the findings of this paper represent an essential contribution to the literature and intend to guide education institutions and policy makers to have more information about the effects of VR training as they expand their offer of digital learning opportunities to students.

Based on the information available, our research shows that VR training is, on average, more effective than traditional training as a mechanism to develop students' technical, practical, and socio-emotional skills. Results are particularly promising in fields related to health and safety, engineering, and technical education. Results reveal that for each additional hour (¼ hour) of VR training, students score 3 percent higher in technical (cognitive) learning assessments than students exposed to the same curricular content delivered through traditional training methods. Results also indicate that students exposed to VR instruction report on average 30 percent higher scores in socio-emotional skills assessments after completing their training than their peers receiving traditional instruction. Results also suggest that students exposed to VR training are up to 30 percent more efficient using inputs, time, and/or avoiding performance errors than students exposed to traditional training, per additional hour of instruction.

The paper is structured as follows. Section II provides a review of the literature on the use of VR for pedagogical purposes. Section III presents the data and methodology used to conduct the meta-analysis. Section IV presents the results of the meta-analysis and intends to quantify the observed effects of VR training on students' learning outcomes and skills development. The conclusion follows in Section V.

II. Literature review

Recent studies assessing the effect of the use of VR simulations in education show promising findings, in different areas, from increased time-on-task, enjoyment, motivation, and learning. Nonetheless, there is not a recent systematic analysis of the effects of VR training on student learning and skills development. A recent review of the literature (Kavanagh, Luxton-Reilly, Wuensche, & Plimmer, 2017) shows that VR simulations are used in many education fields, but primarily in health and surgical education, engineering and technical education, and general education (mainly in STEM related fields). Since each of these fields uses VR with slightly different pedagogical purposes, our literature review discusses each of these fields independently.

2.1. VR training for health and safety

The use of VR training has shown great potential within the field of health, especially in the area of surgery, as it offers trainees the opportunity to practice several surgical procedures in a safe environment and at a comparatively low cost. Simulators provide excellent benefits to surgical trainees by allowing for repeated practice of a specific skill set in a controlled and safe environment, before ever entering the operating room. VR training allows developing surgical training experience that can enable junior trainees to undertake self-directed training while practicing and learning the fundamentals of surgery procedures without putting patients at risk and without needing supervision from an attending surgeon. Also, VR training can provide junior trainees relevant experience at an early stage in their surgical training while giving them an exposure to otherwise scarce educational resources, such as cadaveric parts (Zhao, Kennedy, Yukawa, Pyman, & O'Leary, 2011). A meta-analysis on the effects of VR training for surgery training was first made in 2006 (Haque & Srinivasan, 2006). While the study was limited in scope (only assessed the impacts of VR simulators on task completion time), it concluded that VR simulators did lessen the time trainees take to complete a given surgical task.

Other studies have shown that simulation-based training of surgical skills can improve medical personnel performance in the operating room and diminish complication rates related to inexperience (Gallagher et al., 2005). For instance, VR laparoscopic simulators and robotic surgery have been extensively used in health practice (Gurusamy, Aggarwal, Palanivelu, & Davidson, 2008; Valdis, Chu, Schlachta, & Kiaii, 2016). Laparoscopic and robotic surgery have become a standard approach for many surgical specialties, as they reduce patient's surgical trauma, faster postoperative recovery, shorter hospital stays, and are associated with better cosmetic results. By using virtual reality simulator training, surgeons are expected to improve their proficiency and speed up their learning curve to master these procedures (Larsen et al., 2009).

Similarly, the use of VR training for eye surgery as well as for other uncomfortable procedures for both the patient and the examiner, such as transvaginal examinations and infant sedation, has been promoted extensively in the medical practice (Chao, Chalouhi, Bouhanna, Ville, & Dommergues, 2015; Zaveri et al., 2016). Finally, the use of VR training for some procedures, such as bone surgery and total hip arthroplasty, has shown to be effective addressing limited access to resources that are necessary for making practical training possible, namely real human bones, as well as decreasing surgical errors (such as the incorrect alignment of the hip).[2]

Safety and risk prevention are fields where VR training has also shown significant potential. As disasters and accidents are recurrent in all areas, training on safety and risk prevention is essential to mitigate their incidence and provide a rapid response and minimize casualties. VR training allows participants to emulate situations that may otherwise not be accessible with traditional learning methods. Immersive VR simulators have the potential to expose individuals to situations where high-level performance is critical but difficult to rehearse, such as mass disasters, evacuation drills, firefighting, and other hazardous or toxic conditions (Farra et al., 2018). Training emergency response personnel for catastrophes, for instance, is difficult due to the inability to replicate a given disaster environment comprehensively. In addition, there is an ethical concern about exposing trainees to the emotional and physical stresses encountered in real casualty situations (Andreatta et al., 2010). Available disaster drills often rely on mock patients, and they

[2] Although cadaver temporal bones remain the gold standard of simulated training for temporal bone surgery, their increasing scarcity worldwide has meant that additional training tools are currently being explored. In addition to a shortage of cadaver bones, the increasing workload of attending surgeons has meant that the time that can be devoted to teaching and education has decreased.

can be very costly. Also, available disaster drills do not provide opportunities for on-demand repetitive training. In such contexts, VR training could represent a more cost-effective and accessible alternative than large-scale real-life exercises.

2.2. Engineering, science, and technical education

Applications of VR training in engineering, science, and technical education have been most common in the fields of aviation, design, mechanics, industrial safety, and robotics (Buiu, & Gansari, 2014; Wei, Dongsheng, & Chun, 2013). In these fields, VR training provides students similar to real-life environments and access to state-of-the-art technology and equipment without the need to make significant capital investments in laboratories.

When teaching engineering, science, and technical education, laboratory sessions constitute an essential part of the training. They provide hands-on experiences that allow students to learn the necessary skills required to manage, configure, troubleshoot, repair equipment, specialized instruments, and machinery. Laboratories enable students to practice and acquire skills before performing tasks in real professional situations. Nonetheless, many technical programs fall short in providing practical experience to students, as the set-up and functioning of laboratories require important capital investments in equipment, as well as in maintenance and updates. Also, the necessary equipment needed to perform hands-on labs is not always available or accessible, especially in developing countries and in rural areas.

A proposed solution to address the difficulties to set-up laboratories is to substitute them with virtual laboratories. A virtual lab is an interactive simulation of a real lab. Virtual labs are essentially synthetic environments with attributes that include interactivity and real-time feedback (Lampi, 2013). The purpose of virtual labs is to develop student proficiency in the execution of practical skills. Virtual labs have traditionally been used in fields that require a skills proficiency that guarantees learners' safety before they can operate real equipment. Pilot training, military equipment training, and nuclear power plant training have a long-documented history in utilizing virtual labs (Lampi, 2013). Moreover, virtual labs offer additional advantages such as remote access for distance education, low cost, reliability, security, flexibility, and convenience to the student. The authenticity of the learning experience in a virtual lab depends on the extent to which

the simulation causes learners to engage in cognitive processes comparable to those provided by a real laboratory. VR has been used to improve the fidelity of virtual labs as 3D environments, allowing the possibility of recreating real environments and even developing psychomotor skills when using virtual labs (Stone, Watts, & Zhong, 2011).

2.3. VR training for general education

VR training has also been used to impart general education courses in several areas, such as STEM education, astronomy, anatomy, nurse education, and the arts (Kavanagh et al., 2017). Given its pedagogic potential and increasing market availability, it is crucial to examine the effectiveness of emerging VR technologies to deliver content to students in general education settings. Some advocates for using VR technologies for general instruction claim that immersive learning using VR brings motivational benefits that can lead to improved student learning (Makransky, et al., 2019). This is so because VR simulations can replace or amplify real-world learning environments by allowing students to interact and manipulate objects and parameters, thus promoting constructivist learning. VR simulations can also enable students to observe otherwise unobservable phenomena and provide students a higher sense of physical, environmental, and social presence. Students work harder when they are more interested in the material, either intrinsically (individual interest) or as elicited by the situation (situational interest) (Parong, & Mayer, 2018).

Nonetheless, creating educational applications for VR could be a laborious and costly endeavor, so it is crucial to investigate whether these applications are useful for learning or not (Allcoat, & von Mühlenen, 2018). Unnecessary features introduced by VR simulations may hinder learning compared with traditional methods or compared to less sophisticated multimedia channels, such as videos and well-designed slideshows (Parong et al., 2018). Some authors claim that VR simulations may not adhere to the coherence principle of multimedia, which states that people learn better when extraneous words, sounds, and pictures are excluded from rather than included in the student learning environment. This occurs because VR simulations, especially those that are fully immersive, often add material and features (visual effects, sounds, detailed environments) which could divert attention from the important material. In other words, VR simulations may include content that is not relevant to the instructional goal. Given that learners have a limited amount of cognitive processing capacity, if VR simulations entail unnecessary

detail, learners may not engage adequately with the essential materials that trigger cognitive processing and learning (Mayer, 2009; Mayer, 2014).

Finally, other authors claim that the usefulness of VR for general education might also depend on the type of subject of learning. Indeed, VR simulations may not necessarily be equally suitable for all subject areas. For example, it might be less beneficial for learning to play a musical instrument that requires tactile feedback, such as arts education. Still, it may be particularly helpful for teaching subjects where it is important to visualize the learning materials in 3D (e.g., biology or geometry) (Allcoat, & von Mühlenen, 2018). Other authors argue that technologies themselves, such as VR, do not directly cause learning but can afford specific tasks that themselves may result in learning (Dalgarno, & Lee, 2010).

III. Data and Methodology

A first step for conducting an informative meta-analysis is to gather relevant studies and accurately extract and report information from these primary studies (Uttl, White, & Gonzalez, 2017). Data were collected through a review of available studies assessing the impacts of VR training on learning and skills development. Studies included in the meta-analysis follow some predetermined criteria. First, they need to be published in a peer-reviewed journal or as a doctoral thesis, as a proxy for research quality. Second, to account for significant technological developments in access and quality of VR simulators (hardware and software), the sample only includes studies conducted within the last 15 years (2005-2020 period). Third, studies included in the meta-analysis assess the impact of VR training on student learning through value-added experiments or experimental evaluations that use randomized control trials (RTC).[3] Fourth, studies included in the sample assess skills development using objective and clearly measurable metrics, such as learning assessments or performance evaluations (pre- and post-test). The review of studies relied on web and academic databases, such as ACM Digital Library, IEEE Xplore, Web of Science, ERIC, and Scopus. Data obtained from these primary studies were compiled systematically, including the following information:

[3] RCT experiments are defined as those where individuals are allocated at random (by chance alone) to receive one or several interventions. One of these interventions is the standard of comparison or control. The control may be a standard practice or no intervention at all. Value-added studies quantify changes in desired outcomes (for instance, skills development of student learning) by quantifying these outcomes before and after individuals benefit from the intervention.

- Year of implementation;
- Field of study, using three main categories: (i) Health and safety, (ii) Virtual laboratories for engineering, science, and technical education; and (iii) General Education
- Type of VR training used: (i) immersive; (ii) non-immersive
- Beneficiary grade level: (i) Basic Education (K to 12); (ii) Technical-Vocational Education and Training (TVET), (iii) Higher Education, (iv) On-the-job training[4]
- Number of individuals who participated in the study
- Type of evaluation conducted: (i) RCT; (ii) Value-added
- Type of skills assessed: (i) Cognitive Skills, understood as the acquired knowledge to understand and retain complex ideas, adapt effectively to the environment, learn from experience, and reason; (ii) Technical Skills, understood as the expertise and ability needed to perform a specific job, including the mastery of the materials, tools, or technologies, and time on task [5]; and (iii) Socio-emotional skills, understood as the ability to navigate interpersonal and social situations effectively included leadership, teamwork, cooperation, self-control, self-confidence, self-efficacy, and grit
- A description of the instrument and scale used to assess student's skills
- Evaluation results (e.g. results of pre and posttest and their statistical significance)
- VR exposure time, in hours.

The meta-analysis assesses three main outcomes of VR training courses: learning performance (*L*), value-added (*VA*), and learning efficiency (*LE*). While not all papers report all three outcomes, papers included in the analysis report at least one of them. Learning performance is quantified as the average percentage gain in test-scores obtained after the training is completed (i.e. % difference in posttest scores), between students who receive the VR training, or treatment group (*T*) and students who receive traditional training, or control group (*C*). Learning value added is defined as the net gains accrued as a result of the training, measured by differences in posttest and pretest scores that assess similar competences. Most studies report this information for the treatment group only, but a few allow to assess differences in learning value added between

[4] TVET includes technical basic education, technical higher education, and vocational training programs. Higher Education includes academic undergraduate and graduate programs.
[5] Time on task refers to the time spent to successfully complete a task or procedure.

students in the treatment and control groups. Learning efficiency is measured as the % difference in inputs or time utilization (such as training time, time-on-task, materials used) between students exposed to VR training vs. students exposed to traditional training methods. These outcomes are defined as follows:

$$L = \frac{Posttest^T - Posttest^C}{Posstest^C} \quad (1)$$

$$VA^k = \frac{Postest^k - Pretest^k}{Pretest^k}, \text{ with } k = [T, C] \quad (2)$$

$$LE = \frac{Inputs^T - Inputs^C}{Inputs^C} \quad (3)$$

Since not all test-scores in (1) and (2) reported in the studies use a similar metric, a first step to assure comparability of outcomes across studies is to conduct a monotonic transformation to normalize all test-scores (r) between zero and one and into a comparable metric (s), as follows:

$$s^k = \frac{r^k - min}{max - min}, \text{ with } k = [T, C] \quad (4)$$

Where *min* and *max* are the minimum and maximum allowed values for the test-score of the original metric used.

Since each study (i) included in the meta-analysis display important differences in sample size (N), intervention exposure times in hours (E), and type of skills assessed (j) (cognitive, technical, and socio-emotional); we use weights in our estimations to take into account that larger sample size studies and higher intervention exposure are associated with less sampling error than studies with smaller size and less intervention exposure. As such, to assess average effects by skill type, we compute for each of the outcomes (O) assessed, a sample weighted average and a sample weighted variance, as follows:

$$\bar{O}_j = \sum_i \frac{N_{i,j} \times E_{i,j} \times O_{i,j}}{\sum_i N_{i,j} \sum_i E_{i,j}} \quad (5)$$

$$\sigma_j^2 = \sum_i \frac{(N_{i,j} \times E_{i,i}) \times (O_{i,j} - \bar{O}_{i,j})^2}{\sum_i N_{i,j} \sum_i E_{i,j}}, \text{ with } O = [L, VA, LE] \quad (6)$$

Since the main objective of the study is to identify the extent to which VR contributed to learning and skills development, the instruments used by the primary studies to assess learning and

skills are of prominent importance. The assessment of technical skills often relies on direct observation of the trainee's performance on a predetermined task or procedure. Expert practitioners conduct observations and evaluate student performance based on a series of pre-determined metrics. Observations tend to be anonymous to prevent the observer's bias. The protocols for the observation and the metrics to assess performance are often available and previously applied by the industry to certify professional skills, especially in the fields of medicine and engineering. The assessment of cognitive skills is generally measured using standardized tests, developed by professors and practitioners based on the curricula imparted in the training course. Finally, the assessment of socio-emotional skills is mainly conducted through students' self-reported perceptions of self-efficacy and attitudes towards learning. Tables A2, A3 and A4 in the Annex provide a detailed description of the instruments used by different experiments to assess skills proficiency.

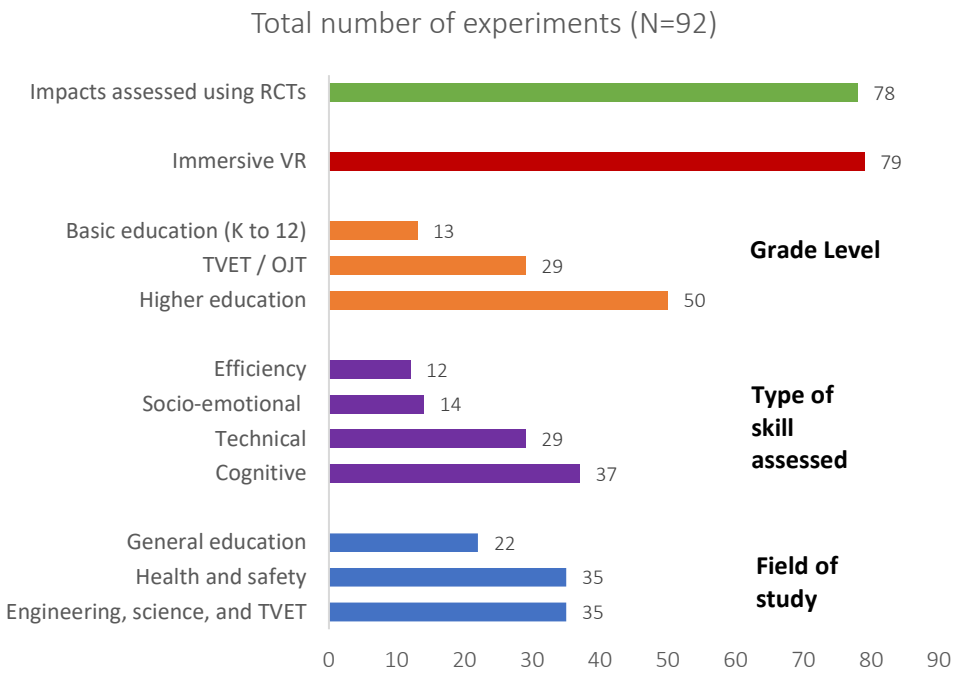
3.1. Descriptive statistics

After conducting a thorough review of the literature, a total of 31 primary studies met the criteria specified above. Most studies (29) were conducted in OECD countries, notably in the United States, the United Kingdom, and Canada (18 of 31). While many studies attempt to assess the effects of VR on learning and skills development, not all conduct a credible evaluation of the impacts and, among those that do, many do not report complete information on the impacts of the VR training and their statistical significance. Nonetheless, the studies that met the criteria (31) include 92 different experiments that assess the effects of VR training on students' skills development. Detailed information about each experiment is provided in Tables A1 to A12 in the Annex section.

Figures 1 and 2 present descriptive statistics of the 92 experiments included in the meta-analysis. A total of 78 experiments assess the impacts of VR training on learning outcomes using RCTs. Most experiments (79 of 92) assess immersive VR training. A total of 50 experiments were conducted in higher education settings, while 37 experiments studied the effects of VR training on cognitive skills and 29 studied the effect on technical skills. Only 13 experiments were conducted in basic education settings (k to 12). In terms of the educational field, the experiments were evenly distributed in health and safety (35) and virtual labs for engineering, science, and technical education (35). A total of 22 experiments focused on general education (Figure 1). While

categorizing experiments across education fields was straightforward for those experiments related to health and safety, there were some topics that overlapped between experiments in the fields of engineering, science, and technical education with those pertaining to general education. The determining factor to sort these studies in one of these two field, was the use or not of a virtual laboratory. If the training imparted aimed to emulate and/or substitute for a real laboratory, the experiment was included in the field of engineering, science, and technical education – independently of the subject (see Table A1).

Figure 1: Descriptive statistics of the primary experiments in the meta-analysis.

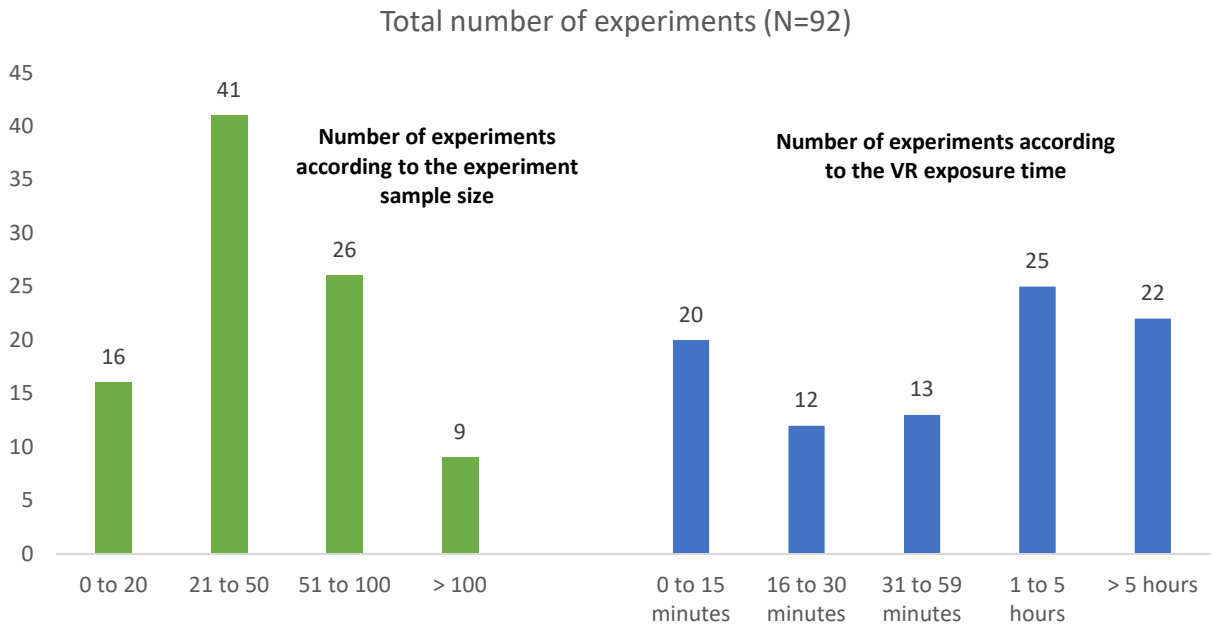


Source: Author's elaboration

Developing comparable and fair metrics constitutes a critical aspect to accurately assess the impacts of VR training on learning and skills development. As mentioned above, in order to provide a fair assessment of the available literature, the meta-analysis weights results based on two aspects, notably, intensity of the treatment (i.e. exposure time to VR training) and experiment sample size (Figure 2). These two variables display important variations in the primary studies included in the analysis. While half of the experiments expose students to more than one hour of VR training, 20 experiments report very short VR exposure (less than 15 minutes), while 22 report an exposure that surpasses 5 hours. Moreover, while most experiments included in the analysis are

medium size (benefiting between 21 and 100 students), some experiments have very limited sample sizes, of fewer than 20 beneficiaries (16 experiments in total), while others (9) include larger scale experiments that reach more than 100 beneficiaries.

Figure 2: Exposure time and sample size of the primary studies in the meta-analysis



Source: Author's elaboration

Figure 3 provides a list of all studies included in the meta-analysis, as well as information about their relative weight (i.e. sample size multiplied by exposure time, normalized to 100%), education field, and training topic. Results in the figure show studies in the field of engineering, science, and technical education tend to provide students with longer VR exposure and be conducted at a larger scale. As such, these studies are given a higher relative weight in the analysis.

Finally, studies included in the meta-analysis cover a diverse range of topics, from surgical education to welding. General education topics are also studied, such as from frog dissection and job interview training. In the field of health and safety, most studies included in the analysis pertained to surgical education (laparoscopic surgery, bone dissection, robotic cardiac surgery, robotic suturing, cataract surgery, and hip arthroplasty), although studies in other areas were also included, such as studies pertained to safety and risk prevention (2), one study on medical procedures in gynecology, and one study on nursing education.

Figure 3: Studies in the meta-analysis according to their relative weight (in %)

Note: studies in bold are in the field of engineering, science, and technical education and underlined studies are in the field of health and safely. Other studies are in the field of general education.

Most studies included in the field of health and safety focused on assessing the impact of VR training on students' technical skills. In the field of engineering, science, and technical education, all studies are focused on assessing the effectiveness of virtual labs (e.g. for welding, physics, chemistry) to develop students' practical and cognitive skills. Most studies pertaining

general education were in STEM related fields and focused on assessing the effects of VR training on students' cognitive skills vis a vis otherwise similar content imparted through more traditional mechanisms, such as slideshows and videos (see Table A5 in the annex).

IV. Main results

This section presents the main results of the meta-analysis. Table 1 presents the observed average effects of VR training on student learning performance, as proxied by the % difference in posttests between students exposed to VR training vs. students exposed to traditional training. The effects are calculated based on information of a total of 42 available experiments, and account for the experiments' sample size and VR exposure time. To foster comparability, we report the effects of VR training on students' technical skills per one hour of training and the effects of VR training on students' cognitive skills per quarter-hour of training. This reporting choice reflects that many experiments that aim to improve students' cognitive skills are of minimal duration (in some cases, they entail an exposure of fewer than five minutes), which would be difficult and misleading to extrapolate more extendedly.

Table 1. Average effects of VR training on student learning performance

	% Difference in posttests results (VR vs traditional training)		
	Technical skills (effect per hour of training)	**Cognitive skills** (effect per ¼ hour of training)	**Socio-emotional skills** (effect per training course)
Average impact	2.95	2.51	29.8
Standard Deviation	10.1	8.07	-
Number of experiments	15.0	24.0	3.0
Registering a positive effect	7.0	17.0	2.0
Registering negative effect	0.0	1.0	0.0
Registering no effect	19 8.0	6.0	1.0
Average VR exposure time (hours)	8.8	1.93	22.0
Average experiment size	57.3	89.6	80.0

Source: Author's elaboration.

A total of (15) 26 out of the 42 experiments included in the meta-analysis show a (neutral) positive effect of VR training on learning performance. Only one experiment indicates that VR training is associated with lower learning performance when compared to traditional training. Findings emanating from these experiments indicate that, on average VR training is indeed more effective than traditional training as a mechanism to develop technical, practical, and socio-emotional skills. Results are particularly promising in fields related to health and safety,

engineering, and technical education. Results reveal that for each additional hour (¼ hour) of training, students exposed to VR training score 3 percent higher in technical (cognitive) learning assessments, when compared to students exposed to the same curricular content delivered through traditional training methods. Results also indicate that students who complete a VR training course report, on average, 30 percent higher scores in socio-emotional skills assessments.

Table 2 presents the observed average effects of VR training on learning value added, as proxied by the % difference between posttests and pretests of students exposed to VR training. These results are indicative of the capacity of VR training to positively improve students' skills. The effects are calculated based on information of a total of 27 available experiments, and account for the experiments' sample size and VR exposure time. Results indicate that, on average, VR training contributes to gains in technical (practical) skills that average 17 (8) percent per hour (¼ hour) of training. Moreover, students who completed a VR training course, report on average, gains averaging 20.5 percent in self-reported socio-emotional skills.

Table 2. Average effects of VR training on student learning value added

	% Difference between posttests and pretests (VR only)		
	Technical skills (effect per hour of training)	**Cognitive skills** effect per ¼ hour of training)	**Socio-emotional skills** (effect per training course)
Average impact	17.3	8.0	20.5
Standard Deviation	20.9	31.3	26.1
Number of experiments	10.0	7.0	10.0
Registering a positive effect	9.0	7.0	7.0
Registering negative effect	0.0	0.0	0.0
Registering no effect	1.0	0.0	3.0
Average VR exposure time in hours	3.5	2.42	0.24
Average experiment size	23.4	89.0	39.4

Source: Author's elaboration.

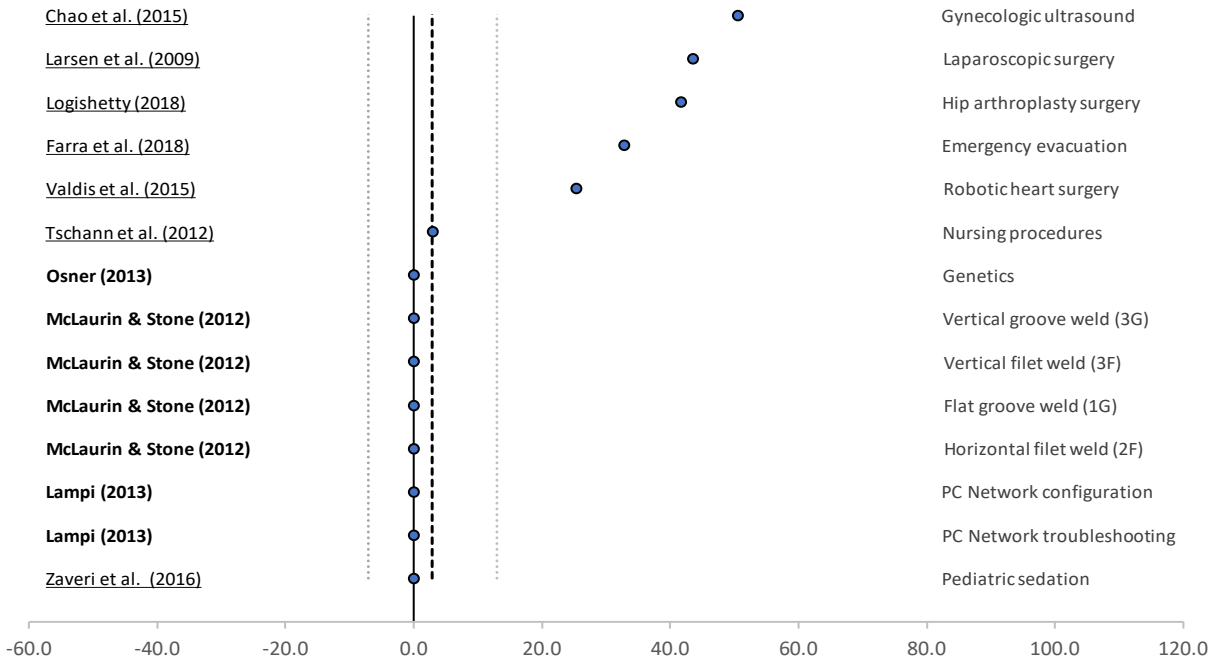
Average results hide important patterns that arise when assessing the effects of VR training by education type of skill and education field. Results in the next subsections are organized by the type of skill assessed by the available experiments (technical, cognitive, and socio-emotional) and, within each skill, if there are any observed patterns, there is brief discussion of the impacts of the training by education field.

4.1. Impacts of VR training on technical skills

Technical skills are often measured using ability tests, whereby students are required to perform specific task and are graded based on how their performance fares against predetermined

standards. The meta-analysis includes a total of 14 experiments that assess student learning of technical skills, proxied as the % difference observed in posttests per hour of training received between students who participated in VR training, or treatment group, and students who participated in traditional training, or control group. The results of these experiments are presented in Figure 4. Results in the chart provide information about the authors who conducted the experiment and the field and topics of the training.

Figure 4: % Difference in posttests between students exposed to VR vs non-VR training per hour of training [Technical Skills]



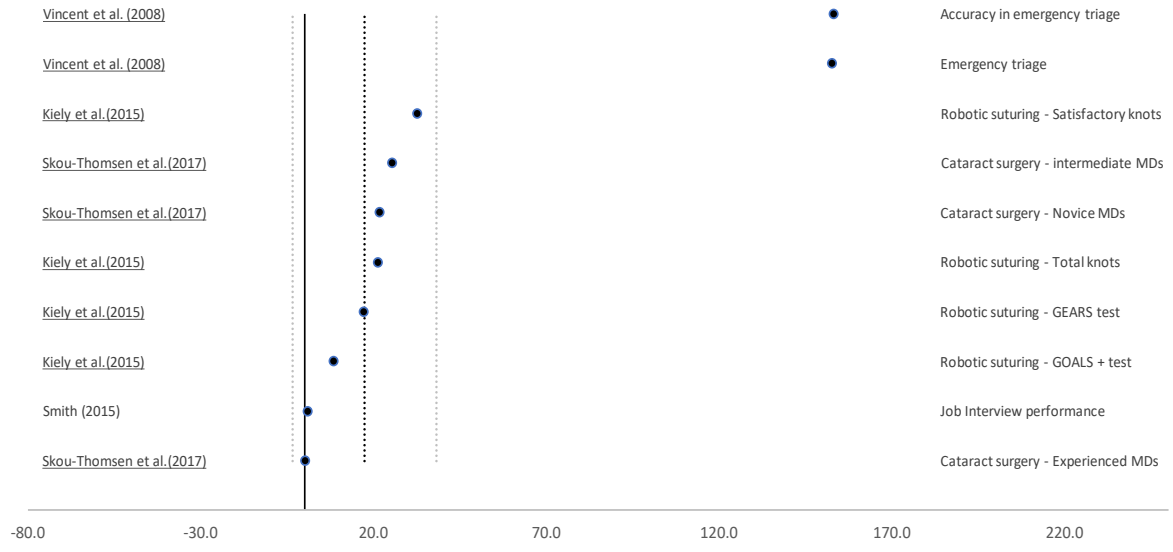
Note: studies in bold are in the field of engineering, science, and technical education. Underlined studies are in the field of health and safely. Other studies are in the field of general education. The bold vertical dotted line represents the average observed effects. Gray vertical dotted lines represent the standard deviation of the observed effects.

Not surprisingly, experiments that assess technical skills pertain mainly to the fields of health and safety (notably surgery performance) and engineering, science, and technical education (i.e. they relate to virtual labs for training students to perform technical tasks). Results indicate that for experiments in the field of health and safety, VR training is generally associated with posttest scores in trainees' ability assessments that are 20 to 60 percent higher per every extra hour of instruction than those observed in students exposed to traditional training. However, this is not the case for studies in engineering, science, and technical education where students who are exposed to VR laboratories fare just as well as students who access traditional laboratories.

At the aggregate level, the average difference in post-test scores between students in the treatment and control groups is 3 percent per additional hour of training, with a standard deviation of 10 percent. This average effect accounts for the fact that experiments in the field of health and safety, where impacts tend to be positive and large, are generally limited in terms of sample size and exposure time to VR training (see Table A1). The fact that students exposed to VR training in the fields of engineering, science, and technical education do not display higher posttest scores than students exposed to traditional training does not indicate that VR training in these fields is not effective. What the result suggests is that VR training is as effective as traditional training methods. As will be discussed below, when the effects of VR training on learning efficiency are assessed, this result is quite relevant because VR training for technical education could be more cost-effective that traditional training in cases where it is cheaper and safer to use simulators than traditional laboratories that are expensive to set up, maintain, and update.

The meta-analysis also includes 10 experiments, most of them in the field of health and safety, that assess the effects of VR training on student learning value added, as proxied by the % difference in post-test minus pre-tests for students who participate for VR training (Figure 5).

Figure 5: Value added (% difference between posttests and pretests) per hour of training for students in the Treatment group [Technical Skills]

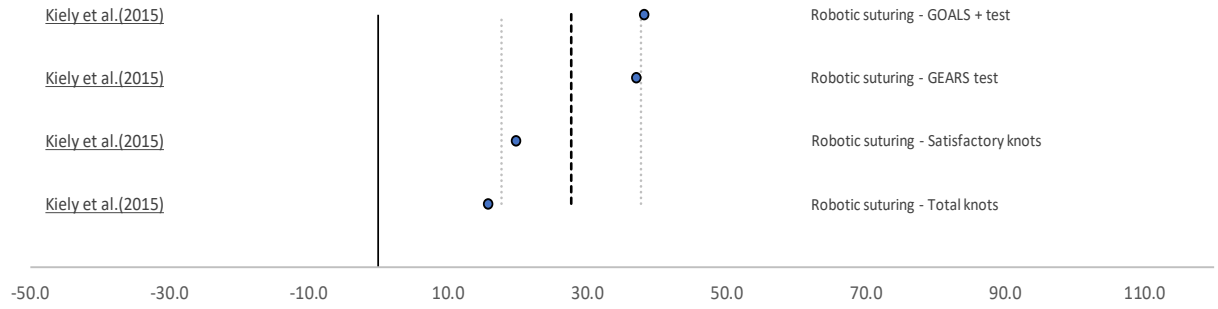


Note: studies in bold are in the field of engineering, science, and technical education. Underlined studies are in the field of health and safely. Other studies are in the field of general education. The bold vertical dotted line represents the average observed effects. Gray vertical dotted lines represent the standard deviation of the observed effects.

Results systematically confirm the effectiveness of VR training to develop technical skills in the field of health and safety. Available experiments indicate that one hour of VR training increases students net learning outcomes by 17 percent on average (with a standard deviation of 18 percent). Some results even indicate that VR training can be conducive to double student learning gains, especially in topics such as emergency response, where it is otherwise hard to provide students with access to real emergency situations. Results also show that the effects of VR training on value added can vary depending on the seniority of the trainees. For instance, results in Skou-Thomsen et al. (2017) indicate that VR training is conducive to higher learning gains for novice and intermediate surgeons. However, such training shows no statistically significant learning gains for more experienced surgeons.

Finally, the meta-analysis results include 4 experiments that assess differences learning value added between students exposed to VR training vs. students exposed to traditional training. In the field of robotics surgery, these experiments suggest that students exposed to VR training display 28 percent higher learning gains per hour of training on average than students exposed to traditional training (Figure 6).

Figure 6: % Difference in value added per hour of training (VR vs. Non-VR) [Technical Skills]



Note: studies in bold are in the field of engineering, science, and technical education. Underlined studies are in the field of health and safely. Other studies are in the field of general education. The bold vertical dotted line represents the average observed effects. Gray vertical dotted lines represent the standard deviation of the observed effects.

4.2. Impacts of VR training on cognitive skills

Cognitive skills are generally measured using standardized tests, whereby students are required to answer a set of questions based on the curricula imparted. The meta-analysis includes a total of 24 experiments that assess the impact of VR training on students' cognitive skills, proxied as the % difference observed in posttests per every ¼ hour of training received between students

who participated in VR training, or treatment group, and students who participated in traditional training, or control group. Contrary to VR training experiments aimed to develop technical skills, available experiments seeking to improve cognitive skills tend to have shorter exposure times to the technology (See Table A1 in the annex). As a result, as mentioned before, while we assess the impacts of VR training per hour of instruction when courses aim to develop technical skills, when developing cognitive skills, it seems more adequate to assess the effects per ¼ hour of instruction.

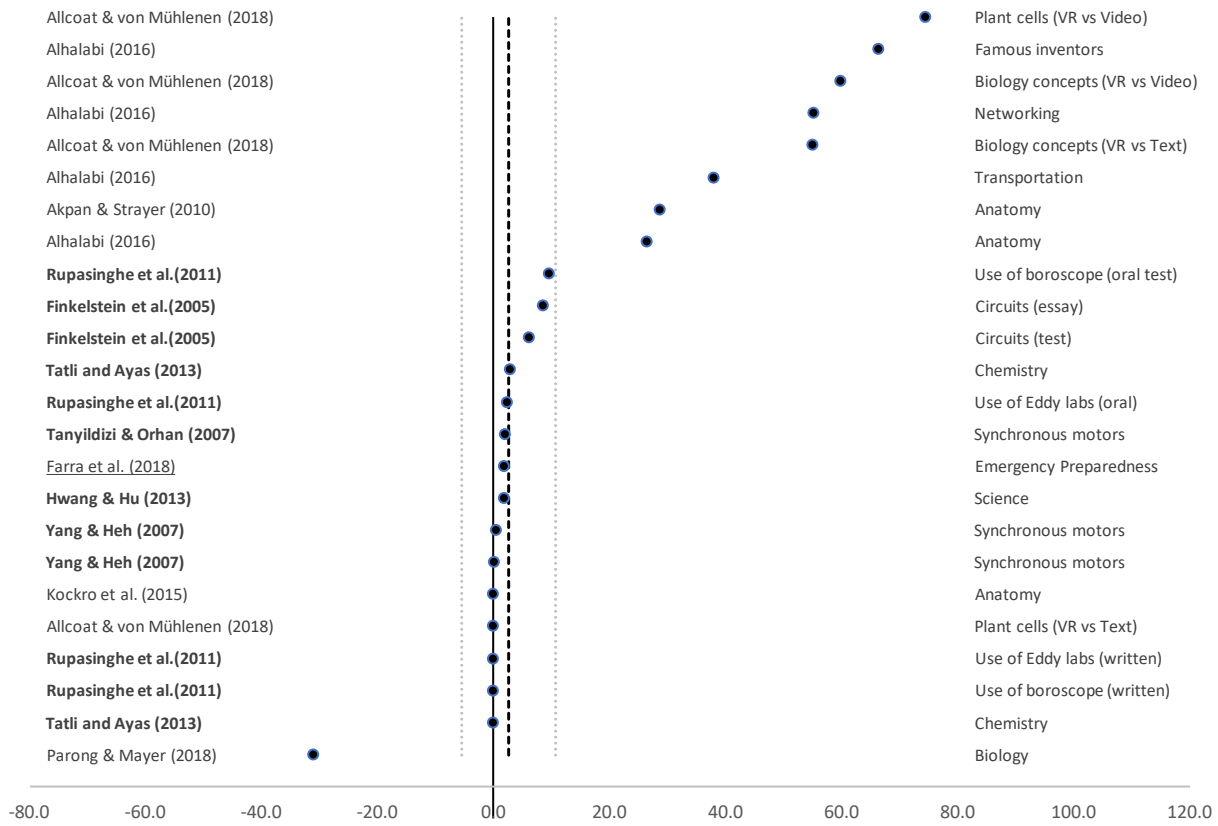
Once factoring out the relative weight of each experiment, results of available experiments reveal that the VR training is associated with a 3 percent higher learning than traditional training per every ¼ hour of instruction, with a standard deviation of 8 percent. Most experiments assessing cognitive skills pertain to the fields of general education (in topics related to STEM), as well as virtual laboratories. The results of these experiments are presented in Figure 7. While results from available experiments display some dispersion, most experiments (18 out of 24) indicate that VR training has positive impacts in student learning. In some experiments (8 out of 24), these impacts are quite high and show that students exposed to VR training have results in cognitive assessments that are 20 to 80 percent higher than those of students exposed to traditional training methods per every additional ¼ hour of instruction. Nonetheless, a total of 6 experiments show no significant effects of VR training on learning vis a vis traditional instruction. Only one study, in the topic of biology, finds that the effect of VR training on learning is negative (Parong & Mayer, 2018).

When assessing by field of study, the results presented in Figure 7 indicate that that students trained in virtual laboratories generally display higher cognitive learning (proxied by test scores) than students exposed to traditional laboratories. The intuition behind this result is that virtual labs allow for illimited repetition of experiments, are self-paced, and generally provide direct feedback to students. Such features are particularly useful for student learning in topics that require understanding of abstract concepts, such as physics (Yang, & Heh, 2007).

Results do not reveal a clear pattern when it comes to general education. Most related experiments aim to assess the impacts VR training has on learning compared to other more traditional instruction methods such as lecture and/or classes that use other type multimedia aid such as a video, a textbook, or a slideshow (Table A5 in the annex provides more details). Some studies indicate that students who receive VR training perform better in cognitive assessments than

students exposed to a traditional lecture or videos (Alhalabi, 2016; Allcoat, & von Mühlenen, 2018).

Figure 7: % Difference in posttests between students exposed to VR vs non-VR training [Cognitive Skills]



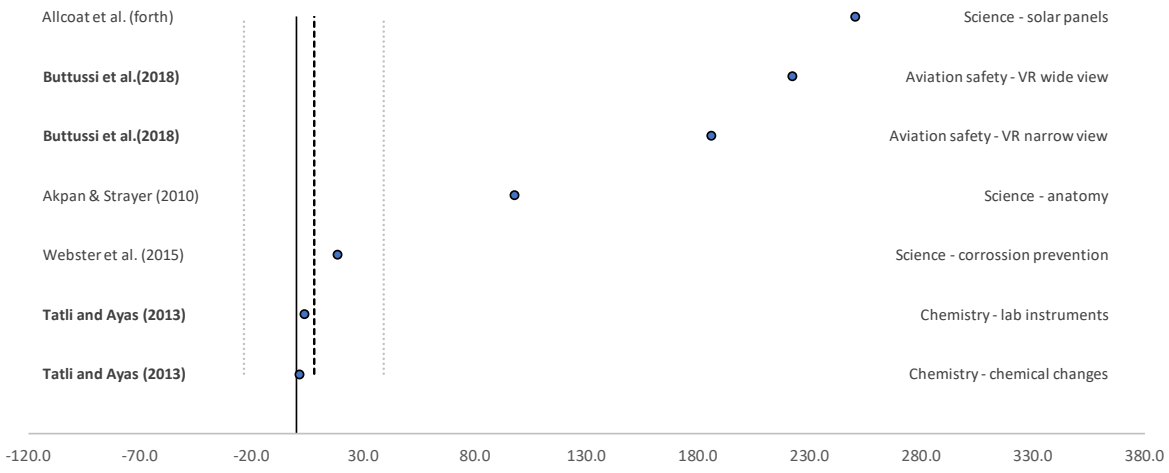
Note: studies in bold are in the field of engineering, science, and technical education. Underlined studies are in the field of health and safely. Other studies are in the field of general education. The bold vertical dotted line represents the average observed effects. Gray vertical dotted lines represent the standard deviation of the observed effects.

Other studies (Allcoat, & von Mühlenen, 2018; Hwang, & Hu, 2013) indicate that students who receive VR training perform just as well as students that use textbooks to complement traditional lectures. Allcoat and von Mühlenen (2018) and Hwang and Hu (2013) find that students exposed to VR training memorize better the parts of a plant cell and display better skills to calculate area and volume of 3D figures than students exposed to similar content but whose lessons were imparted using more traditional learning methods, such as videos and lectures. Parong & Mayer (2018) find that students exposed to slideshows perform better than students exposed to VR training on subjects such as biology. Finally, Kockro et al., 2015 found no significant difference

in students' knowledge of the human ventricular system's anatomy after providing them with a training course using VR content versus a traditional slideshow.

The meta-analysis also includes 7 experiments that assess the effects of VR training on learning value added, as proxied by the % different in post-test minus pre-tests for students who participate for VR training (Figure 8). Results in all available experiments are indicative that VR training is conducive to positive learning gains for the development of cognitive skills. Such gains in some experiments can be as high as 2.5 times the baseline cognitive knowledge, as proxied by standardized tests. Nonetheless, experiments with such acute increases in learning gains are often very small in size and limited in terms in VR exposure (i.e. VR training of less than 5 minutes) (Allcoat, 2021; Buttussi, & Chittaro, 2018). Once accounting for experiment size and exposure time, results indicate that VR training contributes to student learning gains averaging 8 percent per ¼ hour of training, with a rather large standard deviation of 31 percent.

Figure 8: Learning value added (% difference between posttests and pretests) per ¼ hour of training for students in the Treatment group [Cognitive Skills]

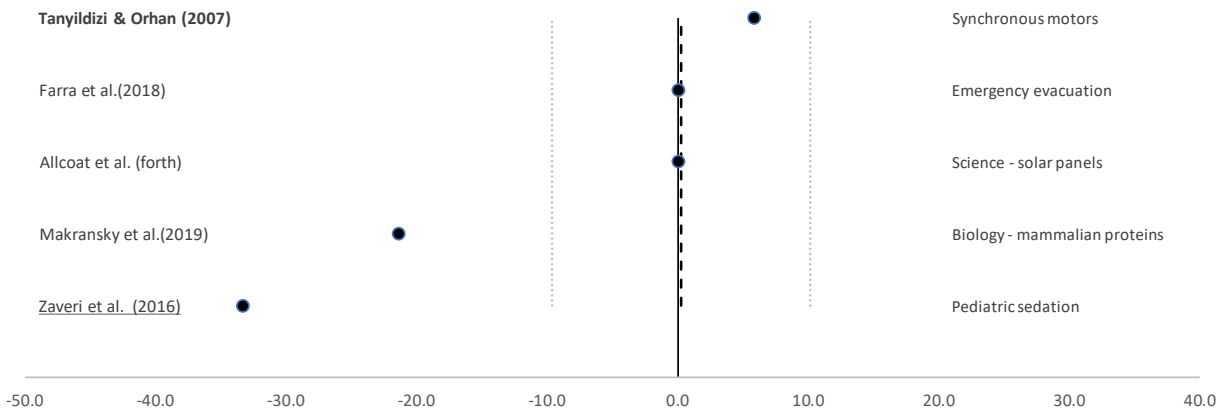


Note: studies in bold are in the field of engineering, science, and technical education. Underlined studies are in the field of health and safely. Other studies are in the field of general education. The bold vertical dotted line represents the average observed effects. Gray vertical dotted lines represent the standard deviation of the observed effects.

Finally, results in the meta-analysis include 5 experiments that assess differences in learning value added between students exposed to VR training vs. students exposed to traditional training. Results are mixed. Some experiments indicate that students exposed to VR training experiment lower learning gains than students exposed to traditional training (Zaveri et al., 2016; Makrasky et al., 2019). Other experiments indicate that VR training contributes to similar or higher

learning gains than traditional training (Allcoat, 2021; Farra et al., 2018; Tanyildizi & Orhan, 2007). Nonetheless, when accounting for experiment size and exposure time, the average observed differences in learning gains per ¼ hour of training between students exposed to VR vis a vis those exposed to traditional learning is close to zero (Figure 9), indicating that VR training is, on average, as effective a mechanism to enhance student's cognitive skills when compared to traditional training. Nonetheless, due to the limited number of experiments that assess learning value added across VR and non-VR recipients, results need to be used with care and may not allow to adequately generalize.

Figure 9: % difference in value added per ¼ hour of training (VR vs. Non-VR) [Cognitive Skills]

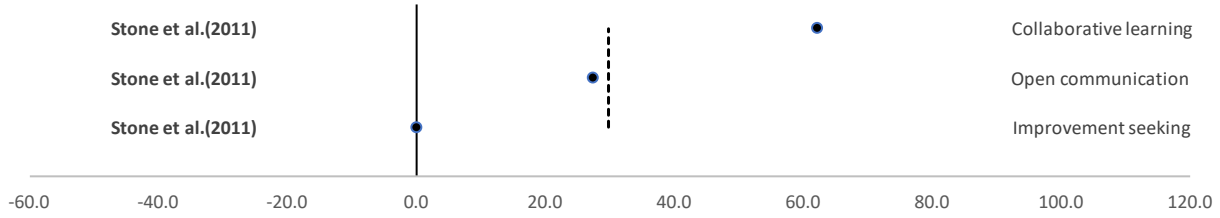


Note: studies in bold are in the field of engineering, science, and technical education. Underlined studies are in the field of health and safely. Other studies are in the field of general education. The bold vertical dotted line represents the average observed effects. Gray vertical dotted lines represent the standard deviation of the observed effects.

4.3. Impacts of VR training on socio-emotional skills

The assessment of socio-emotional skills is often conducted through students' self-reported perceptions of self-efficacy and attitudes towards learning. Such perceptions are often quantified using a Likert scale type of questionnaire that students complete before and/or after they complete their training (see Tables A2 to A4 in the annex). Such perceptions are often indicative of the main channels that explain why VR training can be more effective and more conducive to learning than traditional training. The meta-analysis includes a total of 13 experiments that assess the effects of VR training on the effects of students' socio-emotional skills.

Figure 10: % Difference in posttests per course (VR vs non-VR training) [Socio-emotional Skills]

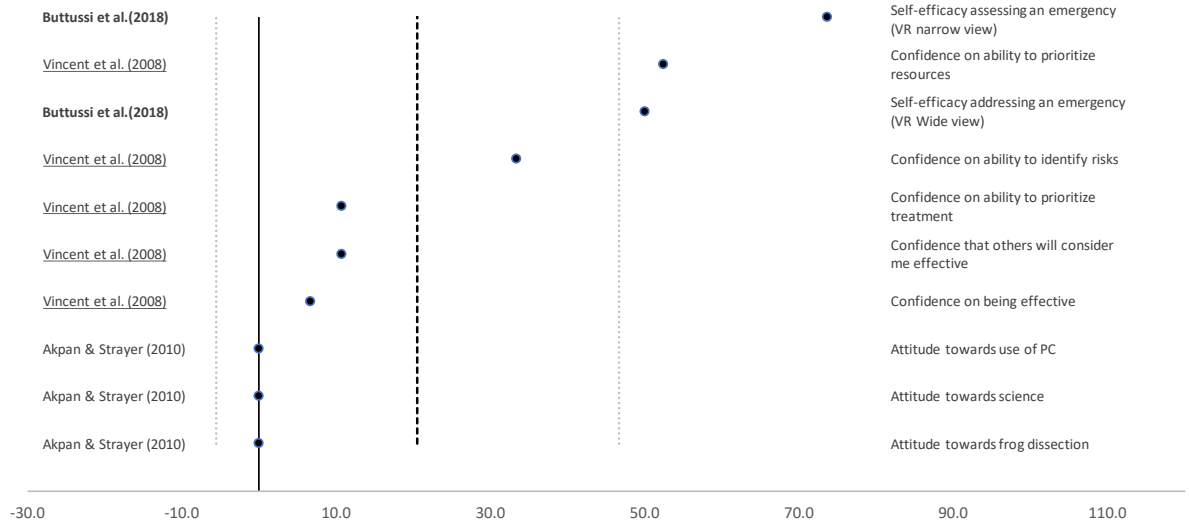


Note: studies in bold are in the field of engineering, science, and technical education. Underlined studies are in the field of health and safely. Other studies are in the field of general education. The bold vertical dotted line represents the average observed effects. Gray vertical dotted lines represent the standard deviation of the observed effects.

Results from Stone et al. (2011) indicate that students who receive welding training in virtual labs report higher self-reported peer communication and collaborative learning in post-test than students who receive traditional training. This may occur because virtual laboratories can allow for digital peer interactions, which can be repeated multiple times, and allow for real-time / automated feedback to students. However, the authors do not find significant differences in student self-reported improvement seeking of students who received their training in a virtual lab (Figure 10). The remaining 10 experiments assess the effects of VR training on student learning self-reported skills value added, as proxied by the % difference in post-test minus pre-tests for students who participate for VR training (Figure 11). Available results indicate that students who complete VR training tend to report 20% higher levels of confidence and self-efficacy towards learning after they complete their courses. For instance, results from Vincent et al. (2008) indicate that after being exposed to VR training, students report higher levels of confidence in several dimensions (e.g. being effective, making decisions, and using resources) when addressing a health emergency.

Interestingly, results for Buttussi et al. (2018) indicate that such self-reported increases in confidence can vary using different types of VR technology. In particular, the authors find that students exposed to narrow view VR technology (which allows them to focus more on features of the simulation) tend to report higher increases in self-reported self-efficacy addressing a plane emergency than students exposed to wide VR technology (which may expose students to more visual distractions).

Figure 11: Learning value added (% difference between posttests and pretests) per completed VR training course [Socio-emotional Skills]



Note: studies in bold are in the field of engineering, science, and technical education. Underlined studies are in the field of health and safely. Other studies are in the field of general education. The bold vertical dotted line represents the average observed effects. Gray vertical dotted lines represent the standard deviation of the observed effects.

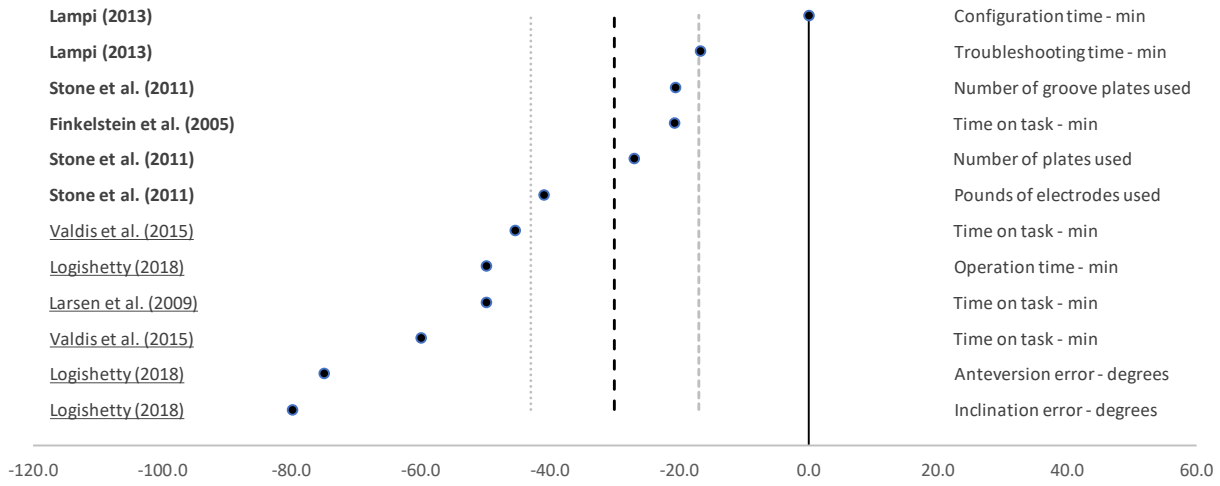
Finally, results from Akpan and Strayer (2010), which use VR training to simulate a frog dissection, do not find a significant effect of the use of VR training in students change in attitude towards learning anatomy and/or using PC assisted instruction.

4.4. Impacts of VR training on learning efficiency

In the context of this study, learning efficiency is defined as any savings in the form of inputs, time, or performance errors that VR training could contribute to. One of the promises of using simulators vis-a-vis traditional instruction is their potential to save training costs and minimize the risks and errors faced when novice students intend to master some skills they will use in real life. Ideally, VR training should contribute to more efficient use of inputs, more expedited completion of tasks (or time-in-task), and fewer performance errors. The meta-analysis includes a total of 12 experiments that study the effects of VR training on the utilization of inputs (such as materials and time to complete a task) and performance errors. A total of 11 out of 12 experiments included in the meta-analysis find that VR training is associated with higher learning efficiency levels when compared to traditional training. In fact, experiment results indicate that, on average, students who are exposed to VR training are, on average, up to 30 percent more efficient (using inputs, time, and/or avoiding performance errors) than students exposed to traditional training per additional hour of instruction. Results from available experiments indicate

that VR training can help welding students to be more efficient using materials, such as plates and electrodes (Stone et al. 2011) and can expedite the time students take to perform surgical procedures such as laparoscopy surgery (Larsen et al. 2009), robotic heart surgery (Valdis et al., 2015) and hip arthroplasty (Logishetty, Rudran, & Cobb, 2018). Results from Logishetty, Rudran, & Cobb (2018) also indicate that students who are exposed to VR training are less likely to make mistakes when performing real surgery procedures (see Table A12 in the annex). Of course, due to the limited number of experiments used to draw these conclusions, these results need to be interpreted with care and should not be generalized. Nonetheless, these findings are indicative of the potential of VR simulators to be not only an effective, but also an efficient learning mechanism, especially in the fields of health and safety and technical education.

Figure 12: % decrease in inputs / performance errors per 1 hour of VR training compared to traditional training [Learning Efficiency]



Note: studies in bold are in the field of engineering, science, and technical education. Underlined studies are in the field of health and safely. Other studies are in the field of general education. The bold vertical dotted line represents the average observed effects. Gray vertical dotted lines represent the standard deviation of the observed effects.

V. Conclusions

The development of students' skills remains a challenge for education systems worldwide. To address this challenge, educators are beginning to explore the possibility of using information technology to create learning experiences that would otherwise not be accessible to students. Simulations that rely on VR technology can provide students access to learning environments that

would otherwise be very difficult, expensive, or impossible to access. VR simulations can provide students practical training opportunities without pressure, danger, and allowing repeated practice. Such opportunities have the potential of accelerating students' learning curve in a simulated environment, reproducing real-life conditions and situations without time or space limitations, and with much fewer risks.

This study constitutes an attempt to assess the effects of VR instruction holistically, as a mechanism to develop students' skills. Given its pedagogic potential and its increasing market availability, it is crucial to examine the effectiveness of emerging VR technologies for pedagogical instruction. Creating educational applications for VR could be a laborious and costly endeavor, so it is essential to investigate whether these applications are useful for learning or not and, to the extent possible, to assess their cost-effectiveness. Results capitalize from a thorough review of 31 primary studies and over 90 experiments that intend to assess the effects of VR instruction on student learning. Our findings reveal that VR instruction is, on average, more effective than traditional training as a mechanism to develop students' skills. Results indicate that for each additional hour (¼ hour) of training, students exposed to VR training score 3 percent higher in technical (cognitive) learning assessments, when compared to students exposed to the same curricular content delivered through traditional training methods. Results also indicate that students exposed to VR instruction report, on average, 30 percent higher scores in socio-emotional skills assessments after completing their training.

Results are particularly promising in fields related to health and safety, engineering, and technical education. Results from available experiments confirm systematically that VR instruction yields positive results as a mechanism to train surgeons and medical personnel. It offers trainees the opportunity to practice medical procedures safely and at a comparatively low cost. Available experiments confirm VR simulators' effectiveness to improve surgeons' proficiency to perform procedures such as laparoscopic surgery, robotic surgery, eye surgery, transvaginal examinations, infant sedation, and bone surgery, to name a few. Some results even indicate that VR training can be conducive to much higher student learning gains, especially in topics such as emergency response, where it is otherwise hard to provide students access to real emergencies.

VR training can provide students similar-to real-life laboratories and equipment without making significant capital investments. Available experiments show that virtual laboratories can

be as effective as real laboratories to develop students' skills, but they can be a more efficient, safe, and cost-effective mechanism of instruction. Our findings indicate that students exposed to VR training are up to 30 percent more efficient using inputs, time, and/or avoiding performance errors than students exposed to traditional training, per additional hour of instruction. The intuition behind this result is that virtual labs allow for illimited repetition of experiments, are self-paced, and generally provide direct feedback to students.

Nonetheless, results do not reveal a clear pattern when it comes to the use of VR instruction for general education. Some studies indicate that students who receive VR training perform better in cognitive assessments than students exposed to a traditional lecture or videos. Other studies indicate that students exposed to other less expensive multimedia platforms, such as slideshows and videos, learn more than students exposed to VR training. As such, VR instruction may not be adequate as a mechanism for instruction in all educational fields. Indeed, VR training may provide students too much information, which may deviate their attention from the aspects of the curricula that matter most.

VR training can also help students develop their socio-emotional skills. Simulations can develop and promote collaborative features that enable peer interaction, active learning, and performance feedback. In turn, these features can promote student motivation and presence, which are two channels that can positively influence student learning. Available experiments show that students exposed to VR training report higher peer communication and collaborative learning in standardized assessments, than students who receive traditional training. Available experiments also show that students who complete VR training report higher confidence and self-efficacy towards learning after completing their courses.

It will be essential to continue to assess the cost-effectiveness of VR training, which is something beyond the scope of this study. While VR training's cost-effectiveness is likely to vary depending on many parameters such as course duration, cost of the actual equipment that VR intends to simulate, educational field, risks of making mistakes in a non-simulated environment, and type of technology used, it is not always assured. Indeed, this type of instruction could be cost-effective only if it provides savings or reduces potentially expensive risks compared to other alternative multimedia or traditional laboratories. Imparting VR courses entails software development and equipment, maintenance, support, and updates, which require sustained

investments. To date, not many studies assessing the effects of VR training have focused on conducting a cost-benefit or cost-effectiveness analysis of VR instruction compared to traditional training methods. Having more such information will be crucial to assess the scalability potential of VR training across education systems.

Finally, this study's results primarily draw their conclusions based on experiments in developed countries. As such, these results may not necessarily hold in all educational settings because several factors necessary for VR training to succeed (e.g., connectivity, availability of equipment and IT support, students' and teachers' dominium of necessary digital skills, among others) are not necessarily assured in many education institutions in developing countries.

In summary, this study finds that VR training tends to be an effective mechanism of instruction to develop students' technical, practical, and socio-emotional skills. Nonetheless, these results cannot be generalized, and it is important to continue to assess the pros and cons of using VR for pedagogical instruction for different subjects as well as its cost-effectiveness and scalability.

References

Akpan, J., & Strayer, J. (2010). Which comes first the use of computer simulation of frog dissection or conventional dissection as academic exercise?. *Journal of Computers in Mathematics and Science Teaching*, *29*(2), 113-138.

Alhalabi, W. (2016). Virtual reality systems enhance students' achievements in engineering education. *Behaviour & Information Technology*, *35*(11), 919-925.

Allcoat, D. (2021). Effects of Virtual Reality on solar-power panel efficiency learning. Department of Psychology. University of Warwick,

Allcoat, D., & von Mühlenen, A. (2018). Learning in virtual reality: Effects on performance, emotion and engagement. *Research in Learning Technology*, *26*.

Andreatta, P. B., Maslowski, E., Petty, S., Shim, W., Marsh, M., Hall, T., ... & Frankel, J. (2010). Virtual reality triage training provides a viable solution for disaster-preparedness. *Academic emergency medicine*, *17*(8), 870-876.

Apostolellis, P., Bowman, D. A., & Chmiel, M. (2018). Supporting social engagement for young audiences with serious games and virtual environments in museums. In *Museum Experience Design* (pp. 19-43). Springer, Cham.

Barbe, W. B., Milone, M. N., & Swassing, R. H. (1988). *Teaching through modality strengths: Concepts and practices*. Zaner-Bloser.

Buiu, C., & Gânsari, M. (2014, April). Designing robotic avatars in Second Life-A tool to complement robotics education. In *2014 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1016-1018). IEEE.

Buttussi, F., & Chittaro, L. (2018). Effects of different types of virtual reality display on presence and learning in a safety training scenario. *IEEE transactions on visualization and computer graphics*, *24*(2), 1063-1076.

Chao, C., Chalouhi, G. E., Bouhanna, P., Ville, Y., & Dommergues, M. (2015). Randomized clinical trial of virtual reality simulation training for transvaginal gynecologic ultrasound skills. *Journal of Ultrasound in Medicine*, *34*(9), 1663-1667.

Cheung, S. K., Fong, J., Fong, W., & Wang, F. L. (Eds.). (2013). *Hybrid Learning and Continuing Education: 6th International conference, ICHL 2013, Toronto, ON, Canada, August 12-14, 2013, Proceedings* (Vol. 8038). Springer.

Dalgarno, B., & Lee, M. J. (2010). What are the learning affordances of 3-D virtual environments?. *British Journal of Educational Technology*, *41*(1), 10-32.

Escueta, M., Quan, V., Nickow, A. J., & Oreopoulos, P. (2017). Education technology: An evidence-based review.

Farra, S., Hodgson, E., Miller, E. T., Timm, N., Brady, W., Gneuhs, M., ... & Bottomley, M. (2019). Effects of virtual reality simulation on worker emergency evacuation of neonates. *Disaster medicine and public health preparedness*, *13*(2), 301.

Ferracani, A., Pezzatini, D., & Del Bimbo, A. (2014, November). A natural and immersive virtual interface for the surgical safety checklist training. In *Proceedings of the 2014 ACM international workshop on serious games* (pp. 27-32).

Finkelstein, N. D., Adams, W. K., Keller, C. J., Kohl, P. B., Perkins, K. K., Podolefsky, N. S., & LeMaster, R. (2005). When learning about the real world is better done virtually: A study of substituting computer simulations for laboratory equipment. *Physical review special topics-physics education research*, *1*(1), 010103.

Gallagher, A. G., Ritter, E. M., Champion, H., Higgins, G., Fried, M. P., Moses, G., ... & Satava, R. M. (2005). Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Annals of surgery*, *241*(2), 364.

Gigante, M. A. (1993). Virtual reality: definitions, history and applications. In *Virtual reality systems* (pp. 3-14). Academic Press.

Gurusamy, K., Aggarwal, R., Palanivelu, L., & Davidson, B. R. (2008). Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery. *British Journal of Surgery*, *95*(9), 1088-1097.

Haque, S., & Srinivasan, S. (2006). A meta-analysis of the training effectiveness of virtual reality surgical simulators. *IEEE Transactions on Information Technology in Biomedicine*, *10*(1), 51-58.

Huang, H. M., Rauch, U., & Liaw, S. S. (2010). Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach. *Computers & Education*, *55*(3), 1171-1182.

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

Kavanagh, S., Luxton-Reilly, A., Wuensche, B., & Plimmer, B. (2017). A systematic review of Virtual Reality in education. *Themes in Science and Technology Education*, *10*(2), 85-119.

Kiely, D. J., Gotlieb, W. H., Lau, S., Zeng, X., Samouelian, V., Ramanakumar, A. V., ... & Press, J. Z. (2015). Virtual reality robotic surgery simulation curriculum to teach robotic suturing: a randomized controlled trial. *Journal of robotic surgery*, *9*(3), 179-186.

Kockro, R. A., Amaxopoulou, C., Killeen, T., Wagner, W., Reisch, R., Schwandt, E., ... & Stadie, A. T. (2015). Stereoscopic neuroanatomy lectures using a three-dimensional virtual reality environment. *Annals of Anatomy-Anatomischer Anzeiger*, *201*, 91-98.

Lampi, E. (2013). *The effectiveness of using virtual laboratories to teach computer networking skills in Zambia* (Doctoral dissertation, Virginia Tech).

Larsen, C. R., Soerensen, J. L., Grantcharov, T. P., Dalsgaard, T., Schouenborg, L., Ottosen, C., & Ottesen, B. S. (2009). Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. *Bmj*, *338*.

Logishetty, K., Rudran, B., & Cobb, J. P. (2019). Virtual reality training improves trainee performance in total hip arthroplasty: a randomized controlled trial. *The bone & joint journal*, *101*(12), 1585-1592.

Lorenzo, G., Pomares, J., & Lledó, A. (2013). Inclusion of immersive virtual learning environments and visual control systems to support the learning of students with Asperger syndrome. *Computers & Education*, *62*, 88-101.

Madathil, K. C., Frady, K., Hartley, R., Bertrand, J., Alfred, M., & Gramopadhye, A. (2017). An empirical study investigating the effectiveness of integrating virtual reality-based case studies into an online asynchronous learning environment. *Computers in Education Journal*, *8*(3), 1-10.

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, *60*, 225-236.

Mayer, R. (2009). Multimedia Learning. *Cambridge University Press*. doi: 10.1017/cbo9780511811678

Mayer, R. (2014). Cognitive Theory of Multimedia Learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (Cambridge Handbooks in Psychology, pp. 43-71). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139547369.005

Mikropoulos, T. A., & Natsis, A. (2011). Educational virtual environments: A ten-year review of empirical research (1999–2009). *Computers & Education*, *56*(3), 769-780.

McLaurin, E. J., & Stone, R. T. (2012, September). Comparison of virtual reality training vs. integrated training in the development of physical skills. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 2532-2536). Sage CA: Los Angeles, CA: SAGE Publications.

Oser, R. R. (2013). *Effectiveness of virtual laboratories in terms of achievement, attitudes, and learning environment among high school science students* (Doctoral dissertation, Curtin University).

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

Rupasinghe, T. D., Kurz, M. E., Washburn, C., & Gramopadhye, A. K. (2011). Virtual reality training integrated curriculum: An aircraft maintenance technology (AMT) education perspective. *International Journal of Engineering Education*, *27*(4), 778.

Sharma, S., Agada, R., & Ruffin, J. (2013, April). Virtual reality classroom as an constructivist approach. In *2013 Proceedings of IEEE Southeastcon* (pp. 1-5). IEEE.

Skou -Thomsen, A., Bach-Holm, D., Kjærbo, H., Højgaard-Olsen, K., Subhi, Y., Saleh, G. M., ... & Konge, L. (2017). Operating room performance improves after proficiency-based virtual reality cataract surgery training. *Ophthalmology*, *124*(4), 524-531.

Smith, M. J., Ginger, E. J., Wright, K., Wright, M. A., Taylor, J. L., Humm, L. B., ... & Fleming, M. F. (2014). Virtual reality job interview training in adults with autism spectrum disorder. *Journal of autism and developmental disorders*, *44*(10), 2450-2463.

Stone, R. T., Watts, K. P., & Zhong, P. (2011). Virtual reality integrated welder training. *Welding Journal*, *90*(7), 136s.

Tanyildizi, E., & Orhan, A. (2007). A virtual electric machine laboratory for synchronous machine application. *Computer Applications in Engineering Education*, *17*(2), 187-195.

Tatli, Z., & Ayas, A. (2013). Effect of a virtual chemistry laboratory on students' achievement. *Journal of Educational Technology & Society*, *16*(1), 159-170.

Tschannen, D., Aebersold, M., Mclaughlin, E., Bowen, J., & Fairchild, J. (2012). Use of virtual simulations for improving knowledge transfer among baccalaureate nursing students. *Journal of Nursing Education and Practice*, *2*(3), 15.

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22-42.

Valdis, M., Chu, M. W., Schlachta, C., & Kiaii, B. (2016). Evaluation of robotic cardiac surgery simulation training: a randomized controlled trial. *The Journal of thoracic and cardiovascular surgery*, *151*(6), 1498-1505.

Vincent, D. S., Sherstyuk, A., Burgess, L., & Connolly, K. K. (2008). Teaching mass casualty triage skills using immersive three-dimensional virtual reality. *Academic Emergency Medicine*, *15*(11), 1160-1165.

Webster, R. (2016). Declarative knowledge acquisition in immersive virtual learning environments. *Interactive Learning Environments*, *24*(6), 1319-1333.

Wei, W., Dongsheng, L., & Chun, L. (2013, September). Fixed-wing aircraft interactive flight simulation and training system based on XNA. In *2013 International Conference on Virtual Reality and Visualization* (pp. 191-198). IEEE.

Yang, K. Y., & Heh, J. S. (2007). The impact of internet virtual physics laboratory instruction on the achievement in physics, science process skills and computer attitudes of 10th-grade students. *Journal of Science Education and Technology*, *16*(5), 451-461.

Zaveri, P. P., Davis, A. B., O'Connell, K. J., Willner, E., Schinasi, D. A. A., & Ottolini, M. (2016). Virtual reality for pediatric sedation: a randomized controlled trial using simulation. *Cureus*, *8*(2).

Zhao, Y. C., Kennedy, G., Yukawa, K., Pyman, B., & O'Leary, S. (2011). Can virtual reality simulator be used as a training aid to improve cadaver temporal bone dissection? Results of a randomized blinded control trial. *The Laryngoscope*, *121*(4), 831-837.

ANNEX

Table A1: Studies included in the meta-analysis

STUDY	Type of skill assessed			Grade Level						
	C	T	SE	K-12	H.E.	TVET /OJT	Immersive VR	RCT	Sample size	VR Exposure in hours
	Health and safety									
Farra et al.(2018)	Yes	Yes	Yes*	No	No	Yes	Yes	Yes	93	0.66
Logishetty (2018)	No	Yes	No	No	Yes	No	Yes	Yes	24	2.00
Skou-Thomsen et al.(2017)	No	Yes	No	No	Yes	No	Yes	No	18	1.50
Zaveri et al.(2016)	Yes	Yes	No	No	Yes	No	Yes	Yes	14	0.50
Chao et al.(2015)	No	Yes	No	No	Yes	No	Yes	Yes	34	0.66
Kiely et al.(2015)	No	Yes	No	No	Yes	No	Yes	Yes	27	5.00
Valdis et al.(2015)	No	Yes	No	No	Yes	No	Yes	Yes	20	9.30
Tschannen et al.(2012)	No	Yes	No	No	Yes	No	Yes	Yes	115	3.00
Zhao et al.(2011)	No	Yes	No	No	Yes	No	Yes	Yes	20	2.00
Larsen et al.(2009)	No	Yes	No	No	Yes	No	Yes	Yes	21	1.00
Vincent et al. (2008)	No	Yes	Yes	No	Yes	No	Yes	No	20	0.25
	Virtual labs for engineering, science, and technical education									
Buttussi et al.(2018)	Yes	No	Yes	No	No	Yes	Yes	Yes	96	0.08
Lampi (2013)	No	Yes	No	No	No	Yes	Yes	Yes	56	4.00
Osner (2013)	Yes	No	No	Yes	No	No	No	Yes	322	5.00
Tatli and Ayas (2013)	Yes	No	No	Yes	No	No	Yes	Yes	90	8.00
McLaurin & Stone (2012)	No	Yes	No	No	No	Yes	Yes	Yes	21	25.00
Rupasinghe et al.(2011)	Yes	No	No	No	No	Yes	Yes	Yes	39	1.00
Stone et al.(2011)	No	Yes	Yes	No	No	Yes	Yes	Yes	22	80.00
Tanyildizi & Orhan (2007)	Yes	No	No	No	Yes	No	No	Yes	73	2.00**
Yang & Heh (2007)	Yes	No	No	Yes	No	No	No	Yes	150	7.50
Finkelstein et al.(2005)	Yes	No	No	No	Yes	No	No	Yes	222	1.00
	General education									
Allcoat et al. (forthcoming)	Yes	No	No	No	Yes	No	Yes	Yes	75	0.17
Makransky et al.(2019)	Yes	No	No	No	Yes	No	Yes	Yes	52	0.50
Allcoat & von Mühlenen (2018)	Yes	No	No	No	Yes	No	Yes	Yes	99	0.12
Parong & Mayer (2018)	Yes	No	No	No	Yes	No	Yes	Yes	55	0.20
Alhalabi (2016)	Yes	No	No	No	Yes	No	Yes	Yes	48	0.33
Kockro et al. (2015)	Yes	No	No	No	Yes	No	Yes	Yes	169	0.30
Smith (2015)	No	Yes	No	No	No	Yes	Yes	Yes	32	10.00
Webster et al. (2015)	Yes	No	No	No	No	Yes	Yes	Yes	140	0.25
Hwang & Hu (2013)	Yes	No	No	Yes	No	No	Yes	Yes	58	4.00
Akpan & Strayer (2010)	Yes	No	Yes	Yes	No	No	No	Yes	34	0.33***

Notes: RCT: Randomized Control Trial. *Type of skill assessed*: C: cognitive, T: technical, SE: socioemotional. *Grade Level*: K-12: Basic education; H.E: Higher education, TVET/OJT: Technical education and on-the-job training.

* While this study includes experiments aiming to assess socio-emotional skills, it does not clearly present the skills assessment instruments. As such, the experiments were not included in the meta-analysis.

** This study does not provide the time of exposure. Since the instruction on synchronous motors within a syllabus of electrical machines takes on average 2.5 hours (including introduction and theory), we estimated the exposure to the virtual environment to be 2 hours approximately.

*** This study does not provide the time of exposure. Since participants were enrolled in six and eight-period life science course, we assumed one of these periods to be dedicated to the topic of dissection and estimate exposure to the simulator in one of these periods to be of 20 minutes approximately.

Table A2: Instruments used to assess learning for health and surgical education

Article	Main Topic	Instrument used for skills assessment
Farra et al. (2018)	Emergency Evacuation of Neonates	**Technical skills** were assessed by having students participate in live evacuation exercises using mannequins of newborns. The research team developed a tool to asses students' performance through direct observation. Psychomotor skills were assessed using a rubric developed in collaboration with disaster experts using the Cincinnati Children's Emergency Preparedness and Response Program. **Cognitive skills** were measured using knowledge assessment developed by the researchers, based upon the course objectives and modules. The assessment included multiple-choice questions to assess students' comprehension of the topic and knowledge of its practical application.
Logishetty (2018)	Total hip arthroplasty	Technical skills were assessed by measuring: (i) the total correct tasks necessary to conduct a successful total hip arthroplasty (out of a total of 30); (ii) the errors in the component orientation in degrees (i.e. inclination with respect of the pelvis) (the higher the degrees, the higher the error); and operation time (in minutes). The task check list was developed by a pool of expert surgeons. Operation time (in minutes) was also registered. [6]
Skou-Thomsen et al. (2017)	Cataract surgery	Technical skills were assessed through performance in the operating room, using the Objective Structured Assessment of Cataract Surgical Skill (OSACSS) rating scale, a tool previously validated by the practice. Participants performed 3 consecutive phacoemulsification surgeries immediately before and after the training intervention. Procedures were recorded in video. Three raters evaluated all anonymized videos independently.
Zaveri et al (2016)	Procedural sedation	Technical skills: After completing the intervention or control module, all residents then immediately participated in a simulated pediatric procedural sedation scenario. All simulations occurred in the Simulation Center with an infant patient simulator. Simulations were video and audio recorded. Each performance video was reviewed by one or two team members blinded to the group allocation. Performance on preparation and management of a complication was assessed using a 32-point checklist, adapted for this sedation scenario from a previously published checklist. The initial checklist was determined by a consensus from a panel of experts in pediatric emergency medicine and pediatric procedural sedation.
Chao et al. (2015).	Transvaginal Gynecologic Ultrasound	Technical skills were assessed by asking participants to produce 4 images (longitudinal and axial sections of the uterus and the ovaries) and measure the uterus and each ovary. Participants were given 5 minutes of scanning time. Images with measurement calipers were stored in a database. Two blinded reviewers (M.D. and G.E.C) assessed the images in a random order two months after the trials were completed.
Kiely et al. (2015)	Robotic suturing	Technical skills were assessed by three blinded raters (two gynecologic oncologists and one gynecologic oncology fellow, all experienced in robotic surgery) using the GOALS+ score. This score is composed of the five domains (each ranging from 1 to 5 points) developed for assessing skill in laparoscopy which include autonomy, efficiency, tissue handling, depth perception, and bimanual dexterity plus two additional metrics developed specifically for robotics, precision and awareness of camera and instruments. The GOALS+ includes 7 domains, 6 of which form the GEARS score. Data allowed also to calculate the Global Evaluative Assessment of Robotic Skill (GEARS) scoring tool, a tool previously validated by the medical

[6] The study also included a procedure-based assessment with a global summary score ranging from an ability to only assist (Level 1a) to advanced competence (Level 4b). However, since the assessment did not include a numeric score, these results are not included in the meta-analysis.

		practice. Other secondary outcomes were the number total knots and satisfactory knots performed during the inanimate model suturing task.[7]
Valdis et al. (2015).	Robotic cardiac surgery	Technical skills were assessed by asking participants to complete a standardized robotic internal thoracic artery harvest and mitral valve annuloplasty performed in porcine models. The de-identified recordings of the procedures were assessed by a single investigator (to control for interobserver variability) using the Global Evaluative Assessment of Robotic Skill (GEARS) scoring tool, a tool previously validated by the practice. Time on task was also assessed.
Tschannen et al. (2012)	Nursing education	Students participated in a mannequin-based simulation. Their performance was evaluated by expert practitioners using an adapted version of the Capacity to Rescue Instrument (CRI), a tool previously validated by the practice. The instrument is designed to capture key elements (assessments, interventions) that are needed to ensure a good outcome for the patient for a specific simulation scenario. For the purpose of this study, the modified CRI consisted of 17 items measuring key concepts: communication (9 items), problem solving (4 items), and priority setting (4 items).
Zhao et al. (2011)	Bone dissection	Technical skills were assessed by asking participants to complete a cortical mastoidectomy on a cadaveric temporal bone. The participants had 1 hour to complete the procedure. Their dissections were captured using a video camera. The videos contained only the hands of the participants. These videos were then presented to 3 otologists who were blinded to whether the participant received traditional or immersive training. The otologists assessed the participants' performance using a standardized assessment tools previously validated by the practice.
Larsen et al. (2009)	Laparoscopic surgery	Technical skills were assessed by asking participants to complete a laparoscopic surgery. Two independent / blinded observers assessed their performance using the Objective Structured Assessment of Technical Skills (OSATS), an instrument previously validated by the practice. A secondary outcome assessed was operation time in minutes (time on task).
Vincent et al. (2008)	Mass casualty triage	Technical and soft skills were measures using the following outcome variables: **Triage score**: A point was given for each correct answer that was selected by the learner in the VR environment: 1) was the main problem correctly identified, 2) was the required intervention correctly identified, and 3) was the triage category correctly identified? Thus, each learner could receive a maximum of 15 points per scenario (5 victims x 3 questions per victim). **Intervention score**: A point was awarded for each intervention that was performed correctly in the VR environment. Thus, each learner could receive a maximum of 5 points per scenario. **Self-efficacy**: Subjects completed a 5 question self-efficacy questionnaire before and after the VR experience. Each question was scored on a 5-point Likert scale with points labeled ''never'' (1) to ''always'' (5).

Note: Authors own elaboration.

[7] For scoring of total knots, if a knot was partially completed at the 10 min stop time, it was scored as follows: 0.4, if the first double throw was completed and cinched down and then, an additional 0.2 for each single throw cinched down. Satisfactory knots were defined as knots that the rater would not cut out and re-suture during live surgery

Table A3: Instruments used to assess learning for engineering, science, and technical education

Article	Main Topic	Instrument used for skills assessment
Buttussi et al (2018)	Aviation safety procedures	To measure cognitive skills about cabin safety, the researchers used a test with nine questions related to: 1) what to do in case of turbulence; 2) what to do in preparation for impact; 3)which exit should be the first choice for evacuation; 4)when it is not possible to use an exit; 5) what to do if the chosen exit cannot be used; 6) what to do if there is smoke in the cabin during evacuation; 7) what to do after using a wing exit; 8) what to do after leaving the aircraft; 9) what to do with luggage. Participants were asked to answer the questions orally to avoid suggesting possible answers. Answers were audio recorded and later rated by the experimenter as correct or wrong, following a codebook that listed the possible answers and their rating (right/wrong). Knowledge was measured as the number of correctly answered questions, and thus ranged between 0 and 9. Self-efficacy was assessed using a questionnaire with six items: 1) I feel able to deal with an emergency evacuation of an aircraft; 2) I would be able to deal with an emergency evacuation even if the aircraft is on fire; 3) I would be able to deal with an emergency evacuation even if one or more exits are blocked; 4) I would be able to deal with an emergency evacuation even if most of the passengers scream or cry; 5) I feel confident of my ability to exit from the aircraft in time; 6) I would be able to help passengers in need. Each item was rated by participants on a 7-point scale (1=not at all, 7=very).
Lampi (2013)	Computer Networks	Technical skills were assessed by giving students the opportunity to configure and troubleshoot local area networks in a physical lab. Proficiency was measured by how quickly and accurately students configured and troubleshot a computer network. The instruments were developed by the researcher based on industrial certification skills objectives of the Cisco Certified Network Associate (CCNA) program (Cisco, 2012), as follow: **Configuration time** measured the amount of time a student took to configure a network. This was measured by recording the time it took to complete a network design as specified in a lab test. Time was measured in minutes. **Configuration accuracy** (0-22 points) was measured by the score a student obtained in a lab test on configuring a network design. The instrument was based on the objective performance measure (Lewis, 1993). It consisted of a check list of items of tasks that had to be completed to determine the eventual score. Correct configuration of an item scored a value of 1 and an incorrect scored a value of 0. **Troubleshooting accuracy** (0-6 points) was measured by the score a student obtained in lab test on troubleshooting a network that had a number of faults. The instrument was an objective performance measure (Lewis, 1993), with a check list of tasks that had to be completed to determine the eventual score. **Troubleshooting time** measured the amount of time a student took to troubleshooting a network that had a number of faults. This was measured by recording the time it took to complete troubleshooting a network specified in a lab test based on a CCNA® certification test. Time was measured in minutes.
Osner (2013)	Genetics	**Achievement test** The Laboratory Assessment in Genetics (LAG) was used in the study. The LAG includes a scale for assessing students' achievement in Genetics. Specifically, it measures the extent to which students understand various concepts, including Mendelian inheritance, the structure of DNA, mutations, cloning, and genetic engineering. All achievement items utilized a multiple-choice answer format with four possible responses from which to choose. Scoring was based on the number of items correctly answered and ranged from zero (0) for no correct answers to ten (10) for all correct answers. The score was then

		divided in half for meaningful comparison with scores from other sections of the LAG, which ranged from zero (0) to five (5).
Tatli and Ayas (2013)	Chemistry lab	Cognitive skills were measured using two posttest examinations: **Chemical changes unit achievement test (CCUA):** the test included 25 items to measure the learning outcomes of the course. **Laboratory equipment test (LET):** This test assessed students' ability to recognize laboratory equipment: Items in the exam, devised by the group of researchers, were prepared in order to cover all laboratory materials and equipment used in primary school science and ninth-grade chemistry courses. The test included 28 items, endorsed by five academics from departments of instructional natural science and chemistry. In addition to these 28 items, a module was added, asking students to enter the names of laboratory materials and equipment into blank spaces beneath color pictures of the related material and equipment.
McLaurin and Stone (2012)	Welding	Technical skills were assessed in two ways: (i) students who completed the training took the welding certification exam imparted the American Welding Society. Certification rates of students could oscillate between (0 and100%); (ii) students submitted their welds to a welding master expert, who would assess the quality of the weld (0-100 points).
Rupasinghe et al. (2011)	Corrosion prevention and control (aircraft Maintenance)	Cognitive skills were measured using two posttest examinations: (i) a written examination (0 to 100 points) consisting of two-tiered multiple-choice questions, fill in the blanks, and essay questions where the students had to describe and apply the concepts and procedures learned. (ii) An oral examination (0 to 100 points) where students were given several inspection scenarios and they had to describe how they would resolve the issues using the most appropriate Non-Destructive Inspection (NDI) tool. These questions were aimed at testing deeper knowledge (higher levels of knowledge using Bloom's taxonomy) on each inspection device / simulator.
Stone et al. (2011)	Welding	Technical skills were assessed using the amount of material used by trainees: amount of overall flat plates (both virtual and real-world plates) used by participants in both groups, the number of groove plates, and the number of electrodes (less usage of materials for a similar task is more desirable). [8] Socio-emotional skills were measured using the Team Learning Questionnaire (TLQ) which tracked three key dimensions of team learning and interaction: (1) Continuous Improvement Seeking (the degree to which a team can learn from previous experiences); (2) Dialogue Promotion and Open Communication (the degree to which open and honest communication is encouraged and takes place within a team); and (3) Collaborative Learning (the degree to which team members are seen and used as sources of knowledge by the rest of the team). Each dimension consists of a series of questions, which the participant answers on a five-point scale.
Tanyildizi and Orhan (2007)	Synchronous Motors	Cognitive skills were measured by pre and posttest assessing cognitive skills of the operation of synchronous motor.
Yang and Heh (2007)	Physics lab	Cognitive skills were measured using two posttest examinations: **Physics Achievement Test** (0-89 points). A 40 items test was developed from a senior high school physics textbook (items in mechanics, optics, and electricity). The content of the test was validated by two senior high school physics teachers and one physics professor. **Science Process Skills Tests** (0-36 points). Assesses the performance of the basic and integrated science process skills of the students. A point is given for every correct item. The highest score is 36 points

[8] The article also assesses the effects of VR training on pure "welding" technical skills but does not report the statistical significance of results due to the limited sample size. As such, these results are excluded from the meta-analysis.

Finkelstein et al. (2005)	Circuits design	**Technical skills:** At the end of each laboratory section, all students completed the challenge worksheet in which they were asked to build a circuit using real equipment with their groups. Teaching Assistants reported the average time used to complete the task. **Cognitive skills (write-up):** Each student in the circuit challenge completed a writeup answering the following question: "*Describe what happens and WHY the bulbs change brightness as they do. You may use words and formulas*". The answers were evaluated by the authors as to their overall correctness using a standardized rubric with a scale from 0 to 3. Zero represented no demonstrated knowledge of the domain, while 3 represented correct and complete reasoning. The research team came to consensus on the grading metric, grading not only for overall correctness, but also for use of concepts, such as current, voltage, power, series or parallel resistance; and mathematics. **Cognitive skills (test):** Three questions on circuits were included in the final examination. Q1: rank the currents through each of the bulbs; Q2: rank the voltage drops across the bulbs in the same circuit; Q3: predict whether the current through the first bulb increased, decreased, or remained the same when the switch was opened. For each student, the share of correct answers in these 3 questions was recorded.

Note: Authors own elaboration.

Table A4: Instruments used to assess learning for general education

Article	Main Topic	Instrument used for skills assessment
Allcoat et al. (forthcoming)	Science (efficiency of solar panels)	Cognitive skills were measured using a battery of 8 questions assessing students' knowledge, comprehension, and knowledge application. The questions were a mix of formats and followed Bloom's Taxonomy. Questions in the knowledge test were marked as correct or incorrect to give a total score of 0 to 8. Participants in the treatment (VR), treatment two (Mixed Reality), and control groups (traditional learning) completed the test before and after the training.
Makransky, Terkildsen, and Mayer (2019)	Biology (mammalian proteins)	Cognitive skills were assessed through two multiple-choice tests: a knowledge test (pretest) and a transfer test (post-test). A group of subject matter experts, including two scientists, two psychologists, and a psychometrician, developed the test questions. The knowledge test consisted of 10 multiple-choice questions designed to assess conceptual and procedural knowledge of essential material presented in the simulation. The questions required that students had a deep knowledge of the content and that they could apply that knowledge to a realistic context. Students received one point for each correct answer and 0 points for selecting an incorrect answer.
Allcoat and von Mühlenen (2018)	Biology (plant cells)	Cognitive skills were assessed using a test that contained 17 biology questions sourced directly from a British AQA Biology From this, 12 questions were related to the remembering of information (memorization), whereas 5 questions were more concerned with the understanding of information.
Parong and Mayer (2018)	Biology (cells in human bloodstream)	To assess cognitive skills, participants completed a posttest on the material they viewed during the lesson. The posttest consisted of 20 questions based on the lesson, including 16 factual questions in multiple-choice format and 4 conceptual questions in short-answer format. The posttest was scored out of 20 points, with a point given for each correct multiple-choice and short-answer question; half-points were given for partially correct answers on short answer question. The short answer questions were scored based on a rubric that indicated the words and phrases required for 1 point or 1/2 point.
Alhalabi (2016)	Science for engineers	Cognitive skills were assessed using a 10 question (100 points) knowledge quiz on each of the 4 topics taught: (i) Astronomy, (ii) Transportation; (iii) Networking; and (iv) Inventors. Questions assessed mostly general facts.
Kockro et al., (2015)	Anatomy of the heart	Cognitive skills were assessed using a test immediately following each teaching session. Participants were asked to complete a short examination consisting of 10 multiple-choice questions related to the content given (i.e. the topographical anatomy of the third ventricle). These questions were developed and agreed on by an expert committee of four neurosurgeons and anatomists. Each correct answer in the examination was awarded one point, with a maximum of 10 points achievable.
Smith (2015)	Job Interview	Technical skills were assessed using a role-play performance of a Job interview. Participants performed two pre-test and two post-test video recording role-play interviews. Each interview was scored 0-100 using and algorithm that assessed the appropriateness of responses based on eight domains: negotiation skills, conveying that you're a hard worker, sounding easy to work with, sharing things in a positive way, sounding honest, sounding interested in the position, behaving professionally, and establishing interviewer rapport. Participants self-reported self-confidence in a pre and post-test. Participants rated their self-confidence at interviewing using a 7-point Likert scale to answer nine questions, with higher scores reflecting more positive views (e.g., "How comfortable are you going on a job interview?"). Total scores at pre-test and post-test had strong internal consistencies ($\alpha = 0.95$ and $\alpha = 0.92$, respectively).
Webster (2015)	Science (corrosion	Cognitive skills were assessed using a test consisting of 22 questions. Questions 1 to 16 had 4 possible answer choices, while questions 17 to 22 had 6. The test served as pre-test and post-test. However, the post-exam had 5 questions that were different from the pre-

	prevention and control)	exam (i.e. 17 common exam questions). Corrosion subject matter experts and instructors validated the content of the exams. The test evaluated five topics: (1) the importance of corrosion prevention and control, (2) corrosion basics, (3) corrosion influences, (4) corrosion types, and (5) basic corrosion prevention. The topics were aligned with six learning objectives: (i) demonstrate knowledge of why CPC is important by identifying and selecting the outcomes of past lack thereof; (ii) demonstrate knowledge of corrosion by identifying and selecting characteristics of the definition; (iii) demonstrate knowledge of the mechanics of corrosion by identifying and selecting the individual components of corrosion and possible influences; (iv) demonstrate knowledge by identifying different types of corrosion by selecting each type; (v) demonstrate knowledge of different types of corrosion by identifying and selecting characteristics of each type; and (vi) demonstrate knowledge of basic CPC techniques, theories, and principles.
Hwang and Hu (2013)	Basic geometry	Cognitive skills were assessed using test Scores (20-100 points in 5 levels): a pretest and a post-test were delivered. The test evaluated four dimensions including the mathematics context, cognitive processes, types of representations, and specific tasks. The test contains rubric at five score levels for examining subjects' understanding of the geometric problems. Level 1: Solutions not related to the problem and no explanation is provided (20 points); Level 2:The process leading to the solution is reasonable but the final answer is incorrect (40 points); Level 3: The answer or equation is correct but without textual or graphical explanation (60 points); Level 4: The answer is correct, textual or graphical explanation of the process leading to the solution is correct but partially provided (80 points); Level 4: The answer is correct, textual or graphical explanation of the process leading to the solution is correct and thoroughly provided (100 points
Akpan and Strayer (2010)	Frog dissection	Cognitive skills were assessed using a Dissection Achievement Test. The test was designed by a life science classroom instructor in cooperation with three science experts. Questions were designed to meet the objectives of dissection as contained in the Modern Biology textbook and the national curricula. Example questions include: "The organ responsible for filtering toxins from the blood is the? (a) spleen (b) kidneys (c) heart (d) liver." The test was used as a pretest and posttest and had 25 multiple choice items (ten focused on identification of organs and fifteen related to the functional knowledge of anatomy and morphology) and a short answer section. Socio-emotional skills were measured using an attitude self-assessment. The assessment measured student's attitudes towards: (i) dissection (9 items), (ii) school/science (4 items); and (iii) computers (10 items). Twenty-three of the items included in the test were adopted from previous research/available instrument. Two items were developed by the researchers.

Note: Authors own elaboration.

Table A5: Impact of VR training on learning performance (cognitive skills) as proxied by results in posttest

Article (1)	Skills Developed (2)	Performance Metric (3)	Posttest Results (4)
Health and Safety			
Farra et al. (2018)	Knowledge on emergency preparedness	Cognitive Assessment. Score range: 0-100 points	Treatment: 74.2; Control: 70.7 (N.S)
Engineering, Science, and Technical Education			
Osner (2013)	Knowledge about genetics	Cognitive Assessment Score range: 0-5	Treatment: 2.78; Control: 2.90 (N.S.)
Tatli and Ayas (2013)	Knowledge about chemical changes	Cognitive Assessment Score range: 0-100	Treatment: 59.33; Control: 55.33 (N.S.)
Tatli and Ayas (2013)	Knowledge about laboratory equipment	Cognitive Assessment Score range: 0-100	Treatment: 67.41; Control: 35.43 (***)
Rupasinghe et al. (2011)	Usage of a borescope to assess aircraft corrosion.	Cognitive Assessment (written) Score range: 0-100	Treatment: 58; Control: 60 (N.S.)
Rupasinghe et al. (2011)	Usage of a borescope to assess aircraft corrosion.	Cognitive Assessment (oral) Score range: 0-100	Treatment: 83; Control: 60 (***)
Rupasinghe et al. (2011)	Knowledge of Eddy current inspection to assess corrosion	Cognitive Assessment (written) Score range: 0-100	Treatment: 71; Control: 75 (N.S.)
Rupasinghe et al. (2011)	Knowledge of Eddy current inspection to assess corrosion	Cognitive Assessment (oral) Score range: 0-100	Treatment: 85; Control: 78 (*)
Tanyildizi and Orhan (2007)	Knowledge of synchronous motors	Cognitive Assessment Score range: 0-30	Treatment: 24.28; Control: 21.00 (***)
Yang and Heh (2007)	Knowledge of physics	Cognitive Assessment Score range: 0-89	Treatment: 61.01; Control: 53.89 (***)
Yang and Heh (2007)	Knowledge of science processes	Cognitive Assessment Score range: 0-36	Treatment: 26.43; Control: 23.69 (***)
Finkelstein et al. (2005)	Knowledge of circuits operation	Cognitive Assessment (written) Score range: 1 to 3	Treatment: 1.86; Control: 1.64 (**)
Finkelstein et al. (2005)	Knowledge of circuits operation	Cognitive Assessment Score range: 0-100	Treatment: 59.3; Control: 47.6 (***)
General Education			
Allcoat and von Mühlenen (2018)	Knowledge of the parts of a plant cell	Cognitive Assessment. Score range: 0-100% (% questions that are correct).	**VR vs video** Treatment (VR): 56.5; Control: 43.9 (***)
Allcoat and von Mühlenen (2018)	Knowledge of the parts of a plant cell	Cognitive Assessment. Score range: 0-100% (% questions that are correct).	**VR vs textbook** Treatment: 56.5; Control:50.2 (N.S)
Allcoat and von Mühlenen (2018)	Memorizing the parts of a plant cell	Cognitive Assessment. Score range: 0-100% (% questions that are correct).	**VR vs video** Treatment: 55.1; Control:40.6 (***)
Allcoat and von Mühlenen (2018)	Memorizing the parts of a plant cell	Cognitive Assessment. Score range: 0-100% (% questions that are correct).	**VR vs textbook** Treatment: 55.1; Control = 43.6 (**)
Parong and Mayer (2018)	Knowledge of how human cells work	Cognitive Assessment. Score range: 0-20 points.	**VR vs slideshow** Treatment: 10.17; Control:13.54 (***)
Alhalabi (2016)	Knowledge of anatomy	Cognitive Assessment. Score range: 0-100 points.	**VR vs lecture** Treatment: 93.0;

			Control: 69.0 (**)
Alhalabi (2016)	Knowledge of transportation	Cognitive Assessment. Score range: 0-100 points.	**VR vs lecture** Treatment: 90.0; Control: 60.0 (**)
Alhalabi (2016)	Knowledge of science networking	Cognitive Assessment. Score range: 0-100 points.	**VR vs lecture** Treatment: 38.0; Control: 22.0 (**)
Alhalabi (2016)	Knowledge of famous inventors	Cognitive Assessment. Score range: 0-100 points.	**VR vs lecture** Treatment: 15.0; Control: 8.0 (**)
Kockro et al., (2015)	Knowledge about the anatomy of the human ventricular system	Cognitive Assessment. Score range: 0-10 points	**VR vs slideshow** Treatment: 5.45; Control: 5.19 (N.S.)
Hwang and Hu (2013)	Calculation of the volume and area of 3D objects	Cognitive Assessment. Score range: 20-100 points.	**VR vs textbook** Treatment: 70.24; Control: 59.17 (**)
Akpan and Strayer (2010)	Knowledge of frog anatomy	Cognitive Assessment. Score range: 0-25 points.	**VR vs dissection** Treatment: 23.33; Control: 16.94 (***)

Note: N.S: Not statistically significant at a 10 percent confidence level. *10% significance level; ** 5% significance level; *** 1% significance level.

Table A6: Impact of VR training on learning performance (technical skills) as proxied by results in posttest.

Article (1)	Skills Developed (2)	Performance Metric (3)	Posttest results (4)
Health and Safety			
Farra et al. (2018)	Performance of emergency evacuation	Ability assessment Score range: 0-100 points	Treatment: 86.5; Control: 71.1 (***)
Logishetty (2018)	Performance of total hip arthroplasty	Ability Assessment. Score range: 0-30 points	Treatment: 22.0; Control: 12.0 (***)
Zaveri et al (2016)	Residents learn how to conduct pediatric procedural sedation.	Ability Assessment Score range: 0-32 points	Treatment: 24.0; Control: 22.5 (N.S)
Chao et al. (2015).	Performance of transvaginal gynecologic ultrasound (experienced surgeons)	Ability Assessment Score range: 0-19	Treatment: 12.0; Control: 9.0 (**)
Valdis et al. (2015).	Performance of robotic internal thoracic artery harvest and mitral valve annuloplasty	Ability Assessment. Score range: 6-30 points	Treatment: 22.8; Control: 11.0 (***)
Tschannen et al. (2012)	Nurses show improvement in the following skills: (i) priority siting (focus on the patient), (ii) communications (with patient, second nurse, and physician), and (iii) problem solving (request assistance when needed).	Ability Assessment. Score: 0 to 22 points	Treatment: 21.9; Control: 20.1 (**)
Zhao et al. (2011)	Performance of cortical mastoidectomy on a cadaveric temporal bone	Ability Assessment. Score range: 0-100 points	Treatment: 67.0; Control: 29.0 (***)

Larsen et al. (2009)	Performance of laparoscopic surgery	Ability Assessment. Score range: 0-100 points	Treatment: 33.0; Control: 23.0 (***)
Engineering, Science, and Technical Education			
Lampi (2013)	Computer network configuration accuracy	Ability Assessment. Score range: 0-22 points	Treatment: 19.36; Control: 18.36 (N.S.)
Lampi (2013)	Computer network troubleshooting accuracy	Ability Assessment. Score range: 0-6 points	Treatment:4.6; Control: 4.4; (N.S.)
McLaurin and Stone (2012)	Performance of horizontal filet weld (2F)	Ability Assessment. Score range: 0-100 points	Treatment: 90; Control: 92 (N.S.)
McLaurin and Stone (2012)	Performance of flat groove weld (1G)	Ability Assessment. Score range: 0-100 points	Treatment: 90; Control: 88 (N.S.)
McLaurin and Stone (2012)	Performance of vertical filet weld (3F)	Ability Assessment. Score range: 0-100 points	Treatment:72; Control:81 (N.S.)
McLaurin and Stone (2012)	Performance of vertical groove weld (3G)	Ability Assessment. Certification rate: 0-100%	Treatment: 10; Control:45 (**)
McLaurin and Stone (2012)	Performance of vertical groove weld (3G)	Ability Assessment. Score range: 0-100 points	Treatment:53; Control: 61 (N.S.)

Note: N.S: Not statistically significant at a 10 percent confidence level. *10% significance level; ** 5% significance level; *** 1% significance level

Table A7: Impact of VR training on learning performance (socio-emotional skills) as proxied by results in posttest

Article (1)	Skills Developed (2)	Performance Metric (3)	Posttest results (4)
Engineering, Science, and Technical Education			
Stone et al. (2011)	Continuous improvement seeking	Socio-emotional Assessment Score range: 1-5 points	Treatment: 4.47; Control: 4.14 (N.S.)
Stone et al. (2011)	Dialogue promotion and open communication	Socio-emotional Assessment Score range: 1-5 points	Treatment: 4.63; Control: 3.85 (***)
Stone et al. (2011)	Collaborative learning	Socio-emotional Assessment Score range: 1-5 points	Treatment: 4.73; Control: 3.30 (***)

Note: N.S: Not statistically significant at a 10 percent confidence level. *10% significance level; ** 5% significance level; *** 1% significance level.

Table A8: Impact of VR training on learning gains (cognitive skills) as proxied by results in posttest minus pretest for treatment group only.

Article (1)	Skills Developed (2)	Performance Metric (3)	Value added (4)
Engineering, Science, and Technical Education			
Buttussi et al (2018)	Knowledge of airplane cabin safety (VR Narrow view)	Cognitive assessment. Score range: 0-9 points	Pretest: 4.7; Posttest: 7.5 (***)
Buttussi et al (2018)	Knowledge of airplane cabin safety (VR Wide view)	Cognitive assessment. Score range: 0-9 points	Pretest: 4.5; Posttest: 7.7 (***)
Tatli and Ayas (2013)	Knowledge about chemical changes	Cognitive Assessment Score range: 0-100	Pretest: 39.66; Posttest: 59.33 (***)
Tatli and Ayas (2013)	Knowledge about laboratory equipment	Cognitive Assessment Score range: 0-100	Pretest: 29.66; Posttest: 67.41 (***)
General Education			
Allcoat et al. (forthcoming)	Knowledge of solar-power panel efficiency.	Cognitive Assessment Score range: 0-8	Pretest: 1.96; Posttest: 5.30 (***)
Webster (2015)	Knowledge of basic corrosion prevention and control.	Cognitive Assessment Score range: 0-100 points	Pretest: 66.9; Posttest: 79.3 (***)
Akpan and Strayer (2010)	Knowledge of frog anatomy	Cognitive Assessment. Score range: 0-25 points.	Pretest: 10.18; Posttest: 23.33 (***)

Note: N.S: Not statistically significant at a 10 percent confidence level. *10% significance level; ** 5% significance level; *** 1% significance level.

Table A9: Impact of VR training on learning gains (technical skills) as proxied by results in posttest minus pretest for treatment group only.

Article (1)	Skills Developed (2)	Performance Metric (3)	Value-added (4)
Health and Safety			
Skou-Thomsen et al. (2017)	Performance of eye cataract removal for novice surgeons (those with less than 75 procedures completed)	Ability Assessment. Score range: 0-53 points	Pretest: 15.33; Posttest: 20.31 (***)
Skou-Thomsen et al. (2017)	Performance of eye cataract removal for intermediate surgeons (75-999 procedures completed)	Ability Assessment. Score range: 0-53 points	Pretest: 25.81; Posttest: 35.58 (**)
Skou-Thomsen et al. (2017)	Performance of eye cataract removal for experienced surgeons (1000+ procedures completed)	Ability Assessment. Score range: 0-53 points	Pretest: 43; Posttest: 43 (N.S.)
Kiely et al. (2015	Performance of robotic suturing (vaginal cuff model)	Ability Assessment (GOALS+) Score range: 0-35 points	Pretest:15.1; Posttest:21.4 (***)
Kiely et al. (2015	Performance of robotic suturing (vaginal cuff model)	Ability Assessment (GEARS) Score range: 6-30 points	Pretest:12.7; Posttest:18.4 (***)
Kiely et al. (2015	Performance of robotic suturing (vaginal cuff model)	Ability Assessment. Total knots completed	Pretest:1.65; Posttest:3.38 (***)
Kiely et al. (2015	Performance of robotic suturing (vaginal cuff model)	Ability Assessment. Satisfactory knots completed	Pretest:1.04; Posttest:2.73 (***)
Vincent et al. (2008)	Performance of mass casualty triage	Ability assessment (triage) Score range: 0-15 points	Pretest: 9.7; Posttest: 13.4 (***)

Vincent et al. (2008)	Performance of mass casualty triage	Ability assessment (accuracy) Score range 0-5 points	Pretest: 3.4; Posttest: 4.7 (***)
General Education			
Smith (2015)	Job Interview performance	Ability assessment Score range: 0-100 points	Pretest: 33.8; Posttest: 36.5 (***)

Note: N.S: Not statistically significant at a 10 percent confidence level. *10% significance level; ** 5% significance level; *** 1% significance level.

Table A10: Impact of VR training on learning gains (socio-emotional skills) as proxied by results in posttest minus pretest for treatment group only.

Article (1)	Skills Developed (2)	Performance Metric (3)	Value added (4)
Health and Safety			
Vincent et al. (2008)	Self-efficacy: I am confident in my ability to prioritize the treatment of patients in a mass casualty situation	Socio-emotional Assessment Score range: 1(never)-5(always)	Pretest: 3.8; Posttest: 4.1 (***)
Vincent et al. (2008)	Self-efficacy: I am confident in my ability to prioritize the use of resources in a mass casualty situation	Socio-emotional Assessment Score range: 1(never)-5(always)	Pretest: 3.1; Posttest: 4.2 (***)
Vincent et al. (2008)	Self-efficacy: I am confident in my ability to identify high risk patients for immediate treatment in a mass casualty situation.	Socio-emotional Assessment Score range: 1(never)-5(always)	Pretest: 3.4; Posttest: 4.2 (***)
Vincent et al. (2008)	Self-efficacy: I am confident that I will learn to be an effective first responder	Socio-emotional Assessment Score range: 1(never)-5(always)	Pretest: 4.0; Posttest: 4.2 (**)
Vincent et al. (2008)	Self-efficacy: I am confident that patients will consider me an effective first responder.	Socio-emotional Assessment Score range: 1(never)-5(always)	Pretest: 3.8; Posttest: 4.1 (***)
Engineering, Science, and Technical Education			
Buttussi et al (2018)	Self-efficacy addressing an emergency (VR narrow view)	Socio-emotional assessment Scope range: 1(not at all), 7(very)	Pretest: 2.9; Posttest: 4.3 (***)
Buttussi et al (2018)	Self-efficacy addressing an emergency (VR Wide view)	Socio-emotional assessment Scope range: 1(not at all), 7(very)	Pretest: 3.0; Posttest: 4.0 (***)
General Education			
Smith (2015)	Self-confidence interviewing	Socio-emotional assessment Score range: 1-7 points.	Pretest: 42.5; Posttest: 50.2 (N.S.)
Akpan and Strayer (2010)	Attitude towards frog dissection	Socio-emotional assessment Score range: 1-5 points.	Pretest: 2.60 Posttest: 2.66 (N.S)
Akpan and Strayer (2010)	Attitude towards science	Socio-emotional assessment Score range: 1-5 points.	Pretest: 2.64 Posttest: 2.73 (N.S.)
Akpan and Strayer (2010)	Attitude towards use of PC	Socio-emotional assessment Score range: 1-5 points.	Pretest: 2.63 Posttest: 2.71 (N.S.)

Note: N.S: Not statistically significant at a 10 percent confidence level. *10% significance level; ** 5% significance level; *** 1% significance level.

Table A11: Differences in learning gains (cognitive, technical and socio-emotional skills) between VR (treatment) vs. traditional training (control)

Article (1)	Skills Developed (2)	Performance Metric (3)	Value added (4)
Farra et al. (2018)	Knowledge on emergency preparedness	Cognitive Assessment. Score range: 0-100 points	Treatment :18.6; Control: 12.0 (N.S)
Zaveri et al (2016)	Knowledge about pediatric procedural sedation	Cognitive Assessment. Score range: 0-20 points	Treatment :1.0; Control: 3.0 (***)
Tanyildizi and Orhan (2007)	Knowledge of synchronous motors	Cognitive assessment (written) Score range: 0-30	Treatment:9.08; Control: 6.18 (**)
Allcoat et al. (forthcoming)	Knowledge of solar-power panel efficiency.	Cognitive Assessment Score range: 0-8	VR vs Slideshow Treatment:3.24; Control: 2.68 (N.S.)
Makransky, Terkildsen, and Mayer (2019)	Knowledge of mammalian transient protein expression	Cognitive Assessment. Score range: 0-10 points	VR vs Desktop Treatment:1.54; Control: 2.69 (***)
Kiely et al. (2015)	Performance of robotic suturing (vaginal cuff model)	Ability Assessment (GOALS+) Score range: 0-35 points	Treatment: 6.4; Control: 2.2 (**)
Kiely et al. (2015)	Performance of robotic suturing (vaginal cuff model)	Ability Assessment (GEARS) Score range: 6-30 points	Treatment: 5.7; Control: 2.0 (**)
Kiely et al. (2015)	Performance of robotic suturing (vaginal cuff model)	Ability Assessment. Total knots completed	Treatment: 1.73; Control: 0.97 (**)
Kiely et al. (2015)	Performance of robotic suturing (vaginal cuff model)	Ability Assessment. Satisfactory knots completed	Treatment: 1.69; Control: 0.85 (**)

Note: N.S: Not statistically significant at a 10 percent confidence level. *10% significance level; ** 5% significance level; *** 1% significance level.

Table A12: Impact of VR training on learning efficiency

Article (1)	Skills Developed (2)	Performance Metric (3)	Learning efficiency (4)
Health and Surgical Education			
Logishetty (2018)	Performance of total hip arthroplasty	Operation time (in minutes)	Treatment: 12; Control: 24 (***)
Logishetty (2018)	Performance of total hip arthroplasty	Inclination error (in degrees)	Treatment: 3; Control: 15 (***)
Logishetty (2018)	Performance of total hip arthroplasty	Anteversion error (in degrees)	Treatment: 4; Control: 16 (***)
Valdis et al. (2015).	Performance of robotic internal thoracic artery	Time on task (in minutes)	Treatment: 342.7; Control: 856.2 (***)
Valdis et al. (2015).	Performance of robotic internal mitral valve annuloplasty	Time on task (in minutes)	Treatment: 139.6; Control: 256.2 (***)
Larsen et al. (2009)	Performance of laparoscopic surgery	Time on task (in minutes)	Treatment: 12; Control: 24 (***)
Engineering, Science, and Technical Education			
Lampi (2013)	Computer network configuration time	Configuration time (minutes)	Treatment:43.5; Control :50.0; (N.S.)
Lampi (2013)	Computer network troubleshooting time	Troubleshooting time (minutes)	Treatment:8.21; Control:9.87 (**)
Stone et al. (2011)	Optimal usage of flat plates	Number of plates used	Treatment:210; Control: 288 (***)

Stone et al. (2011)	Optimal usage of welding materials (groove plates)	Number of groove plates used	Treatment :50; Control :63 (**)
Stone et al. (2011)	Technical – usage of welding materials (electrodes)	Number of electrodes used (in pounds)	Treatment :111; Control: 188 (***)
Finkelstein et al. (2005)	Building a circuit	Time in task (in minutes)	Treatment :14.0; Control: 17.7 (***)

Note: N.S: Not statistically significant at a 10 percent confidence level. *10% significance level; ** 5% significance level; *** 1% significance level.