Doozandeh, P. (2020). From surface realism to training considerations: A proposal for changing the focus in the design of training systems. *Theoretical Issues in Ergonomics Science*, (in press) xx–xx. https://doi.org/10.1080/1463922X.2020.1849442

From Surface Realism to Training Considerations: A Proposal for Changing the Focus in the Design of Training Systems

Pooyan Doozandeh (pooyan.doozandeh@gmail.com)

College of Information Sciences and Technology, The Pennsylvania State University.

Author Note and Acknowledgements

This research was supported by the Office of Naval Research (N00014-18-C-7015) funded through and with Charles River Analytics Inc. For providing helpful comments during the preparation of this article, I am grateful to Frank E. Ritter, Shekoofeh Hedayati, Mathieu Brener, Shota Matsubayashi, Martin Yeh, James Niehaus, Mallory Dixon, and Matthew Norris.

Frank E. Ritter, the co-principal investigator of the project, is required by the Penn State Conflict of Interest Program to include this paragraph [sic]: "I have financial interest with Charles River Analytics Inc., a company in which I provide consulting services and could potentially benefit from the results of this research. The interest has been reviewed and is being managed by The Pennsylvania State University in accordance with its individual Conflict of Interest policy, for the purpose of maintaining the objectivity of research at The Pennsylvania State University."

Abstract

Traditionally, it was believed that realistic, high-fidelity simulation could result in effective training-systems. By showing the potentials of low-fidelity simulations, numerous research projects cast doubt on the traditional belief in high-fidelity simulation. Because the construct of fidelity has guided the training-system design for decades, a growing disbelief in fidelity has created a void in design, leaving designers without resources in their decision-making. This article first presents a historical review that shows how fidelity has been used in research and design, and how it was later challenged by researchers. Then, in filling the void of fidelity, attempts at providing design guidelines, such as trainee- and task-analysis, are reviewed and their strengths and limitations are discussed. Toward the end, the lessons from the review are used to call for the creation of design methods. An example of such a method is discussed that uses the videorecording of expert performance as a resource in design.

Keywords: simulation fidelity, design method, training, expert performance, task-analysis

Relevance to human factors/ergonomics theory

Training-system design has engaged researchers in human factors and ergonomics for decades. This review shows how the theory of training-system design has been dominated by a traditional reliance on high-fidelity simulation. Nonetheless, researchers and designers can no longer rely on the traditional theory due to the challenges to the construct of fidelity. By presenting the lessons from the history of the field, this review is calling for creating new design methods that would shift the focus from the surface realism to important elements of training.

Introduction

Practice is required for skill acquisition. Today, various tasks and occupations need formal training, and as a result, private and public organizations make large investments in training their human resources (Salas, Tannenbaum, Kraiger, & Smith-Jentsch, 2012). A difficulty arises when for some tasks (e.g., flying an airplane), it is dangerous or inefficient to have trainees practice tasks with real equipment in target environments. To enable trainees to practice those tasks, technology tools are used to simulate target tasks in environments and with devices that are safer and cheaper than practicing with real devices in target environments.

Such tools—that are known as training simulators—duplicate target environments to provide trainees with the opportunity of practice. The skills that are developed through practicing with simulators should be transferred to target environments. Domains of practice such as maintenance (e.g., Rouse, 1981), aviation (e.g., Adams, 1979), firefighting (e.g., Williams-Bell et al., 2015), and nursing (Cook et al., 2011) make extensive use of simulators for training. For example, military organizations have used the benefits of training simulators for decades. One study estimated that using simulation-based training saved more than $6 billion of costs for the United States Marine Corps in 2014 (Cooley, Seavers, Gordon, Roth, & Rodriguez, 2015), and this does not include potential savings in training time and reducing dangers of in-field practice for recruits and personnel.

Designing training-systems has been the subject of considerable attention from researchers for more than a century (e.g., Link, 1937; Hays & Singer, 1989; Salas, Wilson, Priest, & Guthrie, 2006; for a historical summary, see Page, 2000). Because designers should simulate target environments, the traditional design practice aimed at creating a realistic duplication of target environments and devices (e.g., Lee, 2005; Smode, 1971, p. 3). In other

words, training-system design was traditionally focused on increasing "simulation fidelity"—i.e., the degree of similarity between the simulation and the target environment (Hays & Singer, 1989; Williges, Roscoe, & Williges, 1973). The belief in high-fidelity simulation from this *traditional theory* was later questioned by a growing number of findings that challenged the sole reliance on high-fidelity simulation for training (e.g., Norman, Dore, & Grierson, 2012; Roscoe, 1991; Swezey, Perez, & Allen, 1991). As a result of this situation, using the construct of fidelity has become problematic in training-system design (Hamstra, Brydges, Hatala, Zendejas, & Cook, 2014). The lack of a systematic alternative to replace fidelity in providing design guidelines has created a confusion for designers. This confusion is reflected in the gap between the theoretical research and design practice that wreaked havoc on the field for decades (e.g., Campbell, 1971; Cannon-Bowers, Tannenbaum, Salas, & Converse, 1991; Fowlkes, Neville, Owens, & Hafich, 2009; Goldstein, 1978; Roberts, Stanton, Plant, Fay, & Pope, 2020).

With the goal of taking a step in resolving the current confusion in design, this article first reviews the historical role of fidelity in training-system design. Unlike most of the past attempts that focused on one domain of practice, the current review adopts a cross-disciplinary approach by focusing on three domains that make extensive used of training-systems: maintenance, aviation, and medical training. This approach has the benefit of unifying, and using the experience of, researchers across domains in defining design problems and proposing solutions. Additionally, the growing use of novel technology tools (e.g., mixed-reality devices) and their respective training benefits is discussed.

The body of the review is composed of more than 100 theoretical and empirical articles that were published during the last century (1920 – present) in psychology, human factors and ergonomics, human-computer interaction, engineering, and healthcare. Various online tools—

including Google, PsycINFO, PubMed, CiteSeer, Google Scholar, Web of Science, Scopus, Microsoft Academic, and Wikipedia—as well as local library tools were used to search the literature. The keywords that were used in the online search included, but were not limited to, "training system", "simulation fidelity", "training device", "aviation training", "simulation in healthcare", "part-task training", "educational technologies", "mixed-reality training", and "science of training". In the inclusion of articles, peer-reviewed theoretical and experimental reports published in well-known journals are prioritized; nonetheless, the results of experiments from technical reports and other sources that met the review's criteria are also included. The table in the Appendix shows a list of articles and reports in the domains of training that constitute the main section of the review. It should also be noted that this review has a broad view on simulation, and as such, the selected articles include both interactive devices (e.g., driving simulators in which trainees use controlling devices to practice) and non-interactive training media (e.g., instructional films and audio). And the terms "training-system" and "training simulator" are used interchangeably to refer to the same group of complex devices that are used for training.

After presenting the review, an outline of the lessons from the literature is presented that would lead us to consider new resources for creating design methods. Toward the end of the article, we will see how studying expert performance, considering trainees' characteristics, and other important elements of training have the potential to fill the void of fidelity in providing design guidelines.

The Traditional Theory

Although using training simulation dates back to the early 1900s (Page, 2000), the modern history of training-systems and the surge in using them can be represented by reviewing

the tools and devices that were introduced during the latter half of 1900s. The experience of two world wars and the Cold War accelerated the use of complex technological devices and systems in the military. To work with and maintain those devices, operators needed special training. This is one of the main reasons that the decades following the 1950's saw investments in the research and design of training programs and devices (for a review, see Koonce, 1984).

The growth and popularity of computational tools also facilitated the production of training-systems (Smode, 1974). Digital computers enabled designers to simulate target environments in training-systems. For instance, "Sim One" was a manikin system that used the latest computational technologies of the 1960's to simulate a human patient for medical education (Abrahamson, Denson, & Wolf, 1969). As another well-known instance of an early modern training simulator, the emergence of the Apple II personal computers during late 1970's inspired the idea of building a computerized simulation to replace the expensive training of tank gunnery skills. This project led to the development of "simulator networking"—or SIMNET—during the 1980's for comprehensive military training (Alluisi, 1991).

Designing training-systems mostly revolved around how to simulate target devices and environments. Without questioning what details of the target environment to simulate and how, research and design projects assumed the effectiveness of realistic simulations and were mostly concerned with increasing the surface realism of training tools. This assumption, and the trend in using simulation in training, can be summarized in an exemplary report by Miller (1954). As shown in Figure 1, Miller formalized the relationship between the degree of simulation (or fidelity in today's terminology) and the effectiveness of training (or transfer). In Miller's view, more effective training-systems were those with higher levels of simulation. But because building realistic simulators was costly, the problem was finding the most efficient level of

simulation in training-systems. The "point of diminishing returns" in Figure 1 is the level of simulation that was argued to be the optimal level for training simulators (for a closely related early discussion on the topic, see Gagne, 1954; and, for a discussion on diminishing returns and cost-effectiveness of using training simulation for pilots, see Povenmire & Roscoe, 1973).

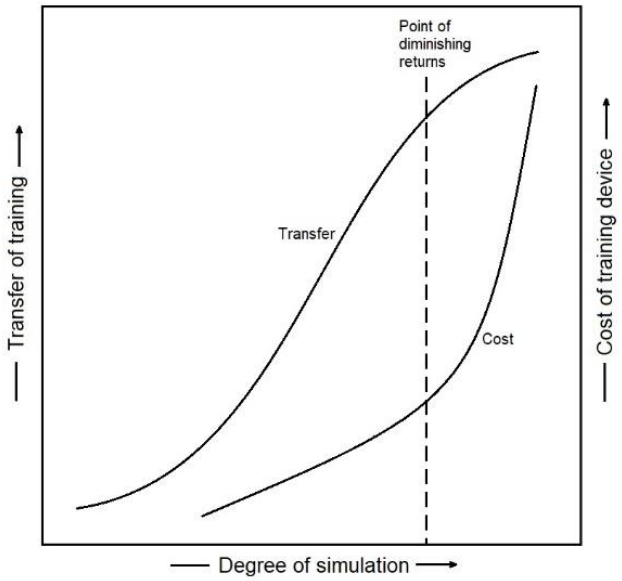


Figure 1. Miller's (1954) argument for the relationship between the degree of simulation and transfer of training. The "point of diminishing returns" is the optimum point for building cost-effective training simulation.

Many studies of the time had similar messages as those of Miller in arguing for the higher training effectiveness of realistic simulation (e.g., Dougherty, Houston, & Nicklas, 1957; Ornstein, Nichols, & Flexman, 1954). The notion that higher similarities between training and target devices could result in effective training had its roots in psychological theories. In particular, the dominant theory of psychology on the subject was the "identical elements theory" of Thorndike and Woodworth (1901). The theory proposes that higher levels of similarity between the trained and the target task would lead to the increased transfer of training. Therefore, training-systems should aim at creating realistic simulations that had high levels of

shared elements with the target environments (for another psychological justification of realistic simulation, see Logan, 1988).

There seemed to be a consensus in the effectiveness of higher levels of simulation in training-systems. For the sake of discussion and future references, this article refers to this belief as *the traditional theory* in the design and evaluation of training-systems. The traditional theory argued that more effective training-systems needed to replicate the target environment as realistically as possible in training devices. Lee (2005) summarized the attitude of the traditional theory in the domain of flight training simulation in the following sentence:

> For the designer, the closer the appearance of the system is to the ideal of full or complete physical fidelity in all its dimensions, the better the simulator design will be and the more effective the flight simulator will serve as an aircraft surrogate. (Lee, 2005; see also Lee, 2018)

The construct of "simulation fidelity" was introduced within the traditional theory and extensively used in research and design (Hays, 1980; Hays & Singer, 1989). High-fidelity training-systems were those that closely resembled the target environment in the eyes of designers and trainees. In this respect, fidelity refers to both *how much detail* of the target environment is included in the simulation, and *how realistically* those details are simulated (Smode, 1971). Although some authors distinguished between "how much detail" and "how realistic" and defined fidelity to refer to the latter (e.g., Williges et al., 1973), this distinction faded over time as fidelity in recent years often refers to the general similarity of simulation to the target environment (e.g., Allen, Hays, & Buffardi, 1986; Hamstra et al., 2014). As a result, for the most part, in discussions on training simulation, "fidelity" replaced "simulation" during the last few

decades, as higher levels of fidelity in today's terminology refer to higher levels of simulation in older literature.

There have been attempts to specify different types or dimensions of fidelity (for a review, see Hays & Singer 1989) such as psychological fidelity (e.g., Miller, 1954), functional fidelity (Fink & Shriver, 1978), physical or equipment fidelity (Allen et al., 1986), and environmental fidelity (Hays, 1980). However, such attempts created further confusion and difficulties in using fidelity in research (e.g., Hamstra et al., 2014; Kyaw Tun, Alinier, Tang, & Kneebone, 2015; Roberts et al., 2020; Roza. 2005). For this reason, as Table 1 shows, in most discussions fidelity as a general concept refers to the degree of surface realism of the simulation. This use of the concept of fidelity is also adopted in this review.

Table 1

Examples of how fidelity, as a general concept, has been used in various domains of training.

Domain	How fidelity was used
Flying	"When simulator engineers speak of fidelity, it is generally the physical replication or physical fidelity of the simulator design to which they refer." (Lee, 2005)
Health	"In training parlance, the term simulation fidelity has traditionally been defined as the degree to which the simulator replicates reality." (Beaubien & Baker, 2004)
Maintenance	"This degree of similarity to the actual equipment is called simulator fidelity." (Allen et al., 1986)
Driving	"fidelity generally refers to the realism or representativeness of the simulation" (Allen et al., 2010)

Because it was believed that higher levels of fidelity could lead to more effective training, and also because high-fidelity systems were costly and inefficient, a popular line of research was in determining the appropriate level of fidelity in training-systems (as shown in Figure 1)—what Hays

and Singer (1989) referred to as the "fidelity question". This required considering contextual factors that could further refine the traditional theory. For example, Alessi (1988) argued for the difference in the effect of fidelity between novice and expert users, and that high-fidelity simulators were mostly effective for expert users. Similarly, subsequent studies refined the traditional theory by considering additional contextual factors—such as trainees' level of expertise, training stage, the training task—in determining the appropriate level of fidelity for training-systems (e.g., Hays & Singer, 1989; Williges et al., 1973).

Finally, the assumption behind using high-fidelity simulation was strengthened by arguments surrounding the necessity of using *whole-task* trainers. Whole-task trainers tried to simulate the target task as a whole in training-systems. As a simple example, if the goal of a simulator is to train driving, the simulator should include the opportunity of practicing all parts of the task such as passing, exiting a highway, parallel parking, and all other parts of the task of driving. As such, the simulator needs to include the necessary details of all parts of the task in the simulation, and this would lead to the need for high-fidelity simulation of the whole task. On the other hand, *part-task* training would divide a task into parts and train each part separately. For example, one trainer (or a module in a trainer) can first train the basics of driving (e.g., initial assessments and adjustments, driving between lanes) and another trainer (or module) can train more difficult parts of driving (e.g., parallel parking). In this case, each simulator (or module) can focus on simulating necessary details only for that part of the task, and therefore, low-fidelity simulation can be sufficient for each part-task trainer.

Although most discussions in part-task training came from aviation research (e.g., Wightman & Lintern, 1985), in various domains and tasks there is the choice of whether to divide a task into parts or to train the whole task in one simulation (e.g., industrial facility tasks: Velotta, 1997; laparoscopic surgery: Spruit, Band, Hamming, & Ridderinkhof, 2014; dual-motor tasks such as

playing piano: Yuksel et al., 2016). Although the debate between proponents of whole- and part-task training has been extensive and still continues (e.g., McGeoch, 1931; Wickens, Hutchins, Carolan, & Cumming, 2013), older research favored the use of whole-task training (e.g., Adams, 1960), and this was another force in relying on high-fidelity simulation (e.g., Needham, Edwards, & Prather, 1980).

In short, it was widely assumed that higher levels of fidelity in simulation would result in more effective training (e.g., Lintern, 1991; Smode, 1971, p. 3). The question of practical significance is: did the traditional trend in using high-fidelity simulation produce effective training-systems? Subsequent research that resulted from the traditional theory should address this question. In the following, the findings in the literature are discussed that would evaluate the training-systems that were built on the premises of the traditional theory in the three domains of interest: maintenance, aviation, and medical training.

Maintenance and Troubleshooting

In one of the early empirical studies on the subject, Spangenberg (1973) conducted experiments in which the instructions of a procedural task were presented to trainees in different formats. For the task of disassembling M85 machine guns, participants (Army enlisted personnel) in the high-fidelity condition watched instructional videos, and those in the low-fidelity condition watched the same instructions with still images. In two experiments Spangenberg showed that participants in the high-fidelity condition were faster than those in the low-fidelity condition in the task of disassembly. The conclusion was that if a task requires motion (e.g., procedural tasks), motion pictures and videos can have training advantage over still images, and this finding had wide-ranging influence on subsequent research. The exact same effect was seen by Johnson and Rouse (1982) in troubleshooting aircraft engines in which authors observed the advantage of instructional videos over low-fidelity training materials such as graphical diagrams and computer instructional training. The

advantage of videos in training certain troubleshooting tasks was also shown recently by two meta-analyses: Berney and Bétrancourt (2016) and Höffler and Leutner (2007).

Lowering the performance time as a result of training with high-fidelity materials was also shown by Allen et al. (1986) who tested the effect of varying degrees of fidelity in electromechanical fault-detection training. Specifically, the high-fidelity trainer in Allen et al. was composed of physical components of a circuit (e.g., relays and pullup panels) and the low-fidelity trainer was a symbolic representation of the reference system drawn on a paper. Participants' performance was tested on a system that was similar to the high-fidelity training materials. Allen et al. found that participants who were trained in the high-fidelity condition were significantly faster in finding faulty components than participants in the low-fidelity condition. The authors mentioned that a possible difference in timing was because unlike those trained in high-fidelity condition, participants in the low-fidelity condition needed more initial time to apply what they learned in the abstract two-dimensional representation to the target physical circuit. In short, in the studies reviewed above, high-fidelity training materials lowered the performance time for maintenance and troubleshooting tasks.

Aviation and Vehicle Training

In the aviation industry which has been using training simulators for more than a century, Lintern, Roscoe, and Sivier (1990) tested the effect of various aspects of a flight training simulator. One of their results showed that training with high-fidelity pictorial displays successfully transferred to the target flight environment, while almost no transfer was seen in training with low-fidelity symbolic displays. Similarly, Gopher, Weil, and Bareket (1994) showed that cadets who passed 10 hours of training with a high-fidelity computer game had better flight performance than cadets without the game experience. And still in another study, Dennis and Harris (1998) tested the effect of using a high-fidelity flight training simulator ("representative set of flight controls") for ab-initio

student pilots before conducting real flight maneuvers, and they found the effectiveness of high-fidelity simulation in pilots' performance. Numerous other studies supported the use of simulation-based and high-fidelity systems for various aspects of training and evaluation in the aviation industry (e.g., Buttussi & Chittaro, 2018; Caro, 1988; Gerathewohl, 1969; Jentsch & Bowers; 1998; Oberhauser, Dreyer, Braunstingl, & Koglbauer, 2018).

In the domain of driving simulation, in one of a few empirical studies on the subject, Allen, Park, and Cook (2010) conducted a comprehensive experiment to analyze the effectiveness of three driving simulators. A total of 554 participants of fourteen to eighteen years of age were divided between three training conditions. In the low-fidelity condition, participants practiced driving with a simple driving simulator that was composed of a single monitor (40 degrees field of view) and a game controller with a steering wheel and pedals. Participants in the mid-fidelity condition used the same game controller of the low-fidelity condition, but with three monitors that were placed adjacent to each other facing participants (135 degrees field of view). The high-fidelity simulator was a real vehicle in which by looking through the windshield, participants could see the computer simulation that was projected on the screen in front of the vehicle and on the side mirrors. Participants' performance was defined based on the frequency of crash and accidents, both immediately after the training sessions in the simulator and 40 months after the training sessions in real-world driving experiences. Allen et al. showed that drivers who were trained with the high-fidelity trainer had significantly lower crash rates both in short-term simulation and in long-term real driving than the other two conditions. Similarly, participants' performance in the mid-fidelity condition was better than the low-fidelity condition. Authors concluded that the size and format of the screen, as well as other surrounding devices directly affected participants' training experience with both short- and long-term effects on performance (see also Roenker et al., 2003).

Within the general domain of vehicle simulation, Lee (2018) reviewed the effect of fidelity on the effectiveness of training simulators and showed that many of the high-fidelity systems used over the last few decades were effective in training. However, because of the costs in increasing fidelity, the efficiency of using those systems was questionable. For example, in an older experiment in vehicle simulation, Hammerton and Tickner (1967) studied the effect of various types of displays (i.e., TV and cathode ray tube with different characteristics) with varying degrees of fidelity for training the task of controlling a trolley on a railway. In their study, using more realistic displays resulted in better and faster transfer of training. In practice, however, because of the high costs of high-fidelity displays, Hammerton and Tickner recommended the use of simple displays for initial stages of training that could result in slower transfer of skills from simulation to the target environment. In a more recent study Taber (2014) reviewed helicopter accident reports to determine the requirements for helicopter underwater egress training-systems, and argued for the use of high-fidelity simulators in respective training programs. Although Taber did not present empirical results to support the use of high-fidelity simulation, the reason behind the argument was that the task of egressing and the associated skills require practicing with real or high-fidelity equipment in training. The overall conclusion of studies reviewed above is that the type of task in flight or driving required high-fidelity training-systems.

Medical Training

The reliance of healthcare education on the traditional theory is wide-spread as various areas of healthcare have made extensive use of high-fidelity simulation in training. This has been shown in several meta-analyses. Cook et al. (2011) conducted a meta-analysis of using training simulators in medical training and found the effectiveness of technology-enhanced simulation (for a historical review, see Rosen, 2008). Similarly, in nurse education, Cant and Cooper (2017) reviewed more than

700 empirical studies, and found the medium- and high-fidelity simulations to be effective in training nurses; their finding has widely been used and implemented in medical training programs (for similar results in nursing education, see Lewis, Strachan, & Smith, 2012; MacLean, Geddes, Kelly, & Della, 2019; Shin, Park, & Kim, 2015). Other examples of healthcare domains that use high-fidelity simulation in training are anesthesiology (Green, Tariq, & Green, 2016), pediatrics (Lopreiato & Sawer, 2015), surgery (Alaraj et al., 2011; Schlickum, Hedman, Enochsson, Kjellin, & Fellander-Tsai, 2009) and dentistry (Buchanan, 2001).

Despite the widespread use of high-fidelity simulation in healthcare training, it was difficult to find empirical studies supporting the use of high-fidelity simulation based on comparisons with low-fidelity simulation. For example, Seymour et al. (2002) compared the training of laparoscopy surgery with and without a virtual-reality (VR) system and showed the benefit of using the VR device; it is not clear what a lower-fidelity simulation would have achieved compared to the VR device. In fact, the experiments and reports that were reviewed in the general domain of medical training did not manipulate fidelity levels to support the effectiveness of high-fidelity simulation, but mostly compared simulation with no simulation conditions. As a result, the comparative effect of high- and low-fidelity simulation on performance was often not shown by studies that supported high-fidelity simulation.

Mixed Reality

Mixed-reality (MR)—i.e., augmented reality (AR) and virtual reality (VR)—systems are used for training with a growing pace (e.g., Dorsey, Campbell, & Russell, 2009; Hughes, Stapleton, Hughes, & Smith, 2005). Because they commonly present training with most of the environmental details, most MR systems can be considered as high-fidelity systems. Many researchers have argued for the effectiveness of MR systems for training in various domains of practice. In addition to the

domain of medical training as mentioned earlier (e.g., Seymour et al., 2002; Schlickum et al., 2009), MR tools have been extensively studied and used in industry. For example, in training industrial maintenance and assembly skills, Webel et al. (2013) and Yuviler-Gavish et al. (2015) provided evidence in supporting the use of MR systems. With a similar message, Langley et al. (2016) reviewed the use of MR systems in the assembly process of the automotive industry, and found potential benefits of virtual training programs in reducing human error in performance. This potential benefit is also shown in automotive technicians' training (Anastassova & Burkhardt, 2009), aircraft maintenance training (De Crescenzio et al., 2011), operating construction equipment (Dunston, Proctor, & Wang, 2014), assembling electronic boards (Westerfield, Mitrovic, & Billinghurst, 2015), and machining skills (Nathanael, Mosialos, Vosniakos, & Tsagkas, 2016).

Finally, the emphasis of the traditional theory in using computer simulation for training drew attention from companies in commercializing products such as "serious game", "brain game", "video-game training" and "brain-training program". These projects aimed at developing computer games and similar entertaining simulations to improve players' skills for tasks other than the game (Freitas & Liarokapis, 2011). There have been recent empirical reports in psychology supporting the use of such computer simulations for general cognitive training (e.g., Bediou et al., 2018; Green & Bavelier, 2007).

Conclusion and Summary

We can infer two general conclusions so far. First, because the traditional theory believed in the realistic replication of target environments for training, the use of technology tools for training was first started by the proponents of the traditional theory. Second, higher levels of fidelity improved the training effectiveness regarding certain aspects of performance. In maintenance and

medical training, the performance improvement was mostly seen in time spent on task, and in vehicle and flight training, high-fidelity simulation resulted in reducing accident rates.

There seems to be various reasons behind the effectiveness of high-fidelity systems. The time reduction in performance is claimed to be largely due to the fact that participants who were trained with low-fidelity systems needed more time to recognize the relationship between the training and the target devices than participants who were trained with high-fidelity materials (Allen et al., 1986). Additionally, for certain tasks (e.g., procedural tasks that involve physical movements), a high-fidelity simulation of the target environment is necessary for training skills because low-fidelity simulation often excludes important information regarding the dynamic or structure of target devices (Allen et al., 2010; Spangenberg, 1973; Taber, 2014). Overall, the dependence on high-fidelity simulation resulted in most of the existing training-systems many of which were effective in practice, especially for tasks that involved timing, motion, and physical structure of tools.

Should Fidelity Determine Training Success?

As we have seen, high-fidelity simulation can have training advantages for certain tasks and circumstances. Nevertheless, solely because of those advantages we cannot declare "success" for the traditional theory. This is because there are no objective criteria as to the desired state of training (Adams, 1979; Kirkpatrick & Kirkpatrick, 2006), and more importantly because there are no alternative approaches or theories for comparison. Moreover, the studies and systems that were reviewed within the traditional theory tried to use high-fidelity simulation for training; many of those systems were not compared with low-fidelity simulations for the same tasks. In fact, except in some cases, those studies did not provide convincing reasons and evidence for the effectiveness of higher levels of fidelity (for a similar problem in medical training, see Moglia et al., 2016).

For those studies that compared the high- and low-fidelity systems, the effectiveness of the high-fidelity systems was shown regarding certain aspects of performance or certain tasks. Because the performance improvement of high-fidelity systems was mostly shown regarding timing, motion, or physical structure of devices, what if our criteria of interest were different aspects of performance (e.g., generality of the trained skills across various tasks)? It is not clear if high-fidelity systems would still work better in those cases.

If it was to be used for designing training-systems in various domains and tasks, and for multiple performance criteria, the traditional theory was not sufficient. What was needed as a theoretical backbone? We should first briefly consider what training-systems need (for a more detailed discussion, see Adams, 1979). A training-system aims to provide the opportunity of practice through which performance in the target task should be improved. In the training cycle, performance is manifested by actions and decisions that a human performer enacts in response to environmental stimuli. The training-system is responsible for providing the environmental stimuli so that trainees can practice appropriate actions and decisions in response. In this respect, trainees' recognition of stimuli is needed to trigger actions and decisions. The question is whether this recognition can be enhanced or influenced by high-fidelity simulation. In other words, should it be important whether the environmental stimuli *look* and *feel* alike their target references? The same question applies to the medium through which trainees apply their actions and decisions. When trainees recognize stimuli, are their responses influenced by the form of input devices? Providing evidence to support a positive answer to these questions was what the traditional theory needed. Nonetheless, these questions were not addressed. Instead, it was assumed that training-systems should present trainees with realistic stimuli and input devices.

One might argue that realistic observation and input devices might affect trainees' experience on factors other than the recognition of the stimuli and implementing actions, that are nonetheless important in performance (e.g., situation awareness or the believability of interfaces). But we do not have evidence to support the notion that realistic interfaces influence trainees' experience regarding other contextual factors. Conversely, as Endsley (2018, p. 727) indicated, a plethora of contextual elements—that can be found in high-fidelity simulations—can undermine trainees' performance and learning. Similarly, Sweller, Van Merrienboer, and Paas (1998) argued that increasing the details of the environment can use up users' cognitive resources; so, trainees' attention would be squandered on unnecessary details of high-fidelity simulations (for sensory moderators of situation awareness, see Hale, Stanney, Milham, Bell Carroll, & Jones, 2009; see also Mayer, 2009, pp. 1–27). In short, because of the increased details of the environment, realistic simulations can pose potential threats to the training experience.

Another argument in supporting the traditional theory comes from the belief in task specificity of training. Often, high-fidelity simulators duplicate the environmental details and devices of a specific task, and the skills that are developed in such a simulator would therefore be transferrable to that specific task and not other tasks. On the other hand, because low-fidelity simulators often do not present the details of the target environment, the skills that are developed in such training can often be transferred to a wider range of similar or dissimilar tasks. Therefore, the training in high-fidelity systems is more task-specific than low-fidelity systems. This subject is discussed in detail in the next section, but for now, it is worth mentioning that the evidence on the specificity of training does not lean toward the traditional theory. It is both because there have been numerous studies supporting task-general training (e.g., Rosa et al., 2020; Rouse,

1981), and because the respective debate in psychology still continues without firm conclusions in deciding between specific and general training paradigms (e.g., Bediou et al., 2018; Sala, Tatlidil, & Gobet, 2018). Similarly, as we will see in the next section, numerous studies support the use of low-fidelity part-task trainers (e.g., Crawford, Hurlock, Padilla, & Sassano, 1976; Wightman & Lintern, 1985). If this possibility exists, high-fidelity whole-task trainers are not always necessary, and this is further evidence that can challenge the reliance of the traditional theory on high-fidelity simulation of whole tasks.

In summary, the traditional theory was not based on validated theoretical grounds that would show whether and how high-fidelity systems could necessarily have training advantages over low-fidelity systems. The effectiveness of past systems does not indicate the validity of the traditional theory. We currently do not have reasons to believe that, solely because of their realism, high-fidelity interfaces can facilitate the recognition of stimuli or trainees' control of systems. As a result, there is neither evidence nor reason to assume that fidelity determines training outcome. In light of this deficiency in providing evidence from the traditional theory, there is a growing body of research that has touched on this issue. The following section reviews a sample of findings that challenged the traditional theory by comparing the benefits of high- and low-fidelity simulators.

Challenging Findings

The primary assumption of the traditional theory was that increased fidelity could improve the training outcome. One way to challenge this assumption was by showing how low-fidelity systems could be as effective in training as high-fidelity systems for the same or similar tasks; numerous research projects tried to show this possibility as well as other disadvantages of

high-fidelity simulators in training. In the following, the evidence in the three domains of practice is reviewed that would challenge the traditional theory.

Maintenance and Troubleshooting

Although the traditional theory was the dominant paradigm in research and design, criticisms on the reliance on high-fidelity tools for training is almost as old as the traditional theory (e.g., Laner, 1954). In one of the early critical articles, Fink and Shriver (1978) reviewed training-systems that were used in the military during the 1950's and 1960's, and questioned the trend toward realistic simulations that the nascent computers of that time could provide. With a focus on maintenance and troubleshooting tasks, their review emphasized the effectiveness of low-fidelity systems in training, especially for trainees with low levels of skill. The unnecessary details and tools in the interface of high-fidelity systems created further confusions for trainees and slowed up their progress. As a result, Fink and Shriver argued, despite many instructors' assumption in the effectiveness of high-fidelity systems, mock-ups and low-fidelity interfaces could be more effective than high-fidelity systems to create initial familiarity with tasks and prepare trainees for advanced training with high-fidelity systems or actual devices.

More than a decade later, Fink and Shriver's position was strengthened by Swezey et al. (1991), in which authors trained undergraduates for the task of electromechanical troubleshooting in two conditions with different training materials: low-fidelity materials (i.e., 35-millimeter slide), and high-fidelity materials (i.e., motion-based videotape). Immediately following the training sessions as well as after a one-week interval, participants' performance was measured in troubleshooting the ignition problem of a diesel engine through various means—i.e., a complex diesel engine simulator, hands-on tasks on a physical system, and knowledge tests. Swezey et al. did not find a significant difference in participants' performance between the low-fidelity and high-fidelity training conditions.

By reflecting on earlier findings of this review, it is likely that the maintenance and troubleshooting training in Swezey et al. did not benefit from the advantages of motion-based instructions as the task did not involve motion.

Discussing an example in more detail can clarify certain benefits of using low-fidelity simulators. In a typical troubleshooting task, the goal is to find one or multiple faulty components in a system. Rouse (1981) introduced three training simulators for a troubleshooting task with varying levels of fidelity: low-, mid-, and high-fidelity systems. As shown in Figure 2, the low-fidelity system contained simple abstract components (i.e., circles) and their connections (i.e., arrows). Each component produces a value of 1 if all inputs to the component are 1, and if the component is not faulty; otherwise, the component produces a value of 0 (similar to an AND gate). Trainees were presented with the output of such a system (i.e., components on the right column in Figure 2), they could test the value of connections, and the goal was to find faulty component(s) that were causing 0s in the output. Fewer steps (i.e., tests of connections) in finding the faulty component(s) defined higher performance for trainees.

Rouse introduced another training-system for the same task but with a higher level of fidelity—the mid-fidelity system. In this condition there were two types of components: rectangular and hexagonal as shown in Figure 3. Similar to the low-fidelity system, the rectangular component acted like an AND gate; however, hexagonal components produced a value of 1 when *at least* one input to a hexagonal component was 1, and the hexagonal component itself was not faulty (similar to an OR gate). Moreover, unlike the low-fidelity system, components in this system are not merely feedforward, and outputs of components in a level might be the input to another component in the same or a higher level.

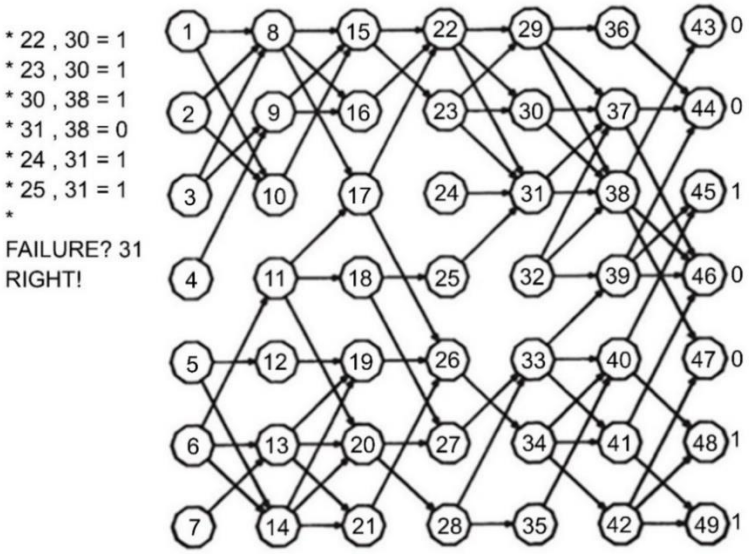


Figure 2. The low-fidelity training simulator of the troubleshooting task in Rouse (1981). The upper left shows the steps a trainee might follow in testing the connections to find a faulty component.

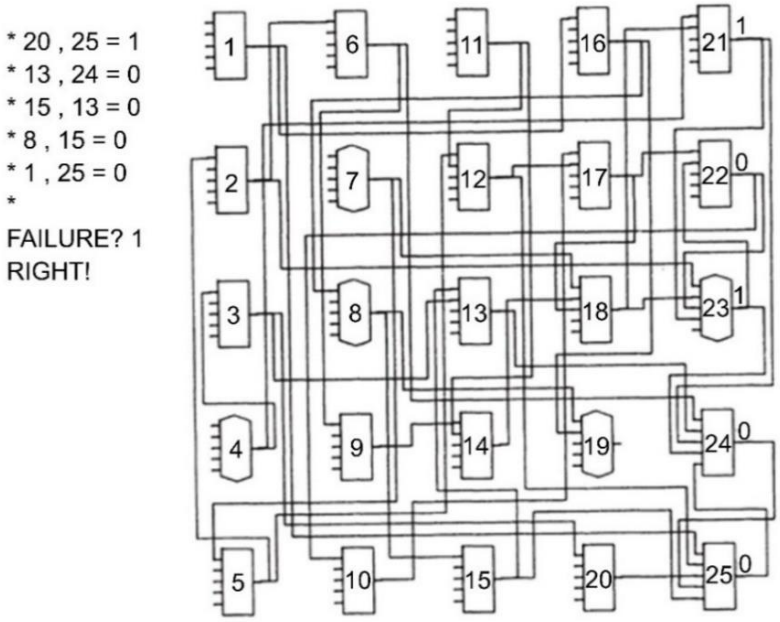
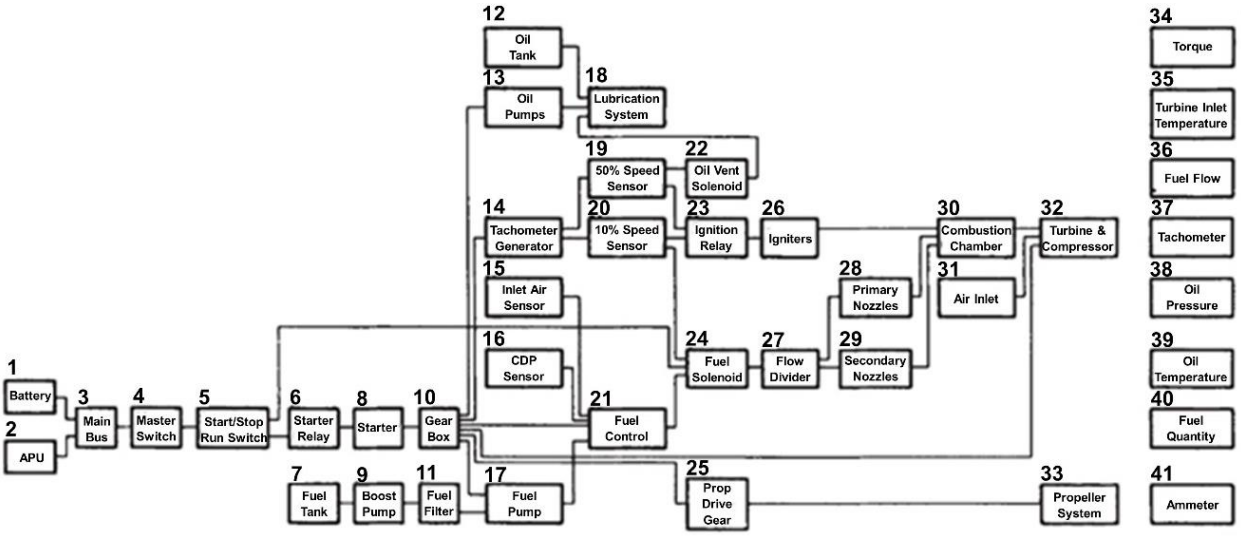


Figure 3. The mid-fidelity training-simulator of the troubleshooting task in Rouse (1981).

Although different in their degree of fidelity, both low- and mid-fidelity systems were task-general and did not require certain knowledge of a specific system for troubleshooting. The high-fidelity training-system, however, used specific knowledge of a realistic troubleshooting task. As Figure 4 shows, participants were presented with the working of a system with various types of components (i.e., Figure 4, top), and the maintenance system was the interface of a control panel software as shown in Figure 4 (bottom). Trainees were first presented with the working state of the system through the control panel, and could gather further information by checking gauges, making observations, or removing components for bench tests; each of these actions was associated with a certain cost. Trainees' overall performance was measured according to the costs associated with their troubleshooting decisions.

In eight experiments, Rouse showed the potentials and benefits of using low- and mid-fidelity simulators. The important contribution of the proposed low- and mid-fidelity systems was that the skills that trainees developed by using those systems improved their performance in a variety of troubleshooting tasks (for a theoretical argument to define troubleshooting on abstract logical inference, see Reiter, 1987). Rouse's findings are contrary to the notion of the traditional theory as they show the benefit of low-fidelity systems that could potentially be effective in a training than can be applied to various tasks (i.e., task-general training). Increasing fidelity increases the specific details in a training, and this narrows the transfer of the skill only to the trained task (for a similar suggestion for pilot training, see Rosa et al., 2020).



System: Turboprop **Symptom: Will not light off**

You have six choices:
1. Observation……OX,Y
2. Information……IX
3. Replace a part……RX
4. Gauge reading……GX
5. Bench test……BX
6. Comparison……CX,Y,Z

(X, Y, and Z are part numbers)

34 Torque	
35 Turbine Inlet Temp	Low
36 Fuel Flow	Low
37 Tachometer	Low
38 Oil Pressure	Normal
39 Oil Temperature	Normal
40 Fuel Quantity	
41 Ammeter	Normal

Your choice …

Actions	Costs	Actions	Costs	Parts Replaced	Costs
4 , 5 Normal	$ 1			14 Tach Generator	$ 199
26 , 30 Abnormal	$ 1				
14 , 20 Not avail	$ 0				
14 Abnormal	$ 27				

Figure 4. The high-fidelity training-simulator of the troubleshooting task in Rouse (1981). Top: the task-specific knowledge of the system that was presented to trainees. Bottom: the maintenance control panel software used in troubleshooting.

In summary, most of the reviewed articles in the domain of maintenance and troubleshooting had a unanimous call to consider training requirements and possibly using low-fidelity systems in training such tasks. One likely reason can be that most tasks in this domain would only need the knowledge of a static state of a system, and as such, the benefits of high-fidelity simulation (e.g., dynamic and motion-based stimuli) are not of much help in, and can even hurt, the training. The main benefit of low-fidelity systems was shown to be in their simplicity for novice trainees and the generality of the acquired skills that could be applied to a variety of tasks.

Aviation and Vehicle Training

Similar to maintenance and troubleshooting, efforts in using inexpensive and low-fidelity systems are almost as old as those of the traditional theory in vehicle training. For example, Denenberg (1954) tested the effectiveness of an inexpensive mock-up trainer for the task of driving the M47 tank, and found it to be more effective in certain aspects of the task (e.g., starting and stopping the tank), and therefore more efficient, than the expensive hull trainer. In another early experiment, Briggs, Fitts, and Bahrick (1957) found no effect of the physical fidelity of manual controlling devices (i.e., force and amplitude of controlling a stick) on transfer in a manual tracking task.

The explosion of the research in aviation training that challenged the traditional theory happened after the mid-1970s (for a review, see Lee, 2005). For example, Hopkins (1975) reviewed the effectiveness of training simulators that were used in the aviation industry and argued against the widespread reliance on motion-based and high-fidelity trainers. Hopkins concluded that the cost-effectiveness of a simulator depends on "(1) your purpose of using it, and (2) your method of using it", even though he did not articulate how these two considerations can determine the technical

specifications of simulators. As another example, in the same study in which Lintern et al. (1990) supported the use of high-fidelity displays in flight training that was discussed earlier, authors found no transfer benefit when they simulated the effect of wind in training. The findings from Lintern et al. (1990) cannot yet lead us to conclude that high-fidelity displays in a simulator without wind would be the most effective and cost-saving combination in flight training-systems. This is because there are still conflicting research reports even from the same researchers.

For example, Lintern, Sheppard, Parker, Yates, and Nolan (1989) manipulated two levels of scene detail and three levels of field of view in training the task of air-to-ground bombing attack. Specifically, the high-fidelity scene detail condition (i.e., "Day Gunnery Scene") contained a multi-colored display showing three-dimensional buildings and guidance lines for the target, while the low-fidelity scene detail condition (i.e., "Dusk Gunnery Scene") contained point lights, running lines, and a concentric set of circles to show the target. The three field-of-view conditions were composed of a small (vertical: –20 to +40 degree; horizontal: –85 to +18 degree), medium (vertical: –16 to +43 degree; horizontal: –95 to +40 degree), and large (vertical: –30 to +50 degree; horizontal: –120 to +40 degree) fields of view from the operating environment of the simulator. The experimental task was a manual delivery of a 25-lb practice bomb from a 30-degree cone pattern (as specified in Naval Air Command's training). Intermediate-level student pilots completed their training sessions in one or two days, and after training, their performance was measured on real bombing tasks by measuring the distance of the bomb impact point from the target. Lintern et al. (1989) found similar transfer effects for the two scene-detail and the three field-of-view levels of the simulation (for similar results, see Westra et al., 1985). This finding is in conflict with the results from Lintern et al. (1990) that found the benefit of using high-fidelity displays in flight training.

In the same line of work, other researchers aimed at showing the benefits of using low-fidelity simulation in flight training. For example, it is well known that a large part of the expertise in controlling airplanes and other critical systems is in showing resilience and adaptive skills in response to system failures and accidents (e.g., Ward, Gore, Hutton, Conway, & Hoffman, 2018). Dahlstrom, Dekker, van Winsen, and Nyce (2009) reviewed a number of flight accidents (e.g., United Airlines 232 in 1989; Pinnacle Airlines 3701 in 2004; Swissair 111 in 1998) in which pilots were following standard procedures to control the airplane under damage but failed in their attempts. Dahlstrom et al. ascribed these accidents mostly to the widely-used pilot training programs in which the emphasis was mostly on creating a photo-realistic simulation of flight environment rather than training resilience skills under stressful situations. The authors argued that the focus on realistic and high-fidelity simulation "may retard or limit the development of skill sets critical for creating safety". Low-fidelity simulations, on the other hand, could thus direct designers' attention and the training toward other important aspects of flight skills such as resilience in crisis management. The reliance on high-fidelity and expensive simulation can have other damaging consequences. For instance, in an empirical study, Stein and Robinski (2012) showed how high-fidelity and motion-based simulators could result in the simulator sickness (e.g., sweating, fatigue, dizziness, and nausea). Although simulator sickness cannot be attributed merely to high-fidelity simulation, it has been often reported in studies that used realistic simulators. For example, the experience of "cybersickness" is reported when using virtual environments (e.g., Kennedy, Lanham, Drexler, Massey, & Lilienthal, 1997).

There have also been a growing number of theoretical papers and literature reviews that argue against the traditional reliance on high-fidelity simulation in aviation. For example, after years of experience in flight simulation research, Roscoe (1991) challenged the widely held reliance on fidelity, and instead, offered measuring training benefits for evaluating simulators. Similarly, Salas,

Cannon-Bowers, and Rhodenizer (1998) questioned the assumption that high-fidelity simulation is necessarily effective in aviation training and proposed to focus on important aspects of training such as trainees' characteristics. Similar messages were given by other reviews and theoretical works such as Noble (2002), Lintern (1991), and Stewart, Johnson, and Howse (2008).

One important result of these efforts in the aviation industry was the recent approval of the United States Federal Aviation Administration (FAA) in using low-fidelity Aviation Training Devices (or ATD, see FAA Advisory Circular No. 61-136B, 2018). These devices, such as Basic Aviation Training Devices(BATD), are different from the traditional high-fidelity flight simulators (known as "full flight simulators", or FFS, and "flight training devices", or FTD) as they are cheaper and often provide trainees with low-fidelity materials for flight training. A list of commercial ATD simulators that were approved by FAA is provided in "FAA Approved Aviation Training Devices" (2019).

Finally, the growing interest in part-task training also helped popularizing the use of low-fidelity simulation for aviation training. For example, Wightman and Lintern (1985) reviewed the evidence in using part-task trainers and showed their effectiveness in training the tasks of tracking and manual control in aviation (see also Crawford et al., 1976; Wightman & Sistrunk, 1987). Although the choice between whole- and part-task training is unresolved and depends on the task and training goals, the possibility of using low-fidelity simulation through part-task training challenged the traditional reliance of the aviation industry on high-fidelity and whole-task trainers (examples of successful use of part-task training in other domains are shown by Spruit et al., 2014; Velotta, 1997; Yuksel et al., 2016). In short, if we aim to train novices, focus on certain aspects of a task, train parts of a task, or reduce the likelihood of simulation sickness, low-fidelity simulation can be effectively used for training in the aviation industry.

Medical Training

In medical training, contrary to the extensive belief in the traditional theory, there is a growing body of evidence that supports the effectiveness of low-fidelity simulation. For example, in training clinical reasoning skills, La Rochelle et al. (2011) recruited 133 medical students who were presented with one of the three types of training materials: paper case, DVD, and live standardized patients (SP). Students in the paper case condition received instructional materials through text (i.e., low-fidelity); those in the DVD condition watched an instructional video of a doctor interviewing and examining a patient (i.e., mid-fidelity); and those in the SP condition were asked to be present in a location where a faculty member interviewed and examined a real SP (i.e., high-fidelity). To increase the empirical rigor, three different subject areas were used: anemia, abdominal pain, and polyuria/diabetes. Each student learned each of the three subject areas only in one level of fidelity, and all students learned all three subject areas (see Figure 5). Authors observed no significant impact of the training format on students' performance, and argued that lowering the fidelity can also decrease the cognitive load for students.

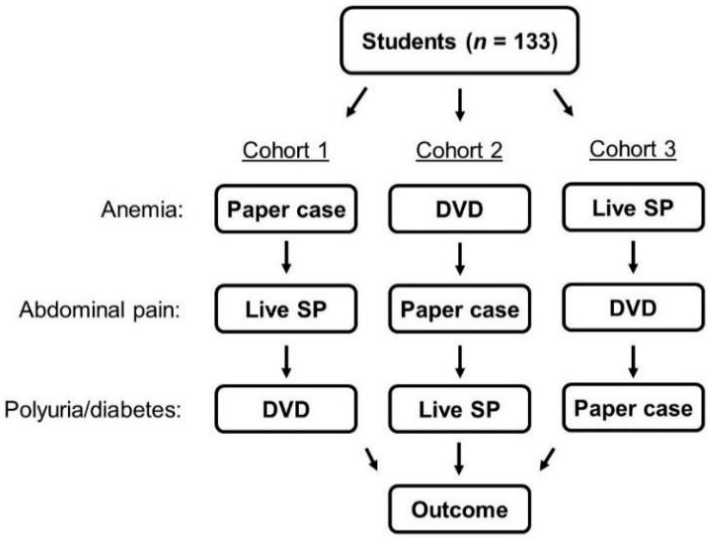


Figure 5. An abstract depiction of the experimental design as used in La Rochelle et al. (2011).

The effectiveness of low-fidelity training simulation is not limited to the skills that depend on declarative knowledge (as in La Rochelle et al.), but is also shown in procedural and more complex manual skills. For example, in a study by Matsumoto, Hamstra, Radomski, and Cusimano (2002), 40 medical students were trained on the manual task of stone-extraction surgery in one of the two conditions of high- and low-fidelity. Students in the high-fidelity condition practiced with complex equipment and materials of surgery, and students in the low-fidelity condition practiced with a mock-up model of surgical equipment such as molded latex, straws, Penrose drain, and a cup. Students in both conditions received the hands-on training under experienced endourologists' supervision. At the end of the training sessions, students' performance was measured on the task of removing a stone from a high-fidelity bench model. Matsumoto et al. found no significant difference in the performance of the two conditions. With considering the cost of the low-fidelity equipment (CA$20) compared to the high-fidelity system (CA$3,700), authors provided compelling evidence to revise the reliance on expensive high-fidelity training materials in procedural tasks in medical training.

In addition to the potentials of using low-fidelity training materials for declarative and procedural tasks, the medical skills that depend on sensory recognition can also benefit from low-fidelity training. As an example, recognizing the regularity of heart sounds and classifying types of irregularities—i.e., auscultation—is an important clinical skill. To train this skill, Harvey®—a $75,000 high-fidelity and life-sized mankin—has been used for decades. In a study by de Giovanni, Roberts, and Norman (2009), 37 medical students were divided between two groups: one received training materials of heart sounds with Harvey® and the other group was trained using recorded sounds via a CD player. Following a six-week interval, trainees' performance in diagnostic accuracy and clinical skills was tested on real patients, and there was no significant difference in clinical or detection skills between the two groups. These findings open the possibility of breaking expensive

and complex whole-task trainers (e.g., Harvey®) into smaller parts, and train each part as a separate task with an inexpensive trainer (e.g., CD player)—that was the idea behind part-task training.

The emergence of empirical investigations was concurrent with a number of reviews and meta-analyses that summarized the effect of fidelity within various areas of medical training. For example, Norman et al. (2012) reviewed 24 studies that compared the effectiveness of low- and high-fidelity simulators for training in three areas of skills: auscultation, surgical, and complex management skills (e.g., critical care such as resuscitation). In each of the three areas, the average training gain in using high-fidelity versus low-fidelity simulation was 2%, 1%, and 1.5% respectively, none of them significant. Norman et al. concluded that for clinical skills that require practice over time (e.g., suturing, measuring blood pressure), low-fidelity simulation can be as effective as—or more effective than—high-fidelity simulation simply because "practice may involve nothing more than mental rehearsal of the steps". Another review with similar results for training laparoscopic surgery skills was reported by Zendejas, Brydges, Hamstra, and Cook (2013) where they reviewed 219 studies and found "in comparison with virtual reality [i.e., high-fidelity training devices], box trainers [i.e., lower-fidelity devices] have similar effects for process skills outcomes and seem to be superior for outcomes of satisfaction and skills time" (see also Moglia et al., 2016).

In summary, in medical training, when the effectiveness of high- and low-fidelity trainers were compared, the results of recent studies often supported low-fidelity systems. The presented results confirmed this trend regarding declarative, procedural, and sensory training. Paying due attention to the potentials of low-fidelity simulation in medical training is crucial to dispel the myth surrounding the necessary effectiveness of high-fidelity simulation (see also, Beaubien & Baker, 2004).

Mixed Reality

Although MR systems have been increasingly used for training in various domains such as industrial assembly (e.g., Yuviler-Gavish et al., 2015) and surgical training (e.g., Seymour et al., 2002), there is a dearth of direct empirical assessments to provide information regarding the proper level of fidelity in such tools (e.g., Dorsey et al., 2009; Moglia et al., 2016). This is because it is widely assumed that MR tools must bring high-fidelity simulation of the environment; therefore, it is difficult to find MR systems that use low-fidelity representations of the environment. Nonetheless, if MR systems are to be used for training, we need to investigate the pros and cons of their high-fidelity simulation.

In an early study on the subject, Kozak, Hancock, Arthur, and Chrysler (1993) compared the effectiveness of a head-mounted VR system in training a simple manual pick-and-place task; they found no significant difference in trainees' performance (i.e., response time) between the VR and the no-training conditions. Although the target task and the characteristics of the VR system might be too distant from today's usage of such systems, Kozak et al. was among the first studies to challenge the growing dependence on MR tools for training. In a more recent study on the subject, to train the manufacturing of an aircraft door, Gonzalez-Franco et al. (2017) recruited 24 participants and equally divided them between two conditions: conventional face-to-face training, and immersive MR training. Following the training session, all subjects were evaluated by tests of knowledge interpretation (i.e., "whether the whole procedure of the assembly was properly captured") and knowledge retention ("multiple-choice format with eight questions"). Participants' performance between two conditions was not different in either of the two criteria of interpretation and retention. Moreover, training duration was higher in the MR condition than the conventional

training. The results in Gonzalez-Franco et al. also questioned the growing reliance of industry on MR systems for training.

In the military, some of the tasks for which training-systems are used are the coordination of infantry soldiers, distance shooting, driving, and precision gunnery; systems that train the skills for those tasks are costly for the military. For example, the acquisition of an M1A2 Abrams tank gunnery trainer can cost more than $127 million in 2015 (GAO Army Training, 2016). Therefore, it is important for the military to find ways to build and use more efficient training-systems. With this goal, Neubauer, Khooshabeh, and Campbell (2017) tested a low-fidelity MR system to train working with tank-commander systems. A tank-commander system allows a commander to communicate to, and coordinate between, several tanks by observing their states, and order movements and shootings. To train the commanding tasks, Neubauer et al. used a simple low-fidelity MR system composed of a personal computer, a joystick, and a head-mounted display. Fifteen participants (consisting of Second Lieutenant trainees, Senior non-commissioned officers, contractors and Captains) tested and evaluated the system. Neubauer et al. found using the simple MR system to be effective in practicing basic communication skills of the tank commanding task. Their system is an initial prototype to replace traditional high-fidelity and expensive tank training simulators.

In addition to experimental studies, there are several review articles investigating the effectiveness of high-fidelity MR systems for training in domains that make extensive use of such systems. For example, in recent years the use of robot-assisted laparoscopic surgery (RAS) is increasing. Because the acquisition of RAS systems for training is costly, virtual reality simulation has been used for training skills that are required in working with RAS systems. In this respect, Moglia et al. (2016) reviewed the effectiveness of MR systems in training RAS skills. Thirty-six studies were reviewed in which five widely-used simulators were used. Moglia et al. reported no

evidence showing the effectiveness of high-fidelity MR training for RAS skills, and subsequently questioned the argument behind the cost-effectiveness of using MR systems in training RAS skills (for similar results in medical training, see also Steigerwald, Park, Hardy, Gillman, & Vergis, 2015; Tan et al., 2012). Moreover, Borsci, Lawson, and Broome (2015) reviewed the use of MR systems in automotive service maintenance training, and showed the lack of concrete empirical evidence to rely on those tools when compared with traditional in-person training techniques. And in a recent interdisciplinary review on the subject, Kavanagh, Luxton-Reilly, Wuensche, and Plimmer (2017) reviewed 99 articles in various domains that used VR tools for decades; they found that such tools are used for reasons that are independent from their transfer success to real-world tasks—e.g., increasing trainees' motivation. Kavanagh et al. concluded that despite the widespread public interest on using such tools, VR systems have been adopted only by a small fraction of training community in certain domains (e.g., general medicine).

Finally, numerous reports and reviews in recent years questioned the use of computer games and simulation in training generalizable skills—the idea that originally stemmed from the reliance of the traditional theory on using computer simulation for training. Although serious games have been used in training (e.g., Williams-Bell et al., 2015) their effectiveness in the transfer of skills from games to other tasks has been brought under doubt in recent years. Specifically, recent research in psychology showed how brain games and similar entertaining products could train skills that were only transferrable to the trained and similar tasks, and not general intelligence or skills (e.g., Sala et al., 2018; Simons et al., 2016). In summary, the benefits of using high-fidelity MR systems and games in training is not sufficiently investigated by the proponents of those systems, and the few studies that did evaluate those systems found no necessary benefits of those tools compared to traditional training techniques and low-fidelity materials.

Conclusion and Summary

It seems that the traditional theory and the belief in the necessary benefits of high-fidelity simulation did not withstand the critical analyses and challenging evidence that subsequent authors presented. The studies mentioned above are among a growing number of reports that are challenging the traditional theory. These studies showed how low-fidelity simulators could surpass high-fidelity simulators in certain aspects of training benefits. In crude terms, in maintenance and troubleshooting the benefit is in the possibility of task-general training, in aviation and vehicle training the benefit is in the simplicity of training materials that could allow novices to focus on important parts of a task, and in medical training the benefit is in the cost-effectiveness of using training devices and improving the performance in various types of skills (i.e., declarative, procedural, and sensory).

There can be various reasons behind the comparative effectiveness of low-fidelity trainers. For example, they can reduce the level of excessive environmental details in the interface of systems, and this opens the possibility of important elements of the task to capture trainees' attention, especially in early stages of training. Moreover, as reviewed earlier, because of the lower level of specificity in low-fidelity simulations, the skills acquired through low-fidelity simulators could be generalized to various tasks, and this was difficult to achieve by high-fidelity simulators. What was shared behind the effective low-fidelity simulators was that they captured those details of tasks that were crucial in training.

Lessons from the Literature

The first lesson from reviewing the literature is that *the central design problem in the three domains of practice is still about fidelity*: what level of fidelity is appropriate for training? Or, in other words:

- what details of the target environment and devices should be simulated? And,
- how realistic should those simulations be for effective training?

If we address the problem of fidelity, we can resolve the central design problem in these domains. Therefore, the fidelity problem can be called *the unified design problem* of training-systems (see also Drews & Bakdash, 2013). In addressing this problem, the second lesson of the review is that *there are potential training benefits to both low- and high-fidelity systems.* Some of the potential benefits of using high-fidelity simulation are:

- lowering the performance time due to the faster recognition of materials in the target environment,
- for some tasks that involve motion, high-fidelity training materials (i.e., motion pictures) can be more effective than low-fidelity materials (i.e., still images),
- if the task requires practicing in certain physical structures, (e.g., egress or vehicle training), high-fidelity tools are likely to be more effective.
- high-fidelity systems provide the opportunity of whole-task trainers where necessary.

And, some of the potential benefits of using low-fidelity simulation are:

- training an abstract (or general) task that can be transferred to specific tasks (e.g., general troubleshooting training),
- simplicity provides the opportunity of focusing on certain training goals (e.g., resilience),
- novices can benefit from starting the practice with simpler systems because of the reduced cognitive load.
- and, in tasks that allow part-task training, using low-fidelity materials can be efficient.

A growing number of the researchers in recent years, especially in the domains of maintenance and aviation, questioned the traditional reliance on high-fidelity systems and advocated the possibility of using low-fidelity simulation for training. The exception was medical training in which—despite the growing number of recent challenging findings—many authors and practitioners still support using high-fidelity systems even without empirical justifications in comparing high- and low-fidelity simulations (e.g., Moglia et al., 2016). Because the nature of the unified design problem is the same across domains, different stances of domains toward the design problem is an evidence showing that these domains are not using each other's knowledge and achievements.

So, the third lesson is that *different domains of practice that use training-systems have much to learn from each other.* Specifically, domains that have more experience in using training simulation (e.g., aviation) can inform other domains with less experience (e.g., medical training) of the possibility of using low-fidelity trainers—an example can be training auscultation skills with low-fidelity materials in de Giovanni et al. (2009). And, domains that are more experienced can also learn from the challenges of less-experienced domains. For example, although MR tools have been used in aviation training in limited scales (e.g., in helicopter training courses), recent research in medical training and manufacturing industry (e.g., Moglia et al., 2016; Webel et al., 2013) can further inform the aviation researchers of the potentials and limitations of MR tools. As Drews and Bakdash (2013) and Roberts et al. (2020) stressed, this emphasizes the importance of cross-disciplinary attempts and reviews in addressing the training-system design (see also Fowlkes et al., 2009).

In addressing the fidelity question, although useful lessons can be learned from the literature, it is obvious that the topic is rife with conflicting evidence and confusions. From one side, there are studies that support high-fidelity simulation, and from another side, opposite arguments are presented in dismissing high-fidelity simulation even for the same tasks. And,

both sides present empirical evidence to support their claims. In addition to the disagreements on fidelity, some studies support task-specific and others support task-general training. It goes without saying that discussions such as the debate between part-task and whole-task training as well as using MR tools suffer from the same dilemma in presenting guidelines to designers and practitioners, and therefore, cannot be of much help in addressing the fidelity question and helping the training-system design.

The traditional theory had the advantage of gathering the researchers around the construct of fidelity, and provided a simple guideline for design: increasing fidelity. However, scientific background and systematic evidence was lacking that should have shown whether and how simulation and surface realism could determine training outcome. As such, in so far as using technologies to build training-systems and motivating designers in their activities, the traditional theory was effective and practical. However, under the shadow of the challenging findings, it has become more and more difficult to rely on the traditional theory in design. The emergence of these challenging findings has made some researchers unwilling to use fidelity altogether in research and discussions on training-systems (e.g., Hamstra et al., 2014) as the concept has faced numerous attempts at redefinition and specification (e.g., Kyaw Tun et al. 2015; Roza, 2005). So, the fourth lesson from the literature is that, *because of the growing number of studies that challenged the traditional theory, fidelity can no longer be the central design construct*. Indeed, if the fidelity cannot determine training outcome, it should not be relied upon as a central design and research construct.

On the other hand, although the emergence of the challenging findings brought about the advantage of increased theoretical and empirical accuracy, these findings have also had their pernicious effects. These challenges are removing the established faith in the traditional theory,

but are not offering alternatives that could guide the design in practice. If the faith in realistic simulation is removed from training-system design, the design question would remain unanswered and there will be no direction to guide the design of training-systems. In this way, the challenging findings have also removed the benefits of the traditional theory without replacements (the benefit of fidelity as an organizing concept was also shown by Hays & Singer, 1989, chap. 3).

The presented challenges to the traditional theory bring the design question on the table again: how to design training-systems? Because of the void coming from the challenges to fidelity, some researchers, designers, and companies remained faithful to the traditional theory and again, focused on and advertised increasing the fidelity in design (e.g., Hambling, 2019; "Saab Receives Order," 2019). The fifth lesson from the literature is that, *if the belief in high-fidelity simulation still exists, it is because fidelity is an easy, well-known, and intuitive construct, and also because there are no existing alternatives to fidelity*. This lack of alternative theory or construct is behind the fact that after nearly five decades of the researchers' position in disagreeing with the traditional theory (e.g., Adams, 1973), the dependence on fidelity and realistic simulation still exists and supported (e.g., Berney, & Bétrancourt, 2016; Höffler, & Leutner, 2007). An illustration of the history of training-system design, as hitherto presented in this article, is shown in Figure 6.

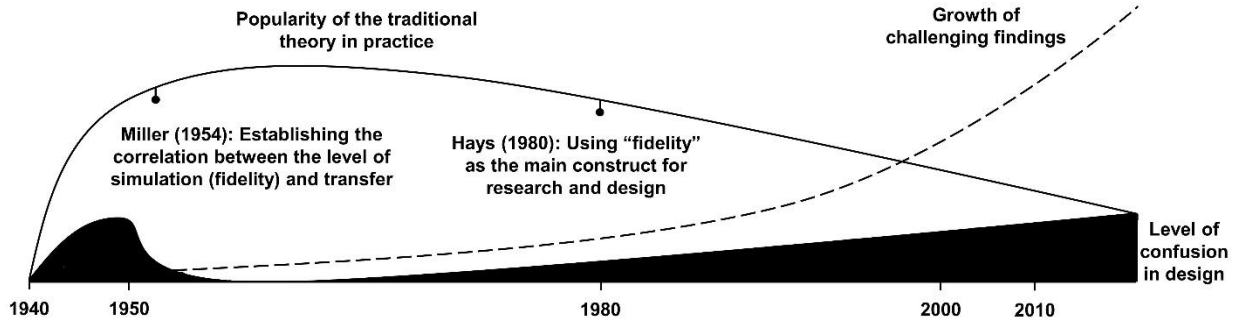


Figure 6. A historical perspective on training-system design as presented in this article. The current state of the field represents a large number of findings that challenge the traditional theory (dashed line), the diminishing popularity of the traditional theory (solid line), and the resulting growth in the confusion for designers (solid dark color).

Despite this short-term remedy in using fidelity, the growing number of challenging findings have made it difficult for practitioners to continue using fidelity in design, improvement, and evaluation of training-systems. As such, it has become difficult for designers to justify their decisions on theoretical and empirical grounds. So, the sixth and concluding lesson is that, *due to the reduced dependence on fidelity by the mainstream research and design community, a conspicuous void exists in providing design guidelines and as a result, confusion in design is the current state of the field*. The increased realism that the modern computational systems—such as MR tools—can provide is not only of little help in building more effective training-systems, but can also pose potential risks such as simulation sickness and the illusions of embodiment (Kennedy et al., 1997; Stein, & Robinski, 2012). The field needs systematic guidelines that can direct the research and design (Fowlkes et al., 2009; Roberts et al., 2020), while organizations are always asking for more effective training-systems (Salas et al., 2012). The next section presents a proposal aiming at filling the void of fidelity and changing the status quo.

Toward Creating Design Methods

Earlier, we concluded that the reason behind the success of the reviewed low-fidelity simulators was that they captured the important elements of a task. The key question is: how can we know what elements of a task are important for training? This section briefly discusses the potentials of new resources for design. In filling the void of fidelity, we need methods that can provide us with specific design guidelines. In creating design methods, one lesson that is clear from the presented review is that instead of simulation technology, we should focus on training elements such as experts, trainees, the task, tools, and other important elements in training. Considering these elements would guide us on how to use the technology in simulating target environments. The following paragraphs will first discuss how past research that focused on trainees- and task-analysis can be used as resources for creating design methods. Later, a discussion of the deficiency of past research is presented that would necessitate the need to consider new training elements as resources of design guidelines.

Information from Past Literature

As reviewed earlier, for years researchers were aware of the problems in depending on fidelity in design. As a result, they approached the design problem with a training-centered approach that focused on two key elements of training: the task and trainees. Here, a description of the emphasis of the past research on these two elements is briefly presented.

Task-analysis. With respect to considering the task in design, a training-system is designed to train a task, and so, we need to know about the task that should be trained. To learn about tasks, standard methods of work/task-analysis have been used for decades (Smith, 1965; Stanton, Salmon, Walker, Baber, & Jenkins, 2018). The importance of such analyses in technical specifications of training-systems is well-established, as researchers showed how the knowledge

from the task-analysis can inform designers of the required fidelity level and other aspects of the training devices and programs (e.g., Hays & Singer, 1989; Smode, 1971, 1972).

For example, one technique that can particularly be useful in design is Mission Essential Competency (MEC) as discussed in Alliger, Beard, Bennett, Colegrove, & Garrity (2007). The MEC systematically analyzes a task by focusing on the aspects of the task that are essential for a specific goal or mission. As such, the MEC is a goal-oriented task-analysis technique. Because the MEC considers both tasks and users, it also provides information on what are needed from users' perspective. This is particularly useful in training-system design as it can help both in determining the requisite characteristics from trainees for a task, and in developing those characteristics. Therefore, the MEC can inform designers of what should be trained in a task. Although the MEC has been developed within the military context, it has the potential to provide important information for other domains such as medical training simulation. Similarly, extensive efforts have been made to develop other task-analysis techniques that are useful in training-system design, and so, past research can provide useful information regarding how to gather information about the tasks that should be trained (e.g., Annett & Duncan, 1967; Jonassen, Tessmer, & Hannum, 1999; Schraagen, Chipman, & Shalin, 2000).

Trainees-analysis. In addition to the task, important lessons can be learned from the past literature regarding considering trainees and their needs in designing training-systems and programs. For example, we now know that training-systems with higher levels of fidelity are more likely to be useful for more advanced trainees, and those with little to no prior skill can often benefit more from low-fidelity simulation (e.g., Alessi, 1988). In another example, Wulf and Lewthwaite (2016) proposed a theory of motor learning in which trainees' motivation and attention are in the focus (to review the effect of trainees' motivation and self-efficacy on

training outcome, see also Colquitt, LePine, & Noe, 2000). Salas et al. (2012) also presented an analysis on the influential factors on training among which various trainees' characteristics are prominent (see also Salas et al., 2006). Similarly, in a well-known line of research that considered trainees' needs, Cognitive Load Theory tells us to include only necessary details of the target devices and environment in the simulation, as extraneous details can consume trainees' cognitive resources (Sweller et al., 1998). Decades of research by educational scientists on designing instructional materials can also provide important information in considering trainees' characteristics (e.g., Kalyuga, 2007; Mayer, 2009).

Considerations of trainees' characteristics in the design of instructional materials are further discussed in educational sciences within subjects such as instructional scaffolding (e.g., Puntambekar & Hubscher, 2005; Wood, Bruner, & Ross, 1976) and Analysis, Design, Development, Implementation and Evaluation (ADDIE) model (e.g., Larson & Lockee, 2013). These resources from educational sciences can be used for our purpose in training-system design. For example, we can use the stepwise and feedback-oriented model of building instructions from the ADDIE model that re-evaluates the design after completion. Similarly, the structured questions in the Analysis phase of the ADDIE model can direct us in gathering information about training needs (see also Allen, 2006; Kalyuga, 2007).

Deficiencies of task- and trainees-analysis. Although these efforts are helpful in creating design methods for training-systems, the main concern of researchers in educational sciences is developing instructional materials that are mostly used in formal educational settings. On the other hand, our concern is mostly technical skill acquisition with the focus on designing training simulators. This distinction between knowledge, learning, and formal education in one hand, and skill, training, and practice in the other hand precludes the research in educational

sciences to resolve central design problems of training-systems (for a more detailed discussion on the topic, see Hays & Singer, 1989, pp. 293–297). As such, the lessons from educational sciences can help but cannot provide specific guidelines that can be used in the design practice.

In a broader sense, although considering the two elements of task and trainees can provide important information for training-system design, it is not sufficient in creating design methods. The presented resources and techniques (e.g., MEC) mostly tell us what to train, but not how to train them via simulation. In fact, if we consider the training cycle, trainees and the task are only one side of the training cycle and as such, they are not sufficient in guiding the design (e.g., Adams, 1979). This deficiency is evident in considering the studies that focused only on trainees and task for training-system design.

For example, Sticha et al. (1990) created a model (Optimization of Simulation-Based Training-systems, or OSBATS) to produce design methods for developing cost-efficient training-systems, and authors tested their model to address design decisions of the training-systems that were used in the military (e.g., Army rotary-wing aviation tasks, troubleshooting M1 Abrams tank). The OSBATS model used device- and task-analysis to determine the elements of the target system that could present designers with the choice of varying degrees of fidelity. Although much can be learned from their modeling approach, due to the excessive use of task- and device-specific knowledge, Sticha et al. recognized the problem of OSBATS in generalizing it to various tasks. For example, fidelity specifications of the model would vary significantly between tasks and domains, and it is not clear how the model or the designer should make technical decisions for each task and domain. These problems of the OSBATS approach in generalizability to other tasks and domains were stressed by subsequent studies from the same line of research (e.g., Willis, Guha, Hunter, & Singer, 1990).

Similarly, Smode (1971, 1972) aimed to specify the technical requirements of training devices through a standardized design method that was mostly based on needs- and task-analysis. Although Smode's and similar analyses are helpful and informative, they suffer from the lack of specificity in addressing design decisions (see Hays and Singer, 1989). And in an attempt at proposing a developmental strategy for training-system design, Cream, Eggemeier, and Klein, (1978) provided suggestions for designers that, albeit informative, does not exceed what was known in trainees- and task-analysis (see also Salas et al., 2006).

In short, the aforementioned attempts, when considered together, do not present a comprehensive resource to guide the design in specific and practical sense. As Hays and Singer (1989, p. 56) wrote, "no systematic guidance exists to translate task analysis information into a form which can facilitate fidelity decisions". The outstanding evidence of the deficiency of the decades of past attempts in trainees- and task-analysis is in the current confusion in design and problems with fidelity. This deficiency can also be found in the lack of attention from practitioners to research community and the long-standing gap that stands between theorical research and design practice (Campbell, 1971; Cannon-Bowers et al., 1991; Goldstein, 1978).

To put the problem in perspective, training-system design needs to consider important elements of training. Trainees and the task are only two elements in training. Because other important elements have not been considered in detail, most recommendations from the past research are insufficient for creating practical design methods. The following paragraphs discuss studying expert performance as an important resource for design that has the potential to resolve many of the current confusions in the field.

Studying Expert Performance

The key lesson from the literature is that successful training-systems embody the important elements of a task, and the subsequent question was in how we should know what elements of a task are important in training. The main goal of training is to transform trainees to perform as good as people who have high levels of skills, or experts. Training should thus be a direction toward expert performance, and as such, skillful performance should be the goal of practicing with training-systems. Therefore, studying expert performance is important, if not crucial, for training-system design. From one side, the designer of training-systems is often a design engineer who is not a subject-matter expert and does not have the experience of the task performance (e.g., Hays & Singer, 1989). From another side, even experts in a task might be unaware of the details of how they perform the task (e.g., Robinson, 1974; Rouse & Morris, 1986). Studying expert performance by the designer or an independent observer is therefore necessary to create a realistic and objective account of expert performance that is needed in design. Without knowing about the characteristics of expert performance, it would be difficult both to design and to evaluate training-systems.

How to study and characterize expert performance? Although this question was studied before (e.g., Ericsson, 2018; Ericsson & Simon, 1980; Lintern, Moon, Klien, & Hoffman, 2018), we need to make modifications to previous methods for our purpose in training-system design. In crude terms, we should first identify individual experts in the task for which we are developing a training-system. Depending on the task and the availability of experts, multiple experts should be recruited, as studying more experts can increase the reliability of our characterization of the expertise. Later, we need to study their performance by videorecording their behaviors when working with tools and systems in the target environment. Although researchers might use

various strategies in studying and eliciting the expertise (e.g., Rosa et al., 2020), the videorecording in target environments provides the maximum information—in the form of a direct and realistic picture—of expert performance. After recording their behaviors, we should analyze the recordings to create a model of expert performance (see Lintern et al., 2018). In other words, we need to derive the shared behaviors among all (or the majority) of the experts that we recorded, and represent those behaviors as the model of expert performance.

Now we should use this model in training-system design. The model of expert performance can be used in structuring the training-system and making design decisions. In fact the focus of a training-system should be on training those behaviors, and in doing so, the characteristics of expert performance can either shape the overall structure of a training-system or inform specific design decisions for improving existing systems. For example, one of the questions of the unified design problem that we reviewed earlier was “what details of the target environment should be included in the simulation?”. The model of expert performance can inform us of the frequency and importance of using tools and devices of a system.

As another example, a designer of a flight training simulator might ask if it is necessary to include a certain control panel in the simulator that would increase the fidelity of simulation. In addition to the analysis of the specific task as the goal of the training, we need to ask what was the frequency and importance of using that control panel in experts’ performance? Does it matter in training how that control panel “looks” and “feels” like to experts? These two examples show how the model of expert performance can inform designers of technical requirements of the simulation, including the appropriate level of fidelity.

Unfortunately, the importance of expert performance has largely been ignored in the research and design community. From one side, studying experts has been the subject of

considerable research in psychology (e.g., de Groot, 1965; Ericsson, & Charness, 1994). From the other side, there have been only scant efforts in using the knowledge of expert performance for training-system design (e.g., Willis et al., 1990). For example, Ward, William, and Hancock (2006) discussed the two issues of expert performance and training simulation almost independently, disregarding the potentials that studying expert performance has in resolving design decisions of training-systems. Similarly, when discussing "Person Analysis" for training design, Salas et al. (2006) merely identify trainees as "Person" (pp. 477–478) and repeat the well-known emphasis on trainees-analysis. It seems necessary to establish experts as an important human element of the training cycle whose performance should be the direction and goal of training programs and systems.

One reason behind paying insufficient attention to experts is the notion that because most experts are not training specialists, they should not be directly involved in the design process. It is argued that experts do not often consider trainees' characteristics and they would often suggest using high-fidelity training simulation with unjustified reasons (e.g., Hays & Singer, 1989, p. 36; Smode, 1971). Nonetheless, using experts should not be limited to asking their opinion in evaluating the design process or the final products of training-systems. Rather, we can study their performance and use the knowledge in training-system design. Expert individuals experienced prolonged practice, and this experience is often represented in how they perform the task, not necessarily in their judgments (Robinson, 1974; Rouse & Morris, 1986).

Although using the knowledge from expert performance in training-system design is not well-established, we can use past research and existing tools that emerged from other related disciplines. For example, in addition to the studies that were mentioned earlier, to derive the shared behavioral characteristics among the recordings of experts' performances, we can use

computational video analysis tools (e.g., U.S. Patent No. 10,482,613, 2019). Moreover, the proposed method of studying expert performance in target environments reminds us of past attempts such as studying naturalistic decision making (Klein, 2008), verbal protocol analysis (Ericsson & Simon, 1980), and ethnographic studies of work (Clancey, 2006; Luff, Hindmarsh, & Heath, 2000) in which researchers tried to study individuals during task performance and in natural settings. Although those attempts were not directed toward addressing the training-system design problems, we can use their experience in studying experts, characterizing expertise, and using them in design. A method of video recording expert performance for training-system design with the aforementioned considerations has recently been under development (Doozandeh, 2020).

Conclusions

Three decades have passed since the publication of Hays and Singer's (1989) book on the topic of simulation fidelity in training-system design. Although this review is a continuation of Hays and Singer's critical position on fidelity, much have changed over the last three decades. The introduction of new training devices (e.g., MR systems), the growing use of training simulation in general and in certain domains (e.g., healthcare), and the recent findings in psychology are among the factors that necessitated a renewal of attention on training-system design. This article reviewed the research on the topic and found that the "fidelity question"—as coined by Hays and Singer—is still unresolved; it is the central design problem that unifies various domains that use training-systems. The literature showed the impossibility of making a universal conclusion of what level of fidelity is appropriate for training simulation, and as such, fidelity cannot be a reliable design construct.

Notwithstanding the growth and availability of simulation technology, we should remind ourselves that the primary goal of training-systems should not be to simulate, but to train. The growing disbelief in fidelity and the void in providing design guidelines should make us consider new resources that have the potential to guide the design of future training-systems. The question of whether to choose high- or low-fidelity training materials should not itself be the subject of investigation, because the answer depends. It depends on training factors such as the task, trainees, and expert performance, and we need to create design methods based on these factors.

In developing design methods, we can use the findings of the past literature that can provide useful information regarding important elements of training. Nonetheless, the past literature is mostly focused on trainees- and task-analysis, and is deficient in providing specific design guidelines. Expert performance as a potential resource has hitherto been largely overlooked. Studying experts by recording their behaviors and modeling their performance compliments the knowledge of trainees and the task, as it provides the direction and the goal for training-systems. By combining the existing knowledge in trainees- and task-analysis with studying expert performance, we can create design methods that can address the long-standing problems in training-system design, including fidelity specifications.

This approach to design, as well as the presented review, are not without limitations. First, the presented review is not using a systematic method of review (i.e., meta-analysis), and as such, we do not have effect sizes and other technical details to compare the effectiveness of simulators. This is due to the nature of this inter-disciplinary investigation and the complexities in measuring constructs such as fidelity and transfer. Such complexities, however, should not make us reluctant in making investigations on the topic, and as presented, practical lessons were gained from the narrative review.

Second, the proposed approach to design is arguing for the use videorecording in capturing experts' skills, and this can open various criticisms. For example, it is argued that videorecording is limited to certain tasks that have external representations of performance (e.g., Lintern et al., 2018). For instance, mental skills cannot be captured in video. Although this is a limitation, we should consider that various elements of performance in mental tasks—such as implementing the decisions, as well as the activities and processes during problem-solving—can still be captured via video, and these elements can have training values. Additionally, there are still many tasks and activities that can use the benefits of videorecording (e.g., flight, driving, military skills). And as another example of limitation, because the proposed approach is using the performance of multiple experts for modeling the performance, it might be argued that individual differences in performance might pose difficulty in creating the performance model (e.g., Ackerman, 1988). This is again a limitation; however, this would highly depend on the task. Based on our experience, there are numerous shared behavioral characteristics that can be found in various tasks (e.g., truck driving). Overall, we believe that the benefits and limitations of the proposed approach should be assessed only after the creation of design methods which is a central message of this article.

The call is to create informed design methods with considering important elements of training such as expert performance and trainees- and task-analysis. Although this approach might seem costly at first, spending sufficient resources to the design can prevent the expenses, and potential dangers, of relying on high-fidelity simulation in the long run. To build effective training-systems, we need to integrate the findings from various domains as it helps us defining training-system design as a holistic practice in which researchers can learn from the experiences and challenges of each other and share their findings.

References

References marked with asterisks indicate studies included in the review section that compared between the traditional theory and challenging findings.

* Abrahamson, S., Denson, J. S., & Wolf, R. M. (1969). Effectiveness of a simulator in training anesthesiology residents. *Journal of Medical Education, 44*, 515–9.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117,* 288–318.

Adams. J. A. (1960). Part trainers. In G. Finch (Ed.), *Educational and training media: A symposium* (Publication 789). Washington, DC: National Academy of Science, National Research Council.

Adams, J. A. (Ed.). (1973). Flight simulation [Special issue]. *Human Factors, 15*(6).

Adams, J. A. (1979). On the evaluation of training devices. *Human Factors, 21*(6), 711–720.

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

Alessi, S. M. (1988). Fidelity in the design of instructional simulations. *Journal of Computer-Based Instruction, 15*(2), 40–47.

Allen, W. C. (2006). Overview and evolution of the ADDIE training system. *Advances in Developing Human Resources, 8,* 430–441.

* Allen, J. A., Hays, R. T., & Buffardi, L. C. (1986). Maintenance training simulator fidelity and individual differences in transfer of training. *Human Factors, 28*(5), 497–509.

* Allen, R. W., Park, G. D., & Cook, M. L. (2010). Simulator fidelity and validity in a transfer-of-training context. *Transportation Research Record: Journal of the Transportation Research Board, 2185*, 40–47.

Alliger, G. M., Beard, R., Bennett Jr, W., Colegrove, C. M., & Garrity, M. (2007). *Understanding mission essential competencies as a work analysis method* (Report No. AFRL-HE-AZ-TR-2007-0034). Mesa, AZ: Air Force Research Laboratory.

Alluisi, E. A. (1991). The development of technology for collective training: SIMNET, a case history. *Human Factors, 33*(3), 343–362.

* Anastassova, M., & Burkhardt, J.-M. (2009). Automotive technicians' training as a community-of-practice: Implications for the design of an augmented reality teaching aid. *Applied Ergonomics, 40*(4), 713–721.

Annett, J., & Duncan, K. D. (1967). Task analysis and training design. *Occupational Psychology 41*, 211–221.

* Beaubien, J. M., & Baker, D. P. (2004). The use of simulation for training teamwork skills in health care: How low can you go? *Quality and Safety in Health Care, 13*(1), 51–56.

* Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin, 144*(1), 77–110.

Berney, S., & Bétrancourt, M. (2016). Does animation enhance learning? A meta-analysis. *Computers & Education, 101*, 150–167.

* Borsci, S., Lawson, G., & Broome, S. (2015). Empirical evidence, evaluation criteria and challenges for the effectiveness of virtual and mixed reality tools for training operators of car service maintenance. *Computers in Industry, 67*, 17–26.

* Briggs. G. E., Fitts, P. M., & Bahrick, H. P. (1957). Effects of force and amplitude cues on learning and performance in a complex tracking task. *Journal of Experimental Psychology, 54*(4), 262–268.

* Buchanan, J. A. (2001). Use of simulation technology in dental education. *Journal of Dental Education, 65*(11), 1225–1230.

* 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.

Campbell, J. P. (1971). Personnel training and development. *Annual Review of Psychology, 22*, 565–602.

Cannon-Bowers, J. A., Tannenbaum, S. I., Salas, E., & Converse, S. A. (1991). Toward an integration of training theory and technique. *Human Factors, 33*(3), 281–292.

* Cant, R. P., & Cooper, S. J. (2017). Use of simulation-based learning in undergraduate nurse education: An umbrella systematic review. *Nurse Education Today, 49*, 63–71.

* Caro, P. (1988). Flight training and simulation. In E. Weiner & D. Nagel (Eds.), *Human Factors in Aviation* (pp 229–261). San Diego, CA: Academic.

Clancey, W. J. 2006. Observation of work practices in natural settings. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 127–146). Cambridge, UK: Cambridge University Press.

Colquitt, J., LePine, J., & Noe, R. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85*, 678–707.

* Cook, D. A., Hatala, R., Brydges, R., Zendejas, B., Szostek, J. H., Wang, A. T., Erwin, P., & Hamstra, S. (2011). Technology-Enhanced simulation for health professions education: A systematic review and meta-analysis. *Journal of American Medical Association, 306,* 978–988.

Cooley, T., Seavers, G., Gordon, S., Roth, J., & Rodriguez, J. (2015). Calculating simulation-based training value: Cost avoidance and proficiency. In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference*. Orlando, FL.

* Crawford, A. M., Hurlock, R. E., Padilla, R., & Sassano, A. (1976). *Low cost part-task training using interactive computer graphics for simulation of operational equipment* (No. NPRDC-TR-76TQ-46). San Diego, CA: Navy Personnel Research and Development Center.

Cream, B. W., Eggemeier, F. T., & Klein, G. A. (1978). A strategy for the development of training devices. *Human Factors, 20*(2), 145–158.

* Dahlstrom, N., Dekker, S., van Winsen, R., & Nyce, J. (2009). Fidelity and validity of simulator training. *Theoretical Issues in Ergonomics Science, 10*(4), 305–314.

* De Crescenzio, F., Fantini, M., Persiani, F., Di Stefano, L., Azzari, P., & Salti, S. (2011). Augmented reality for aircraft maintenance training and operations support. *IEEE Computer Graphics and Applications, 31*(1), 96–101.

* de Giovanni, D., Roberts, T., & Norman, G. (2009). Relative effectiveness of high- versus low-fidelity simulation in learning heart sounds. *Medical Education, 43*(7), 661–668.

de Groot, A. (1965). *Thought and choice in chess*. The Hague, Netherlands: Mouton.

* Denenberg, V. H. (1954). *The training effectiveness of a tank hull trainer* (Technical Report HUMRRO-TR-3). Alexandria, VA: Human Resources Research Office, George Washington University.

* Dennis, K. A., & Harris, D. (1998). Computer-based simulation as an adjunct to Ab Initio flight training. *The International Journal of Aviation Psychology, 8,* 261–277.

Doozandeh, P. (2020). *Videorecording of experts as a method of training-simulator design* [Manuscript submitted for publication]. College of Information Sciences and Technology, The Pennsylvania State University.

Dorsey, D., Campbell, G., & Russell, S. (2009). Adopting the instructional science paradigm to encompass training in virtual environments. *Theoretical Issues in Ergonomics Science, 10*(3), 197–215,

* Dougherty, D. J., Houston, R. C., & Nicklas, D. R. (1957). *Transfer of training in flight procedures from selected ground training devices to the aircraft.* (Tech. Rep. NAVTRADEVCEN-TR-71-16-16). Port Washington, NY: Office of Naval Research, Naval Training Device Center.

Drews, F. A., & Bakdash, J. Z. (2013). Simulation training in health care. *Reviews of Human Factors and Ergonomics, 8*(1), 191–234.

* Dunston, P. S., Proctor, R. W., & Wang, X. (2014). Challenges in evaluating skill transfer from construction equipment simulators. *Theoretical Issues in Ergonomics Science, 15*(4), 354–375.

Endsley, M. R. (2018). Expertise and situation awareness. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *Cambridge handbook of expertise and expert performance.* Cambridge, UK: Cambridge University Press.

Ericsson, K. A. (2018). Capturing expert thought with protocol analysis: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *Cambridge handbook of expertise and expert performance.* Cambridge, UK: Cambridge University Press.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*(8), 725–747.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251.

FAA Advisory Circular No. 61-136B (2018). *FAA Approval of Aviation Training Devices and Their Use for Training and Experience*. Retrieved from https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_61-136B.pdf

FAA Approved Aviation Training Devices (August 2019). *U.S. Department of Transportation, Federal Aviation Administration: Airmen Training and Certification Branch*. Retrieved from https://www.faa.gov/about/office_org/headquarters_offices/avs/offices/afx/afs/afs800/afs810/media/FAA_Approved_Airplane_ATDs.pdf

* Fink, C., & Shriver, E. (1978). *Simulators for maintenance training: Some issues, problems and areas for future research* (Rep. No. AFHRL-TR-78-27). Lowery Air Force Base, CO: Air Force Human Resources Laboratory.

Fowlkes, J. E., Neville, K. J., Owens, J. M., & Hafich, A. J. (2009). Challenges to the development of pedagogically driven engineering requirements for complex training systems. *Theoretical Issues in Ergonomics Science, 10*(3), 217–229.

* Freitas, S., & Liarokapis, F. (2011). Serious games: A new paradigm for education? In M. Ma, A. Oikonomou, & L. C. Jain (Eds.), *Serious Games and Edutainment Applications* (pp. 9–23). London, UK: Springer.

Gagne, R. M. (1954). Training devices and simulators: Some research issues. *American Psychologist, 9*(3), 95–107.

G. A. O. Army Training (2016). *Efforts to adjust training requirements should consider the use of virtual training devices* (Technical Report GAO-16-636). Government Accountability Office: Washington, DC. Retrieved from https://www.gao.gov/assets/680/679104.pdf

* Gerathewohl, S. J. (1969). *Fidelity of simulation and transfer of training: A review of the problem* (Report AM 69-24). Washington, DC: Department of transportation, Federal Aviation Administration, Office of Aviation Medicine.

Goldstein, I. L. (1978). The pursuit of validity in the evaluation of training programs. *Human Factors, 20*(2), 131–144.

* Gonzalez-Franco, M., Pizarro, R., Cermeron, J., Li, K., Thorn, J., Hutabarat, W., Tiwari, A., & Bermell-Garcia1, P. (2017). Immersive mixed reality for manufacturing training. *Frontiers in Robotics and AI, 4*(3), 1–8.

* Gopher, D., Weil, M., & Bareket, T. (1994). Transfer of skill from a computer game trainer to flight. *Human Factors, 36*, 387–405.

* Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science, 18*(1), 88–94.

* Green, M., Tariq, R., & Green, P. (2016). Improving patient safety through simulation training in anesthesiology: Where are we? *Anesthesiology Research and Practice, 2016*, 1–12.

Hale, K. S., Stanney, K. M., Milham, L. M., Bell Carroll, M. A., & Jones, D. L. (2009). Multimodal sensory information requirements for enhancing situation awareness and training effectiveness. *Theoretical Issues in Ergonomics Science, 10*(3), 245–266.

Hambling, D. (2019). The sciences of simulation: A primer on the visual, motion, audible and haptic cues necessary for immersive environments. *Military Simulation & Training Magazine, 3*, 6–11.

* Hammerton, M., & Tickner, A. H. (1967). Visual factors affecting transfer of training from a simulated to a real control situation. *Journal of Applied Psychology, 15*(1), 46–49.

Hamstra, S. J., Brydges, R., Hatala, R., Zendejas, B., & Cook, D. A. (2014). Reconsidering fidelity in simulation-based training. *Academic Medicine, 89*(3), 387–392.

Hays, R. T. (1980). *Simulator fidelity: A concept paper* (Report No. 490). Alexandria, VA: Army Research Institute for the Behavioral and Social Sciences.

Hays, R. T., & Singer, M. J. (1989). *Simulation fidelity in training-system design: Bridging the gap between reality and training*. New York, NY: Springer Science & Business Media.

Höffler, T. N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction, 17*(6), 722–738.

* Hopkins, C. O. (1975). How much should you pay for that box? *Human Factors, 17*(6), 533–541.

* Hughes, C. E., Stapleton, C. B., Hughes, D. E., & Smith, E. M. (2005). Mixed reality in education, entertainment, and training. *IEEE computer graphics and applications, 25*(6), 24–30.

* Jentsch, F., & Bowers, C. A. (1998). Evidence for the validity of PC-based simulations in studying aircrew coordination. *International Journal of Aviation Psychology, 8*, 243–260.

* Johnson, W. B., & Rouse, W. B. (1982). Training maintenance technicians for troubleshooting: Two experiments with computer simulations. *Human Factors, 24*(3), 271–276.

Jonassen, D. H., Tessmer, M., & Hannum, W. H. (1999). *Task analysis methods for instructional design*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509–539.

* 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.

Kennedy, R. S., Lanham, D. S., Drexler, J. M., Massey, C. J., & Lilienthal, M. G. (1997). A comparison of cybersickness incidences, symptom profiles, measurement techniques, and suggestions for further research. *Presence: Teleoperators & Virtual Environments, 6*(6), 638–644.

Kirkpatrick, D. L., & Kirkpatrick J. D. (2006). *Evaluating training programs*. San Francisco, CA: Berrett-Koehler.

Klein, G. (2008). Naturalistic decision making. *Human Factors, 50*(3), 456–460.

Koonce, J. M. (1984). A brief history of aviation psychology. *Human Factors, 26*(5), 499–508.

* Kozak, J. J., Hancock, P. A., Arthur, E. J., & Chrysler, S. T. (1993). Transfer of training from virtual reality. *Ergonomics, 36*(7), 777–784.

Kyaw Tun, J., Alinier, G., Tang, J., & Kneebone, R. (2015). Redefining simulation fidelity for healthcare education. *Simulation & Gaming, 46*, 159–174.

* La Rochelle, J. S., Durning, S. J., Pangaro, L. N., Artino, A. R., van der Vleuten, C. P., & Schuwirth, L. (2011). Authenticity of instruction and student performance: A prospective randomised trial. *Medical Education, 45*, 807–817.

* Laner, S. (1954). The impact of visual aid displays showing a manipulative task. *Quarterly Journal of Experimental Psychology, 6*(3), 95–106.

* Langley, A., Lawson, G., Hermawati, S., D'Cruz, M., Apold, J., Arlt, F., & Mura, K. (2016). Establishing the usability of a virtual training system for assembly operations within the automotive industry. *Human Factors and Ergonomics in Manufacturing & Service Industries, 26*(6), 667–679.

Larson, M., & Lockee, B. B. (2013). *Streamlined ID: A practical guide to instructional design*. New York, NY: Routledge.

Lee, A. T. (2005). *Flight simulation: Virtual environments in aviation.* New York, NY: Routledge.

* Lee, A. T. (2018). *Vehicle simulation: Perceptual fidelity in the design of virtual environments.* Boca Raton, FL: CRC Press.

* Lewis, R., Strachan, A., & Smith, M. M. (2012). Is high fidelity simulation the most effective method for the development of non-technical skills in nursing? A review of the current evidence. *Open Nursing Journal, 6*, 82–89.

Link, J. E. A. (1937). *U.S. Patent No. 2,099,857*. Washington, DC: U.S. Patent and Trademark Office.

Lintern, G. (1991). An informational perspective on skill transfer in human-machine systems. *Human Factors, 33*(3), 251–266.

Lintern, G., Moon. B., Klein. G., & Hoffman. R. R. (2018). Eliciting and representing the knowledge of experts. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, M. Williams (Eds). *The Cambridge handbook of expertise and expert performance*. New York, NY: Cambridge University Press.

* Lintern, G., Roscoe, S. N., & Sivier, J. E. (1990). Display principles, control dynamics, and environmental factors in pilot training and transfer. *Human Factors, 32*(3), 299–317.

* Lintern, G., Sheppard, D. J., Parker, D. L., Yates, K. E., & Nolan, M. D. (1989). Simulator design and instructional features for air-to-ground attack: A transfer study. *Human Factors, 31*(1), 87–99.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*(4), 492–527.

* Lopreiato, J. O., & Sawyer, T. (2015). Simulation-based medical education in pediatrics. *Academic Pediatrics, 15*(2), 134–142.

Luff, P., Hindmarsh, J., & Heath, C. (2000). *Workplace studies: Recovering work practice and informing system design*. Cambridge, UK: Cambridge University Press.

* MacLean, S., Geddes, F., Kelly, M., & Della, P. (2019). Realism and presence in simulation: Nursing student perceptions and learning outcomes. *Journal of Nursing Education, 58*(6), 330–338.

McGeoch, G. O. (1931). Whole-part problem. *Psychological Bulletin, 28*, 713–739.

* Matsumoto, E. D., Hamstra, S. J., Radomski, S. B., & Cusimano, M. D. (2002). The effect of bench model fidelity on endourological skills: A randomized controlled study. *The Journal of Urology, 167*, 1243–1247.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York, NY: Cambridge University Press.

* Miller, R. B. (1954). *Psychological considerations in the design of training equipment* (WADC Report No. 54-563, AD 71202). Springfield, OH: Carpenter Litho & Prtg. Co.

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

* Nathanael, D., Mosialos, S., Vosniakos, G. C., & Tsagkas, V. (2016). Development and evaluation of a virtual reality training system based on cognitive task analysis: The case of CNC tool length offsetting. *Human Factors and Ergonomics in Manufacturing & Service Industries, 26*(1), 52–67.

Needham, R. C., Edwards, B. J., and Prather, D. C. (1980). Flight simulation in air-combat training. *Defense Management Journal, 16*(4), 18–23.

* Neubauer, C., Khooshabeh, P., & Campbell, J. (2017). When less is more: Studying the role of functional fidelity in a low fidelity mixed-reality tank simulator. In *International Conference on Applied Human Factors and Ergonomics* (pp. 220–229). Springer, Cham.

* Noble, C. (2002). The relationship between fidelity and learning in aviation training and assessment. *Journal of Air Transportation, 7*(3), 33–54.

* Norman, G., Dore, K., & Grierson, L. (2012). The minimal relationship between simulation fidelity and transfer of learning. *Medical Education, 46,* 636–647.

* Oberhauser, M., Dreyer, D., Braunstingl, R., & Koglbauer, I. (2018). What's real about virtual reality flight simulation? *Aviation Psychology and Applied Human Factors, 8*(1), 22–34.

* Ornstein, G. N., Nichols, I. A., & Flexman, R. E. (1954). *Evaluation of a contact flight simulator when used in an Air Force primary pilot training program, Part II: Effectiveness of training on component skills.* (Tech. Rep. AFPTRC-TR-54-110) Lackland Air Force Base, TX: Air Force Personnel and Training Research Center.

Page, R. L. (2000). Brief history of flight simulation. In *SimTecT 2000 Proceedings* (pp. 11–17). Lindfield, Australia: Simulation Industry Association of Australia.

Povenmire, H. K., & Roscoe, S. N. (1973). Incremental transfer effectiveness of a ground-based general aviation trainer. *Human Factors, 15*(6), 534–542.

Puntambekar, S., & Hubscher, R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist, 40*(1), 1–12.

Radwin, R., Wang, X., Hu, Y. H., Difranco, N. (2019). *U.S. Patent No. 10,482,613*. Washington, DC: U.S. Patent and Trademark Office.

* Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence 32*(1), 57–95.

Roberts, A. P. J., Stanton, N. A., Plant, K. L., Fay, D. T., & Pope, K. A. (2020). You say it is physical, I say it is functional; let us call the whole thing off! Simulation: an application divided by lack of common language. *Theoretical Issues in Ergonomics Science, 21*(5), 507–536.

Robinson, W. L. (1974). Conscious competency—The mark of a competent instructor. *The Personnel Journal—Baltimore, 53*, 538–539.

* Roenker, D. L., Cissell, G. M., Ball, K. K., Wadley, V. G., & Edwards, J. D. (2003). Speed-of-processing and driving simulator training result in improved driving performance. *Human Factors, 45*, 218–233.

Rosa, E., Dahlstrom, N., Knez, I., Ljung, R., Cameron, M., & Willander, J. (2020). Dynamic decision-making of airline pilots in low-fidelity simulation. *Theoretical Issues in Ergonomics Science*, 1–20.

Roscoe, S. N. (1991). Simulator qualification: Just as phony as it can be. *The International Journal of Aviation Psychology, 1*(4), 335–339.

* Rosen, K. R. (2008). The history of medical simulation. *Journal of Critical Care, 23*, 157–166.

* Rouse, W. B. (1981). Experimental studies and mathematical models of human problem solving performance in fault diagnosis tasks. In J. Rasmussen & W. Rouse (Eds.), *Human detection and diagnosis of system failures* (pp. 199–216). New York, NY: Plenum.

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin, 100*(3), 349–363.

Roza, Z. C. (2005). *Simulation fidelity theory and practice: A unified approach to defining, specifying and measuring realism of simulations* (Doctoral thesis, Delft University of Technology, Delft, The Netherlands). Retrieved from https://repository.tudelft.nl/islandora/object/uuid:a5afd816-4b04-459d-b54c-da93e3b8a0d7/datastream/OBJ

Saab receives order to expand Czech Republic tactical training system. (2019, February). Saab Press Release. Retrieved from https://saabgroup.com/globalassets/cision/documents/2019/20190226-saab-receives-order-to-expand-czech-republic-tactical-training-system-en-0-3218402.pdf

* Sala, G., Tatlidil, K. S., & Gobet, F. (2018). Video game training does not enhance cognitive ability: A comprehensive meta-analytic investigation. *Psychological Bulletin, 144*(2), 111–139.

* Salas, E., Cannon-Bowers, J. A., & Rhodenizer, L. (1998). It is not how much you have but how you use it: Toward a rational use of simulation to support aviation training. *International Journal of Aviation Psychology, 8,* 197–208.

Salas, E., Wilson, K. A., Priest, H. A., & Guthrie, J. W. (2006). Design, delivery, and evaluation of training systems. In G. Salvendy (Ed), *Handbook of human factors and ergonomics* (pp. 472–512). Hoboken, NJ: Wiley and Sons.

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

* Seymour, N. E., Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K., et al. (2002). Virtual reality training improves operating room performance: Results of a randomized, double-blinded study. *Annals of Surgery, 236,* 458–463.

* Schlickum, M. K., Hedman, L., Enochsson, L., Kjellin, A., & Felländer-Tsai, L. (2009). Systematic video game training in surgical novices improves performance in virtual reality endoscopic surgical simulators: A prospective randomized study. *World Journal of Surgery, 33,* 2360–2367.

Schraagen, J. M., Chipman, S. F., & Shalin, V. L. (2000). *Cognitive task analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

* Shin, S., Park, J-H., & Kim, J-H. (2015). Effectiveness of patient simulation in nursing education: Meta-analysis. *Nurse Education Today, 35,* 176–182.

* Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. (2016). Do “brain-training” programs work? *Psychological Science in the Public Interest, 17*, 103–186.

Smith, B. J. (1965). *Task analysis methods compared for application to training equipment development* (Report No. NAVTRADEVCEN 1218-5). Port Washington, NY: U.S. Naval Training Device Center.

Smode, A. F. (1971). *Human factors inputs to the training device design process* (Report No. NAVTRADEVCEN 69-C-0298-1). Orlando, FL: Naval Training Device Center.

Smode, A. F. (1972). *Training device design: Human factors requirements in the technical approach* (Report No. NAVTRAEQUIPCEN-71-C-0013-1). Orlando, FL: Naval Training Equipment Center.

Smode, A. F. (1974). Recent developments in instructor station design and utilization for flight simulators. *Human Factors, 16*(1), 1–18.

* Spangenberg, R. W. (1973). The motion variable in procedural learning. *AV Communication Review, 21*(4), 419–436.

* Steigerwald, S. N., Park, J., Hardy, K. M., Gillman, L. M., & Vergis, A. S. (2015). Does laparoscopic simulation predict intraoperative performance? A comparison between the Fundamentals of Laparoscopic Surgery and LapVR evaluation metrics. *The American Journal of Surgery, 209*(1), 34–39.

* Spruit, E. N., Band, G. P., Hamming, J. F., & Ridderinkhof, K. R. (2014). Optimal training design for procedural motor skills: A review and application to laparoscopic surgery. *Psychological Research, 78*(6), 878–891.

Stanton, N. A., Salmon, P. M., Rafferty, L. A., Walker, G. H., Baber, C., & Jenkins, D. P. (2018). *Human factors methods: A practical guide for engineering and design*. New York, NY: Routledge.

* Stein, M., & Robinski, M. (2012). Simulator sickness in flight simulators of the German armed forces. *Aviation Psychology and Applied Human Factors, 2*(1), 11–19.

* Stewart, J., Johnson, D., & Howse, W. (2008). *Fidelity requirements for army aviation training devices: Issues and answers* (ARI Research Report 1887). Arlington, VA: U.S. Army Research Institute.

Sticha, P. J., Blacksten, H. R., Buede, D. M., Singer, M. J., Gilligan, E. L., Mumaw J. R., & Morrison, J. E. (1990). *Optimization of simulation-based training systems: Model description, implementation, and evaluation* (Technical Report No. 896). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296.

* Swezey, R. W., Perez, R. S., & Allen, J. A. (1991). Effects of instructional strategy and motion presentation conditions on the acquisition and transfer of electromechanical troubleshooting skill. *Human Factors, 33*(3), 309–323.

* Tan, S. C., Marlow, N., Field, J., Altree, M., Babidge, W., Hewett, P., & Maddern, G. J. (2012). A randomized crossover trial examining low-versus high-fidelity simulation in basic laparoscopic skills training. *Surgical endoscopy, 26*(11), 3207–3214.

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

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. (I). *Psychological Review, 8,* 247–261.

Velotta, J. C. (1997, June). Part-task trainers or back-to-basics simulation. In *Proceedings of the 1997 IEEE Sixth Annual Human Factors Meeting* (17, pp. 7–10). Orlando, FL.

Ward, P., Williams, A. M., & Hancock, P. A. (2006). Simulation for performance and training. In K. A. Ericsson, N. Charness, R. R. Hoffman, & P. J. Feltovich (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 243–262). Cambridge, UK: Cambridge University Press.

Ward, P., Gore, J., Hutton, R., Conway, G. E., & Hoffman, R. R. (2018). Adaptive skill as the conditio sine qua non of expertise. *Journal of Applied Research in Memory and Cognition, 7*(1), 35–50.

* Webel, S., Bockholt, U., Engelke, T., Gavish, N., Olbrich, M., & Preusche, C. (2013). An augmented reality training platform for assembly and maintenance skills. *Robotics and Autonomous Systems, 61,* 398–403.

* Westerfield, G., Mitrovic, A., & Billinghurst, M. (2015). Intelligent augmented reality training for motherboard assembly. *International Journal of Artificial Intelligence in Education, 25*(1), 157–172.

* Westra, D. P., Lintern, G., Sheppard, D. J., Thomley, K. E., Mauk, R., Wightman, D. C., and Chambers, W. S. (1985). *Simulator design and instructional features for carrier landing: Transfer study* (Tech. Report NAVTRASYSCEN 85-C-0044-2). Orlando, FL: Naval Training Systems Center.

Wickens, C. D., Hutchins, S., Carolan, T., & Cumming, J. (2013). Effectiveness of part-task training and increasing-difficulty training strategies: A meta-analysis approach. *Human Factors, 55*(2), 461–470.

* Wightman, D. C., & Lintern, G. (1985). Part-task training for tracking and manual control. *Human Factors, 27*(3), 267–283.

* Wightman, D. C., & Sistrunk, F. (1987). Part-task training strategies in simulated carrier landing final-approach training. *Human Factors, 29*(3), 245–254.

* Williams-Bell, F. M., Murphy, B. M., Kapralos, B., Hogue, A., & Weckman, E. J. (2015). Using serious games and virtual simulation for training in the fire service: A review. *Fire Technology, 51*, 553–584.

Williges, B. H., Roscoe, S. N., & Williges, R. C. (1973). Synthetic flight training revisited. *Human Factors, 15*(6), 543–560.

Willis, R. P., Guha, P., Hunter, D. R., & Singer, M. J. (1990). *The optimization of simulation-based training systems: Model data collection and utilization* (Report No. AD-A231 436). Orlando, FL: Training Research Laboratory.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 17*(2), 89–100.

Wulf, G., & Lewthwaite, R. (2016). Optimizing performance through intrinsic motivation and attention for learning: The OPTIMAL theory of motor learning. *Psychonomic Bulletin & Review, 23*, 1382–1414.

Yuksel, B. F., Oleson, K. B., Harrison, L., Peck, E. M., Afergan, D., Chang, R., & Jacob, R. J. (2016, May). Learn piano with BACh: An adaptive learning interface that adjusts task

difficulty based on brain state. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5372–5384).

* Yuviler-Gavish, N., Gutiérrez, T., Webel, S., Rodríguez, J., Peveri, M., Bockholt, U., & Tecchia, F. (2015). Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks. *Interactive Learning Environments, 23*(6), 778–798.

* Zendejas, B., Brydges, R., Hamstra, S. J., & Cook, D. A. (2013). State of the evidence on simulation-based training for laparoscopic surgery: A systematic review. *Annals of Surgery, 257*, 586–593.

Appendix

The studies included in the review that compared between the traditional theory and the challenging findings.

Author(s)	Year	Report type
Traditional theory		
Maintenance and Troubleshooting		
Spangenberg	1973	Experimental
Johnson & Rouse	1982	Experimental
Allen et al.	1986	Experimental
Aviation and Vehicle Training		
Miller	1954	Theoretical
Ornstein, Nichols, & Flexman	1954	Experimental
Dougherty, Houston, & Nicklas	1957	Experimental
Hammerton & Tickner	1967	Experimental
Gerathewohl	1969	Theoretical
Caro	1988	Theoretical
Lintern, Roscoe, & Sivier	1990	Experimental
Gopher, Weil, & Bareket	1994	Experimental
Jentsch & Bowers	1998	Theoretical
Dennis & Harris	1998	Experimental
Roenker et al.	2003	Experimental
Allen et al.	2010	Experimental
Taber	2014	Theoretical
Lee	2018	Theoretical
Buttussi & Chittaro	2018	Experimental
Oberhauser et al.	2018	Theoretical
Medical Training		
Abrahamson, Denson, & Wolf	1969	Experimental
Buchanan	2001	Theoretical
Seymour et al.	2002	Experimental
Rosen	2008	Theoretical
Schlickum et al.	2009	Experimental
Alaraj et al.	2011	Theoretical
Cook et al.	2011	Theoretical
Lewis, Strachan, & Smith	2012	Theoretical
Lopreiato & Sawer	2015	Theoretical
Shin, Park, & Kim	2015	Theoretical
Green, Tariq, & Green	2016	Theoretical
Cant & Cooper	2017	Theoretical
MacLean et al.	2019	Experimental

Appendix *(Continue)*

Author(s)	Year	Report type
Mixed Reality		
Hughes et al.,	2005	Theoretical
Green & Bavelier	2007	Theoretical
Anastassova & Burkhardt	2009	Experimental
De Crescenzio et al.	2011	Theoretical
Freitas & Liarokapis	2011	Theoretical
Webel et al.	2013	Experimental
Dunston, Proctor, & Wang	2014	Experimental
Yuviler-Gavish et al.	2015	Experimental
Westerfield, Mitrovic, & Billinghurst	2015	Experimental
Langley et al.	2016	Theoretical
Nathanael et al.	2016	Experimental
Bediou et al.	2018	Theoretical
Challenging findings		
Maintenance and Troubleshooting		
Laner	1954	Experimental
Fink & Shriver	1978	Theoretical
Rouse	1981	Experimental
Reiter	1987	Theoretical
Swezey et al.	1991	Experimental
Aviation and Vehicle Training		
Denenberg	1954	Experimental
Briggs, Fitts, & Bahrick	1957	Experimental
Hopkins	1975	Theoretical
Crawford et al.	1976	Experimental
Westra et al.	1985	Experimental
Wightman & Lintern	1985	Theoretical
Wightman & Sistrunk	1987	Experimental
Lintern et al.	1989	Experimental
Lintern et al.	1990	Experimental
Salas et al.	1998	Theoretical
Noble	2002	Theoretical
Stewart, Johnson, & Howse	2008	Theoretical
Dahlstrom et al.	2009	Experimental
Stein & Robinski	2012	Experimental

Appendix *(Continue)*

Author(s)	Year	Report type
Medical Training		
Matsumoto et al.	2002	Experimental
Beaubien & Baker	2004	Theoretical
de Giovanni, Roberts, & Norman	2009	Experimental
La Rochelle et al.	2011	Experimental
Norman et al.	2012	Theoretical
Tan et al.	2012	Experimental
Zendejas et al.	2013	Theoretical
Spruit et al.	2014	Theoretical
Steigerwald et al.	2015	Experimental
Moglia et al.	2016	Theoretical
Mixed Reality		
Kozak et al.	1993	Experimental
Borsci, Lawson, & Broome	2015	Theoretical
Williams-Bell et al.	2015	Theoretical
Simons et al.	2016	Theoretical
Kavanagh et al.	2017	Theoretical
Gonzalez-Franco et al.	2017	Experimental
Neubauer, Khooshabeh, & Campbell	2017	Experimental
Sala et al.	2018	Theoretical

Note. The order of reports is chronological within each domain. The "Theoretical" report type refers to literature reviews, meta analyses, and theoretical reports.