Ritter, F. E., Yeh, M. K., Stager, S. J., McDermott, A. F., & Weyhrauch, P. W. (2023). The effect of task fidelity on learning curves: A synthetic analysis. *International Journal of Human-Computer Interaction, 39*(11), 2253-2267.

The Effect of Task Fidelity on Learning Curves: A Synthetic Analysis

Frank E. Ritter (frank.ritter@psu.edu)
Martin K. Yeh (martin.yeh@psu.edu)
College of IST, Penn State
University Park, PA

Ashley F. McDermott (ashley.f.mcdermott@gmail.com)
Peter W. Weyhrauch (pweyhrauch@cra.com)
Charles River Analytics
Cambridge, MA
Word count: 9983 all in

27 mar 2022

20aug 2022, v13b, final copy, updated to note where figures come from
19 july 2022, v13, final copy lightly revised.

Abstract

There have been discussions about the value of fidelity since simulation-based training systems have been created. A primary question, which has yet to be fully answered, is what is the effect of level of simulation fidelity on learning on a target task? We present a new analysis method and use it for several analyses of a training simulation for an electronic maintenance task with two levels of fidelity: a high-fidelity simulation that basically takes as much time as the real-world task and a low-fidelity simulation with minimal delays and many actions removed or reduced in fidelity and time. The analyses are based on the Keystroke-Level Model (KLM) and the power law of learning. The analyses predict that the performance on the low-fidelity simulation initially takes between one quarter and one eighth of the time of the high, and thus starts out providing between four and eight times as many practice trials in a given time period. The low-fidelity curve has a lower intercept and a steeper slope. Learners that move from low to high appear to not be adversely affected. For a small number of practice trials, this makes a significant difference. We also explore the effect of missing subtasks in the low-fidelity simulation. This effect varies with the tasks included: If the low-fidelity simulation does not train an important task, learners can be slower when they transfer. We also analyze a simulation that we have built and are studying. These analyses demonstrate that using lower fidelity training situations help most where there is less time to practice, and that if there is extensive time to practice full fidelity has nearly the same outcome (but perhaps not the same costs or risks). We discuss how this analysis approach could be extended to help choose the level of fidelity of training simulations.

Keywords: fidelity, training systems, training simulations, learning curves

Acknowledgements: Grace Good, Jake Graham, Jong Kim, Jacob Oury, Clare Robson, Fred Ryans, Shan Wang, Steve Zimmerman, and two anonymous reviewers provided useful comments. James Niehaus suggested how to improve the schematic. This work was supported by ONR, N00014-18-C-7015 and N00014-15-1-2275. A previous version was published at the International Conference on Cognitive Modeling, and the reviewers there provided useful comments.

Frank Ritter is required by Pennsylvania State University [sic] to include this paragraph: "Frank E. Ritter, the co-author of this paper, have financial interest with Charles River Analytics Inc.; a company in which Frank E. Ritter provides consulting services and could potentially benefit from by the results of this research. The interest has been reviewed and is being monitored 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." [sic]

Introduction

What is the effect of varying the level of simulation fidelity on learning? What happens to learning when a learner practices in a less visually complex simulation and then moves to a more realistic or higher fidelity simulation or the actual task? In this paper we explore how task fidelity affects how fast a task is learned in an example task with three different analyses examining learning a single task on two different interfaces. We present a methodology for examining the effect of fidelity on learning generally using two commonly used modeling techniques, the keystroke level model (Card, Moran, & Newell, 1983; Card, Moran, & Newell, 1980) and the power law of learning (Crossman, 1959; Newell & Rosenbloom, 1981; Ritter & Schooler, 2001) that have not been previously combined and applied to training simulations to our knowledge.

We consider these questions by using an example of a maintenance training simulation with two levels of fidelity: (a) a full simulation that basically takes as much time as the real-world task (high fidelity), and (b) a simple simulation with shorter delays and with many non-task specific actions removed or reduced in fidelity and time (low fidelity). The low-fidelity simulation starts out taking less time to complete (depending on the simulation), and thus enables the trainee to perform more practice trials in a given time period. This work may ultimately lead to a better method for determining optimal simulation training time based on predicting performance improvement through learning curves.

The results of this study demonstrate how task complexity of simulated tasks influences the time to perform the task, and this in turn modifies the associated learning curves, both in the intercept and learning rate over training time. Thus, this difference in trial time can make a significant difference in learning outcomes because more practice trials occur with a faster task. We also show that for some simulations that do not support all the subtasks, learning gains can be less impressive or in the opposite direction. Finally, the amount of training on the tasks also influences the choice of trainer fidelity as well.

After briefly reviewing the effect of training system fidelity we introduce a maintenance task that we developed to study learning and retention. We base our analysis on a simple non-information-processing model of learning based on ACT-R, Soar, and the power law of learning. The learning curves from this analysis allow us to draw some conclusions about the effect of fidelity on the effectiveness of training, notably that using lower fidelity training situations helps most where there is less time to practice, and that if there is extensive time to practice full fidelity has nearly the same outcome (but perhaps not the same costs or risks) as does starting with a simple simulation and moving to the complex simulation. We also explore how subtask inclusion and transfer influences the learning curves.

This work was initially reported in a short conference paper (Ritter & McDermott, 2020). We have extended the work to look at individual trials instead of blocks of learning, modified the model to take better account of previous practice and transfer, and have included more example analyses.

Brief Review of Fidelity on Training

There is a long-standing debate of the effects of fidelity on training with simulation. The early research on fidelity was based on the natural assumption that higher fidelity simulations would necessarily lead to better learning, because the simulation would more closely resemble the actual system (e.g., Allen, Hays, & Buffardi, 1986; Miller, 1954; Noble, 2002). However, much of the research supporting this notion was conducted from the 1950s to 1980s, so it had a lower ceiling for how representative high simulation fidelity could be at the time. There is also a body of research showing that higher fidelity is not always desirable to maximize learning (e.g., Dahlstrom, Dekker, van Winsen, & Nyce, 2009; Havinghurst, Fields, & Fields, 2003; Lesgold, Lajoie, Bunzon, & Eggan, 1992; Swezey, Perez, & Allen, 1991).

Delving into this literature leads to the question of what fidelity actually means. The most common distinction is between surface or physical fidelity versus operational or task fidelity (Allen et al., 1986; Liu, Macchiarella, &

Vincenzi, 2009). Within physical fidelity there are still many dimensions, including visual layout, amount of visual clutter (e.g., non-targets on a radar or amount of detritus and white (civilian) forces on a street), auditory fidelity, and haptic fidelity. All of these dimensions have the potential to affect both speed of learning and amount of transfer to the real task. Some of these dimensions, however, are not relevant to the task being taught. To properly learn a task, the simulation should have reasonably high fidelity on task-relevant dimensions (Prophet & Boyd, 1970; Thorndike & Woodworth, 1901), but irrelevant dimensions may best be kept at a lower fidelity to minimize distraction from the task (Alessi, 1988).

An additional factor that affects task time and transfer of learning is the experience level of the learner (Alessi & Trollip, 1991). The experience of the learner will affect the cognitive load associated with higher fidelities and the dimensions of fidelity that could be considered task relevant (Alessi, 1988). For example, an expert who is familiar with the actual task will likely experience less cognitive load with a nearly full fidelity simulation for additional training than a novice learning about the task and interface for the first time. Similarly, the age or technology experience of the learner can affect what sorts of simulations are easily usable. Simulations could introduce interactions (e.g., joy sticks, touch displays, VR) that are natural for younger adults who are experienced interacting with new technology but novel or slower for older adults, or vice versa depending on the interaction interfaces and their prevalence (John & Jastrzembski, 2010).

The question of when higher fidelity is better for learning continues to be debated because it is not clear why or when lower fidelity simulations provide the most advantage. As we have discussed, experience of the learner and cognitive load are considered to be two important contributing factors, as is the type of task.

In this paper we propose an additional factor, the number of repetitions of the task (or subtask) that a learner is able to complete while training. We examine the effect of initial task time on the learning curve, and through a series of analyses find and illustrate several important factors. In the discussion we note further factors that came to light while performing the analyses.

Method

In this section we attempt to cast the simulation into a traditional psychology method section, noting the apparatus, the task as procedure (which we present second for clarity), and the model of learning as the participant.

Apparatus: A Simple Task and Model of Learning and Fidelity to Use as an Example

The Ben Franklin Rader

To examine the effect of fidelity on number of repetitions and thus on learning we use an example simulation. The Ben-Franklin (BF) Radar (Ritter, Tehranchi, Brener, & Wang, 2019) system was designed to support troubleshooting within the confines of a study: to be closer to a real maintenance task than the Klingon Laser Bank task (Friedrich & Ritter, 2020; Kieras & Bovair, 1984; Ritter & Bibby, 2008), but not so complex that it would take more than an hour to learn.

The schematic in Figure 1 shows its relative complexity. The system has 35 replaceable components that can have faults, 15 switches, and a power supply that cannot have faults. These components are organized in five subsystems that vary in their complexity and connectivity. Blue lines are power connections; red lines are information; purple lines are both. The schematic also identifies certain components that have their status displayed on the front panel of the BF Radar. The system is based on the Klingon Laser Bank task and on a functional radar system (Charvat, 2011). This system has been realized in several ways with different levels of assumed fidelity. Each single-fault error state creates a unique light configuration and is always solvable.

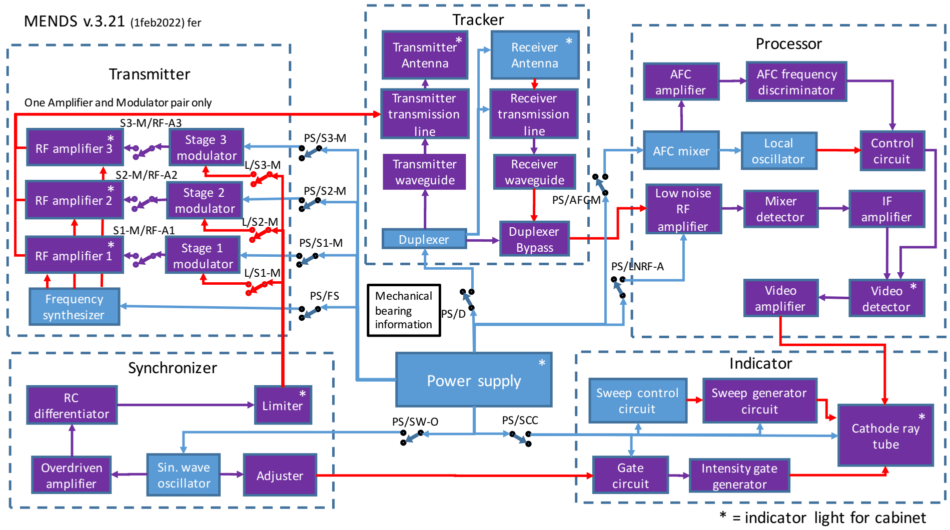

Figure 1. Schematic of the Ben-Franklin Radar simulation.

The Ben Franklin Radar Simulations

In our analysis we examine two potential implementations of the BF Radar device. The first (MENDS[1] High Fidelity) is realized in software and can be realized in hardware, which has been partially built. The second (MENDS Low fidelity) is realized in software alone. How the radar works and how the faults works are the same in both. They are available under license.

High-Fidelity Simulation. Figure 2 depicts a higher fidelity version of the MENDS simulation, designed to be realized in an approximate 2 ft. physical metal cube with components using Raspberry Pi's. To run remotely during the pandemic we used a virtual reality version implemented in Unity 3D. It has a cabinet holding trays for each subsystem. The front panel provides a summary of the system. The other trays each hold one subsystem.

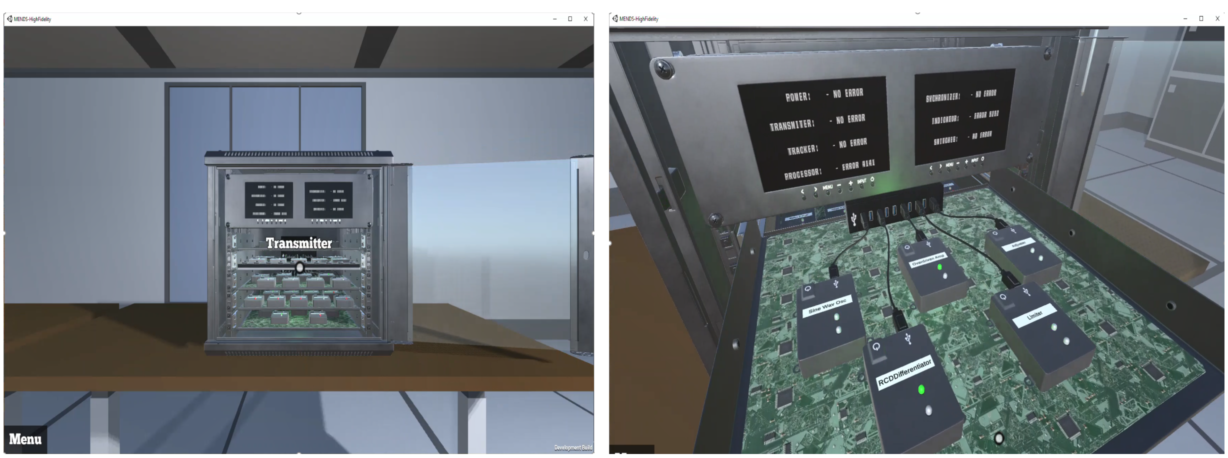

Figure 2. The MENDS high-fidelity realization of the BF Radar (left), showing the cabinet (with its door opened), and the Synchronizer tray (right).

The task is to troubleshoot the radar. For the actual troubleshooting task, the user must first put on a grounding strap, and then examine the lights, open the cabinet door, pull out a tray, and examine its contents. They must then choose the broken component, find the replacement part, and replace the broken component. To set up a trial, the experimenter must have the user look away, replace a working component with a broken component, and then close the tray and the door.

[1] Maintenance Enhancement with Next Generation-Development of Skills

Low-Fidelity Simulation. Figure 3 shows the general layout of the MENDS Low-fidelity simulation of the BF radar system (Ritter, Tehranchi, Brener, et al., 2019). The simulation is implemented in Unity. The front panel (left image) shows the subsystems; the lights in the upper right corner of each square show which subcomponents are working. An individual tray (right image) shows the components that are working (yellow and white lights) and the components that are not working (red light and grayed out component).

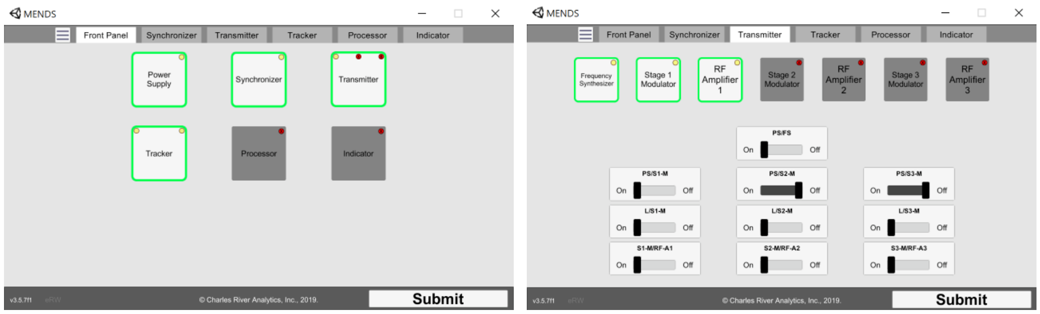

Figure 3. The interface of the MENDS low-fidelity simulation. At left is the initial screen, which represents the front panel and provides an overview of the radar status. By selecting a tab (shown on right), the status of the subsystem tray can be seen. In this case, the Transmitter tab has been selected, and the line replaceable units and switches of the Transmitter are shown. In these pictures, the fault is in the Processor subsystem.

The Task

This section describes the task that is used in the analyses. Multiple tasks can be done with the system; it could be turned on, or switches could be flipped to get it working correctly or to avoid a broken component. In this case the task is to find and replace a broken part (a fault). We assume that the user has been taught the structure of the BF Radar schematic and has it available, either in their head or on a sheet of paper, but has not performed the task before. That is, they know how to do the task but have not practiced it. Step 1 (Insert fault) is performed by an experimenter or the system, to insert a faulted component. This task uses the Overdriven Amplifier fault, an early component in the system, as the example fault for the analyses presented here.

Table 1. Task used to illustrate learning troubleshooting. The fault modeled is the Overdriven Amplifier fault, which is in the Synchronizer tray. Subtasks are grouped into four categories for clarity and illustration purposes.

Step	Stage and Name
	Setup
1	Insert fault
2	Approach System
3	Open cabinet
4	Ground yourself
	Find subsystem
5	Check front panel
6	Choose tray
7	Open tray
	Tray loop
8	Check light
9	If off, have answer
	If on, go to next light
10	Check light - recognize Overdriven Amp fault
11	Recognize fault
	Replacement
12	Indicate fault / replace component
13	Confirm done

Human Participant Data on Tasks Related to the Ben Franklin Rader Troubleshooting Task

There is some reason to believe that troubleshooting a fault in this simulation is a useful task for studying the effect of fidelity on learning and transfer because we have three sets of data related to this task. On the original Klingon Laser Bank task, Ritter and Bibby (2008) saw reaction times ranging from 20 s on the first trial, to around 7 s when practiced over 20 trials. Friedrich and Ritter (2020) reported similar times. When the response time is less than 7 s it difficult to see differences between interfaces and training because of inherent noise in reaction times. Thus, this task is too fast to study over a longer period of training time, but otherwise shows that the task is learnable. Subtask times used in the model are consistent with times from Keystroke-Level Model (KLM) and GOMS (Goal-Operator-Method-Selection rules) task analysis methods (Burns, Ritter, & Zhang, 2016; Card et al., 1983; Card et al., 1980; John & Kieras, 1996).

Using the MENDS Low-fidelity simulation, Ritter et al. (2019) had a participant with four practice sessions with 5 practice trials per session that went from 60 s to 22 s. Their learning rate across the sessions in that study was found to be 1.05, slightly higher than the learning rates noted earlier. The learning rates within the trials in the four sessions in that study were 1.54, 0.22, 0.089, and 0.26, respectively, using the formula from Rosenbloom and Newell (1987). This learning rate likely varied because it was a single participant, and the sessions were not continuous. The participants in Ritter and Bibby (2008) doing the Klingon Laser Bank task had a learning rate of 0.477 for all trials, and 0.769 if the first two warm up trials were excluded. These numbers are not from a canonical situation, but show the range of learning rates and provide justification that the times and structure of the task are reasonable.

Participant model: A Theory of Learning of the Task

This section describes the task analysis that was used to compute the task performance time, which is used as a measure of learning. Other measures could be used, but this is a commonly used measure of learning. The performance time includes a learning component, and it can be adjusted to account for different interfaces.

The time to perform a task is broken down into three types of time: time to perform the skills to be learned, time to perform skills that are already learned (or, essentially learned), and system response time (time for the system to respond to the user and display further information or to compute and display information to the user). Skills that are to be learned get faster with practice. On this task, learnable skills include: recognizing the lights and their implications. Skills that are essentially already learned are moving the mouse and clicking; system response times include replacing faults or inserting faults by the system (these are combined in the analyses).

These steps and their times are taken from the Keystroke-Level Model (Card et al., 1980). Each step takes time and is assumed to be error free. The operator times from the KLM have been used numerous times to predict task times (e.g., Batran & Dunlop, 2014; Card et al., 1983, multiple studies; Gong & Kieras, 1994; John, Prevas, Salvucci, & Koedinger, 2004, 3 tasks; Katsanos, Karousos, Tselios, Xenos, & Avouris, 2013, 6 tasks).

In Table 2, the *Setup* tasks are setting up to perform the task. *Find subsystem* finds the subsystem and tray the fault is in. The *Tray loop* works through the components in that subsystem until the fault is found. Table 2 shows the two steps for the Overdriven Amplifier fault.

These times are used to compute the time to do the task using Equation 1. This equation is consistent with Soar's (Newell, 1990; Rosenbloom & Newell, 1987) and ACT-R's (Anderson, 2007; Ritter, Tehranchi, & Oury, 2019) learning theories and mechanisms, and the learning curve theory (Crossman, 1959; Ritter & Schooler, 2001; Seibel, 1963; Snoddy, 1926). All these learning theories treat repetitions on the task as being the driver of the learning question and have a good fit to learning data.

$$\text{(1)}\quad \text{Time} = \sum \left(\text{Fixed time} + \text{learned time} * (\text{Trial})^{-\alpha}\right) \text{ for all subtasks}$$

A response time is thus computed for each trial. We do this for a series of 100 trials. Longer lengths of trials could be done, but we can illustrate our points with this sequence. The model does not include stochastic components, so it is run once, not multiple times (Ritter, Schoelles, Quigley, & Klein, 2011).

The learning rate α (alpha) = 0.4 was chosen. This value of α is consistent with values from Newell and Rosenbloom (1981, 0.06 - 0.81, a variety of tasks); and similar to values from Delaney, Reder, Staszewski, and Ritter (1998, 0.265 - 0.510, mental arithmetic); and Kim and Ritter (2016, 0.4 – 1.2, spreadsheet tasks). We will also show that this analysis approach is somewhat robust to learning rate changes.

We also computed time to perform the task if users were to start in the low-fidelity simulation and practice for a period of time and then move to the high-fidelity simulation and do a single trial. These times in the high-fidelity simulation are thus not a learning curve, but show how well the learner would perform on the first trial in the high-fidelity simulation after variable amounts of practice in the low-fidelity simulation.

Equation 2 shows how that time is computed. This equation arises from applying the power law of learning (Newell & Rosenbloom, 1981; Ritter & Schooler, 2001; Rosenbloom & Newell, 1987) to the subtask times. The first term is the fixed tasks and system response times in the high-fidelity task. The second term is the task time in the high-fidelity simulation where the subtasks are the same in the high- and low-fidelity simulations. In this case, learning transfers and further practice is added to previous practice. The third term is where the knowledge does not transfer or was not learned in the low-fidelity simulation; learners are at trial 1 when knowledge does not transfer—this is just the initial subtask time without learning.

$$\begin{aligned}\text{(2)}\quad \text{Time} = \sum \Big(& \text{Fixed task times in High} \\ &+ (\text{Task time learnable in High}) * (\text{Trial})^{-\alpha} \quad \text{[where learning transfers]} \\ &+ (\text{Task time learnable only in High}) * (1)^{-\alpha} \Big) \quad \text{[where learning does not transfer]}\end{aligned}$$

Thus, the time to perform a task will vary based on the subtasks in the two systems, and how much practice the learner receives with both simulations. This approach is basically the same as our earlier predictions (Ritter & McDermott, 2020), but more complete in that it models every trial instead of blocks of trials, and the theory is revised to keep track of the transferability of learning for each subtask. This model has several simplifications; it does not predict variance in the task strategy or performance time, nor errors. It is a simplification supporting comparisons of designs.

Procedure: Application of the Theory of Learning to Instances of the Interfaces

To examine the effect of fidelity on learning and to illustrate its use, we applied the learning theory to different hypothetical instances of the MENDS interfaces and to the interfaces as built. We included instances: (a) where the simple interface taught all tasks but took less time to use and to load faults, (b) where the simpler interface did not train an important and learnable task, and (c) how the interfaces have been built for use in a study. We also examine (d) the effect of varying the learning rate. The next sections are essentially results sections presenting the predictions of the model of the user learning in different training simulations.

Analysis 1: Analyzing A Low-Fidelity Training Simulation that Is Faster than High

In this analysis we examine a hypothetical low-fidelity simulation that includes all the subtasks in the MENDS Low-fidelity simulation where all the skills can be transferred to the high-fidelity simulation, and the tasks are simply faster in the low-fidelity simulation than with the high-fidelity simulation. The low-fidelity can be faster because of a faster system response time, less complex interaction with the low-fidelity interface, or because nonlearnable subtasks are not included. Table 2 defines the task through the subtask times for the two hypothetical simulations for a novice to complete. Times for subtasks were assigned by the authors using the KLM (e.g., mouse click, Mental Operation, Mop) based on the two simulation trainers. System response times and well-learned subtasks were put into the fixed times.

Table 2. Task analysis for when the low-fidelity training simulation trains all tasks and is faster on all subtasks. The fault modeled is the Overdriven Amplifier fault. (Times are in s.)

		Training System High Fidelity		Training System Low Fidelity	
		Skills		**Skills**	
		Learnable	**Fixed**	**Learnable**	**Fixed**
	Sub total time (s)	**14.75**	**33.50**	**14.75**	**7.55**
	Total Task Time	**48.25**		**22.3**	
	Initial Trials per min.	1.2		2.7	
	Setup	**Learnable**	**Fixed**	**Learnable**	**Fixed**
1	Insert fault		12		0.05
		insert fault		click"Next"	
2	Approach System		5		1.5
3	Open cabinet	1.35	3	1.35	1.5
		lift lid		Mop, move, click to open	
4	Ground yourself	0	0	0	0
			(not required)		
	Find Subsystem				
5	Check front panel	4.8		4.8	
		(scan first panel, 0.25/light + Mop)			
6	Choose tray	1.35	0	1.35	0
		Mop		Mop	
7	Open tray	1.35	4	1.35	1.5
		reach etc.		move+click	
	Tray Loop				
8	Check light	0.25		0.25	
		(scan first tray, 0.25/ light)			
9	If off, have answer	1.35		1.35	
	If on, go to next light				
10	Check light	0.25		0.25	
11	Recognize fault	1.35		1.35	
	Replacement				
12	Indicate fault	1.35	8	1.35	1.5
		replace physically		Mop+move+click	
13	Confirm done	1.35	1.5	1.35	1.5
		say "done"		Mop+move+click	

remove excess vertical lines on second lines of subtasks

Figure 4 shows learning curves for the high- and low-fidelity simulations. For low and high simulations, these are standard learning curves showing response time (time to find the fault) vs. trial number. The low-fidelity curve starts out lower and stays lower than the high-fidelity learning curve because the subtasks initially take less time and with practice continue to take less time.

We would also like to see the effect of a mixed strategy where learning starts on the low-fidelity simulation and moves to the high-fidelity simulation. To examine this effect, we generate an “outcome curve” by examining a learning sequence consisting of a variable number of trials on the low-fidelity simulation and then a single trial on the high-fidelity simulation. The independent variable of the outcome curve is the total number of trials, i.e., number of low plus one for high. The dependent variable is the response time in the final trial in high.

Using this approach, the mixed-fidelity training strategy can be evaluated against an all-high strategy by comparing the high-fidelity learning curve and the outcome curve. Each point on the outcome curve represents a potential choice of when to transition from low to high. This provides insight into whether a mixed strategy is useful, and when the transition might be best done to optimize learning. In the current analysis, as shown in Figure 4, the outcome curve has the same values as the high-fidelity curve. For example, the value of the mixed outcome curve at trial 11 is the time that the learner takes to do the high-fidelity simulation after spending 10 trials in the low fidelity simulation, and this is also the same result as spending the previous 10 trials in the high-fidelity simulation.

Looking at task time by trials, the mixed learning curve outcomes are exactly the same as the learning curve for the high-fidelity simulation. In other words, when taken trial by trial, a trial in the low-fidelity simulation does not provide any additional learning above the learning in a trial in the high-fidelity simulation. This graph based on trials does not suggest that lower fidelity training has any advantage given the same number of trials. With trial number as the x-axis, the high-fidelity learning curve and the mixed outcome curve are identical.

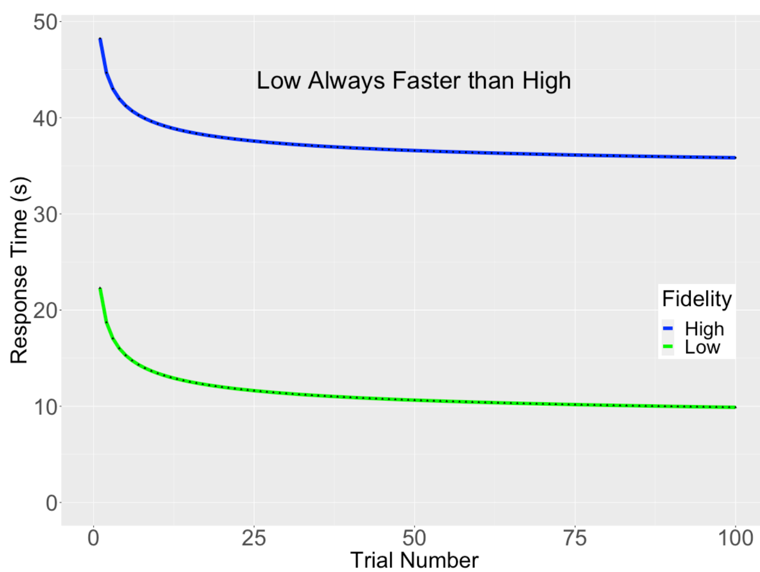

Figure 4. The learning curves for the high- and low-fidelity simulations by trial for 100 trials. The low to high transitions (mixed learning curves) overlay the High curve.

[should redraw to show the transition lines, every 10 faults]

However, in a real system, we will typically not be comparing learning over trials, but learning over *time on task,* or cumulative time. We modify the definition of an outcome curve: the independent variable is no longer number of trials, but instead time taken to perform those trials. We adjust the graph by plotting response time vs. cumulative time, we see a difference between the two strategies. Computing the time at which each trial occurs at (rather than trial number) is a slightly more complicated analysis, which might be why this graph is rarely seen in the learning curve literature.

Figure 5 shows how response time decreases with cumulative time on task when using low- and high-fidelity simulations, as well as a mixed-outcome curve. 100 trials with the high-fidelity simulation takes about 3700 s; 100 trials with the low-fidelity simulation takes about 1200 s. Figure 5 shows that the learner can learn to perform the task using the low-fidelity simulation in approximately one third of the time of the high-fidelity simulation. Thus, decreasing the time that learners spend either waiting for the system (from system response time) or spending doing tasks that are not of interest (not shown in this example), will increase task repetitions that lead to learning. Thus, decreasing system response time and removing tasks not useful to train are a good idea.

The graph of the outcome curve (the open circles) shows that when learners on the low-fidelity simulation move to the high-fidelity simulation, they are faster at that point in time. The curve shows that at larger values of trials the two curves will become closer. The slant on the lines showing the transfer from low to high represents the time to perform the task on the high-fidelity simulation for that data point. Initially, these times to perform the task in the high-fidelity simulation are relatively long, so the slant is greater for the beginning of the curve.

The difference in response time for the group moving from low to high may or may not be important. Where there are time-critical environments, this difference in time might be a particularly large difference. If resources are costed not by time but by trials, then there is no savings.

The effect of getting more practice trials with a lower fidelity trainer over a higher-fidelity trainer is more pronounced when there are fewer trials. The difference also is influenced by the choices made to make the low-fidelity simulation faster. Changing how much faster the low-fidelity simulation is could increase or decrease the differences. While the graphs suggest the absolute response time may not be significantly lower, the graphs suggest that time to train to a given criteria may be significantly faster when the low-fidelity simulation is faster to use than the high-fidelity simulation. We will next explore a situation where the low-fidelity simulation does not include all the tasks in the high-fidelity simulation.

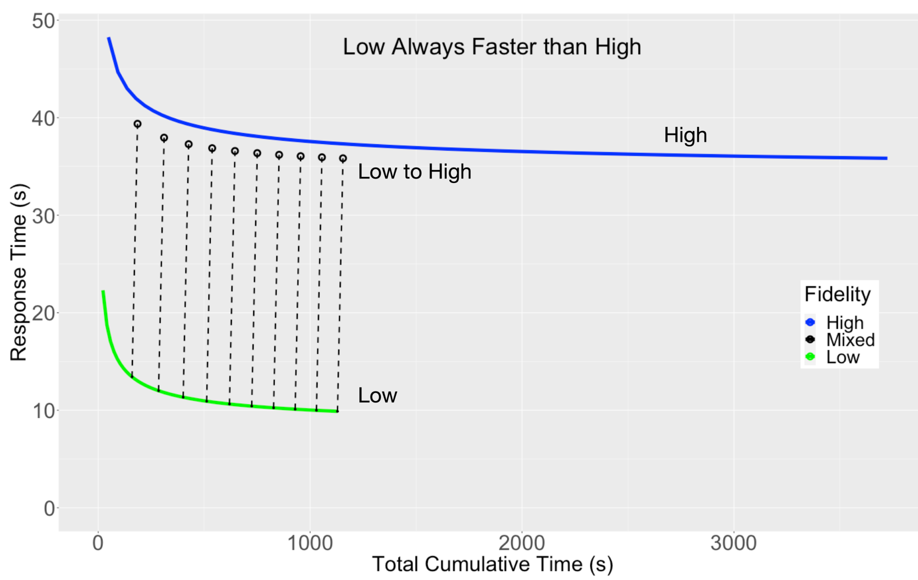

Figure 5. The learning curves for the low- and high-fidelity simulations plotted by cumulative time for 100 trials. The open black circles indicate the response time on the high-fidelity simulation when a learner would move to that simulation from the low-fidelity simulation at that point in time. These mixed times at transfer are shown for every 10 trials, starting at trial 10.

Analysis 2: Analyzing a Low-Fidelity Training Simulation that Does Not Include a Required Subtask

What if there is an important subtask not trained in the low-fidelity training simulation? This analysis examines what happens in a hypothetic implementation of a low-fidelity simulation that does not include an essential and learnable subtask. In this case, the subtask assumes that the use of a grounding strap is learnable with little fixed knowledge and that it is not chosen to be included in the low-fidelity training simulation. Table 3 defines the task shows how the difference in time arises; the difference is a 15 s learnable subtask performed only in the high-fidelity simulation. The missing subtask lets us examine one aspect of learning transfer between low-fidelity and high-fidelity simulations, where the low-fidelity simulations does not teach elements of the overall task to be learned.

Table 3. Task analysis. The fault modeled is the Overdriven Amplifier fault. (Times are in s). The training simulation does not include the ground yourself subtask, which is assumed to be slightly long and learnable. The only change between the previous analysis and this one is shown in circled with a bold dashed red line.

		Training System High Fidelity		Training System Low Fidelity	
		Skills		Skills	
		Learnable	Fixed	Learnable	Fixed
	Sub total time (s)	**29.75**	**33.50**	**14.75**	**7.55**
	Total Task Time	**63.25**		**22.3**	
	Initial Trials per min.	0.9		2.7	
	Setup	**Learnable**	**Fixed**	**Learnable**	**Fixed**
1	Insert fault		12		0.05
		insert fault		click "Next"	
2	Approach System		5		1.5
3	Open cabinet	1.35	3	1.35	1.5
		lift lid		click to open	
4	Ground yourself	15	0	0	0
		apply strap		not required	
	Find Subsystem				
5	Check front panel	4.8		4.8	

remove excess vertical lines on second lines of subtasks

Figure 6 shows the learning curves for the high- and low-fidelity simulations, as well as the mixed outcome curve (black circles) for the mixed strategy. Performance on the high-fidelity simulation is slower than the low-fidelity simulation. The outcome curve shows that when learners transfer to the high-fidelity simulation early on, they are about the same speed as if they had spent the same time on the high-fidelity simulation. But, with additional practice time, the lack of practice on the grounding strap subtask makes them increasingly slower compared to those trained on the high-fidelity simulation when they transition. After 100 trials those that saw just the high-fidelity simulation will take about 38 s to perform the task, and training will take about 4000 s. Learners who do 100 trials of the low-fidelity training simulation will take about 54 s on the high-fidelity task, and will take about 1200 s to perform these trials. In this case, starting and staying in the high-fidelity simulation will lead to faster performance on the task.

For low-fidelity simulations that do not support all the subtasks, the gains from a lower fidelity simulation in learning the other tasks can be offset by not spending practice time with the unincluded but important subtasks or other factors such as cost. This effect must be kept in mind.

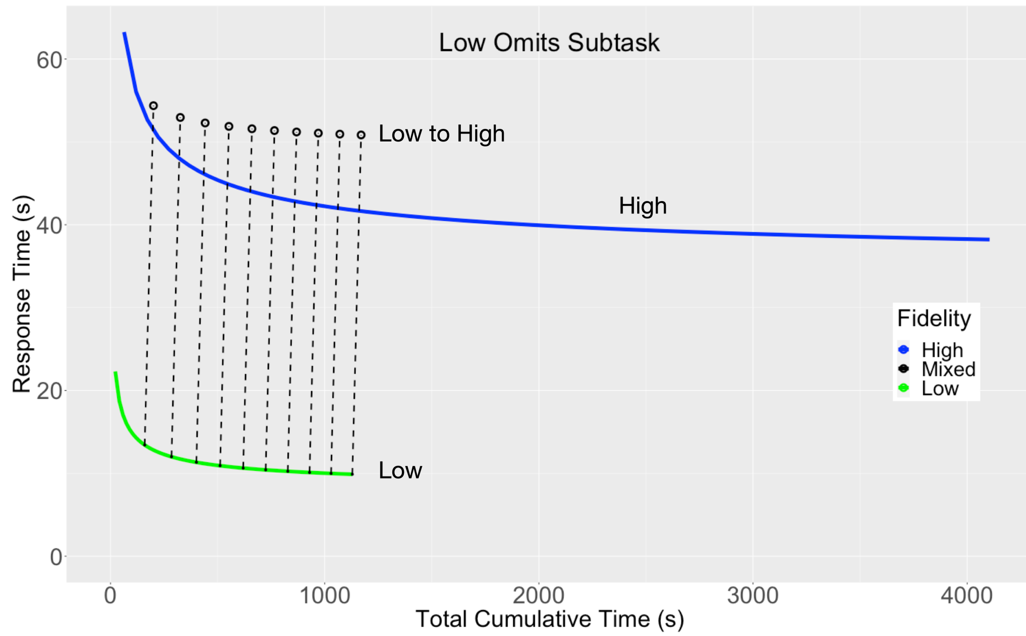

Figure 6. The learning curve for the high- and low-fidelity training simulations for 100 trials. The outcome curve (open circles) indicates the response time on the high-fidelity simulation when learners would move to that simulation from the low-fidelity simulation at that point in time. These mixed times at transfer are shown for every 10 trials, starting at trial 10.

There may be reasons to prefer the low-fidelity simulation. We take this up in more detail in the discussion, including that it might be worth it for price or training on other tasks or as an adjunct. It might also be worth creating an additional trainer for the untrained task if the price or risks are worth it.

Analysis 3: Analyzing the MENDS Training Simulation as Built

Here, we present predictions for the simulations used in a study in progress to explore empirically the effects of fidelity on learning. The study and simulations were designed with a preliminary version of this analysis.

Table 4 shows that the low-fidelity simulation as built does not include all tasks, and is predicted to be much faster. Some of the tasks not included in the low-fidelity interface are not very learnable and include much longer system response times. For example, the time to replace the fault in the high-fidelity simulation starts higher, at 120 s, because the simulation's interface and interactions are complex.

Table 4. Task analysis for the STRUDEL study we are running. The fault modeled is the Overdriven Amplifier fault. (Times are in s.)

		Training System **High** Fidelity		Training System **Low** Fidelity	
		Skills		**Skills**	
		Learnable	**Fixed**	**Learnable**	**Fixed**
	Sub total time (s)	**149.75**	**75.05**	**21.5**	**4.55**
	Total Task Time	**224.80**		**26.05**	
	Initial Trials per min.	0.3		2.3	
	Setup	**Learnable**	**Fixed**	**Learnable**	**Fixed**
1	Insert fault	1.50	0.05		0.05
		(merge + system response time)			
2	Approach System	1.35	4.00	1.35	0
3	Open cabinet	1.35	3	0	0
		lift lid		already open	
4	Ground yourself	0	0	0	0
			(not required)		
	Find Subsystem				
5	Check front panel	4.8		4.8	
		(scan first panel, 0.25 / light plus Mop)			
6	Choose tray	1.35	0	1.35	0
		Mop		Mop	
7	Open tray	1.35	4	1.35	1.5
		reach etc.		move + click	
	Tray Loop				
8	Check light	0.25		0.25	
		(scan first tray, 0.25 / light)			
9	If off, have answer	1.35		1.35	
	If on, go to next light				
10	Check light	0.25		0.25	
11	Recognize fault	8.1		8.1	
	Replacement	Mops		Mops	
12	Indicate/replace fault	120	60	1.35	1.5
		replace physically		Mop+move+click	
13	Confirm done	8.1	4	1.35	1.5
		Check and say "done"		Mop+move+click	

remove excess vertical lines on second lines of subtasks

Figure 7 shows the learning curves for the high- and low-fidelity simulations along with the mixed outcome curve. The low-fidelity simulation is even faster than the example in Figure 5. The additional fixed times that the high-fidelity simulation requires has a large effect on task time and number of practice trials possible in a given timeframe, allowing a greater amount of practice for the low-fidelity simulation in the same time period. The transitions from the low-fidelity simulation to the high-fidelity simulation (using the mixed strategy) shows that learners of the low-fidelity simulation will be faster than their counterparts who were only trained on the high-fidelity simulation at all points of time because they acquired a large number of practice trials when they were using the low-fidelity simulation.

Figure 7 shows that when a learner starts on the low-fidelity simulation and moves to the high-fidelity simulation after trial 1 (the first slanted line), they are very much like those who started on the high-fidelity simulation. However, after the first 100 trials, training on the low-fidelity has allowed learning to progress much more rapidly because a practice trial takes less time with the low-fidelity interface. Within the time span of 500 trials, learners on the low-fidelity simulation will stay much faster than those using the high-fidelity simulation, but at an even higher number of trials the two curves will become closer because the power law compresses the effect of practice.

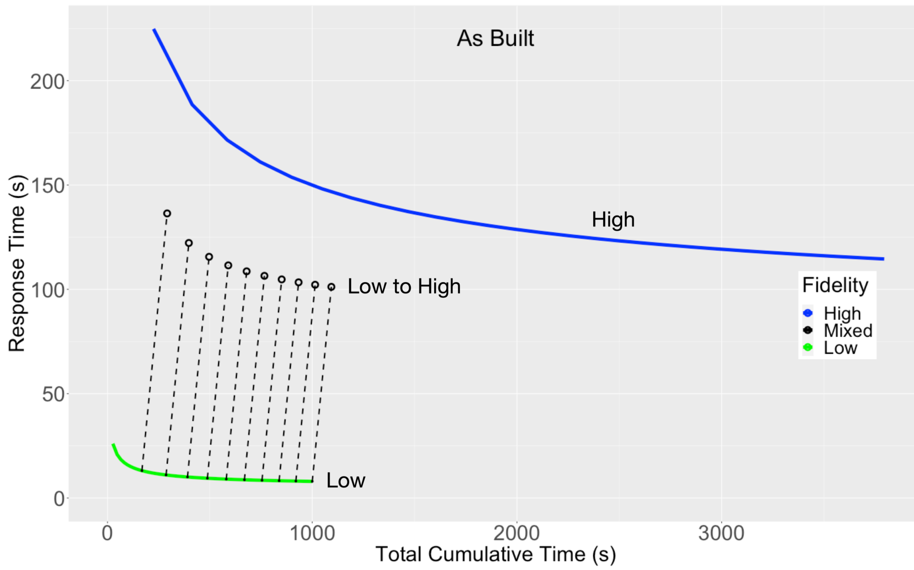

Figure 7. The predicted learning curve for the low- and high-fidelity training simulations for 100 trials. The outcome curve (open circles) indicates the response time on the high-fidelity simulation when learners would move to that simulation from the low-fidelity simulation at that point in time. These mixed times at transfer are shown for every 10 trials, starting at trial 10.

The empirical study will have learners train with the high- or low-fidelity simulations for 15 minutes (900 s). On the first trial, the high-fidelity simulation response time will be around 242 s and the low-fidelity response time around 26 s. This is shown in Figure 8 as points 1 and 4. After 15 min. (900 s) the low-fidelity users will have seen about 34 practice trials (point 5), and the high-fidelity users will have seen about 4 trials (point 2). The high-fidelity trained learners will be at 160 s, whereas the low-fidelity trained learners will be at approximately 20 s on the low-fidelity simulation and at 95 s on the high-fidelity simulation.

Training will then continue for both groups with the high-fidelity simulation (points 2 and 6). With an additional 15 min. of training, the high-fidelity learners will be at approximately 135 s after about 28 trials (point 3), and the low- to high-fidelity learners will be at approximately 80 s after about 300 trials (point 7). These are thus seven testable predictions shown with stars in Figure 8 that we hope to test and report in a later study.

Based on file Fig8.png, by Martin, 27mar22

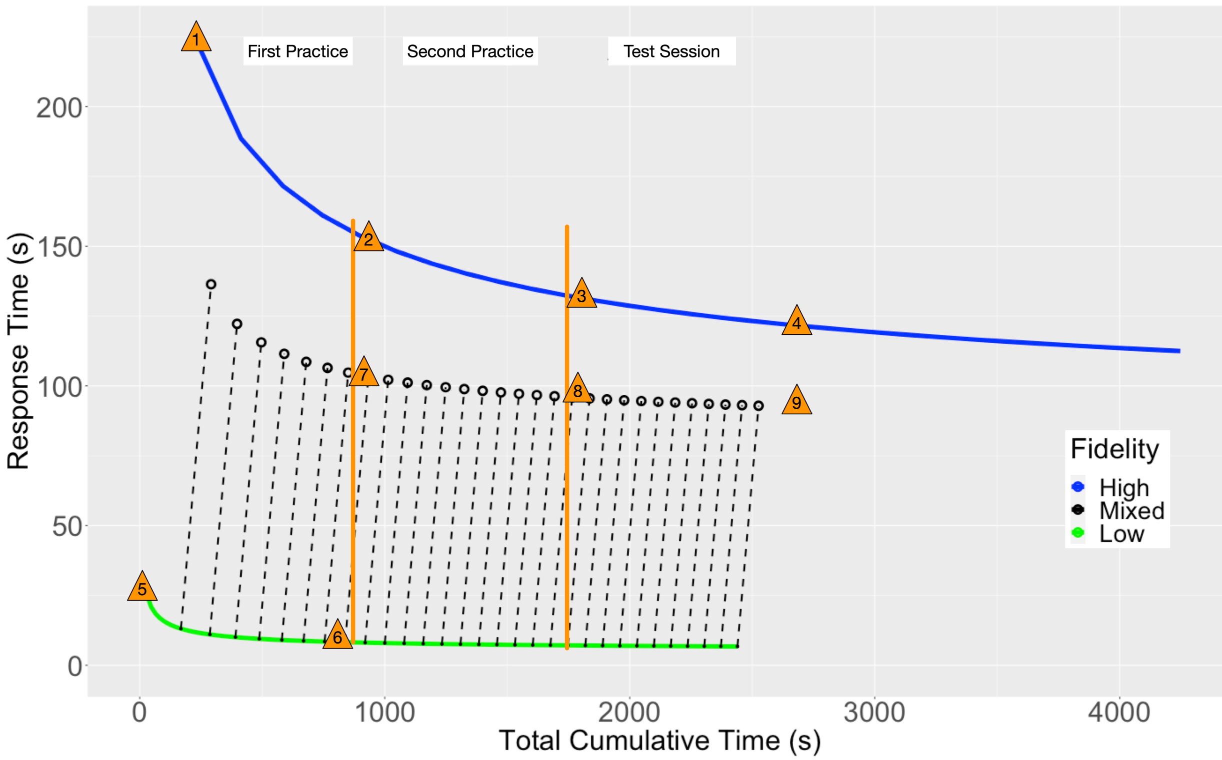

Figure 8. Testable predictions (shown with stars) of the response times at the beginning and end of the first 15 min. training session, and at the end of the second and third 15 min. training sessions. At 900 s the high-fidelity group will have had about 4 practice trials, and the low-fidelity group will have had about 87 trials.

Analysis 4: Effects of Varying Learning Rate

We next examine how sensitive this analysis is to learning rate. We generated plots for the three analyses above for learning rates from 0.1 to 1.2 at 0.1 intervals. The curves for 0.1 and 1.2 are shown in Figure 9. Changing the learning rate alters the curves' slopes and the relative performance. The changes in learning rate do not alter the outcomes about which interface has a relatively faster learning curve; it just alters how far apart they are and how quickly the two curves converge.

This analysis suggests that this analysis approach is relatively immune to differences in learning rate. In these three examples, when the low-fidelity simulation includes all the tasks and is faster, the result when moving to the high-fidelity simulation is faster than training only with the high-fidelity simulation. When the low-fidelity simulation leaves out some important subtask, it remains slower than using the high-fidelity simulation. And, for both learning rates, 0.1 and 1.2, it predicts that a study will find that the group trained initially with the low-fidelity simulation will end up faster than a group that uses the high-fidelity simulation only.

Learning rate: 0.1 **Learning rate: 1.2**

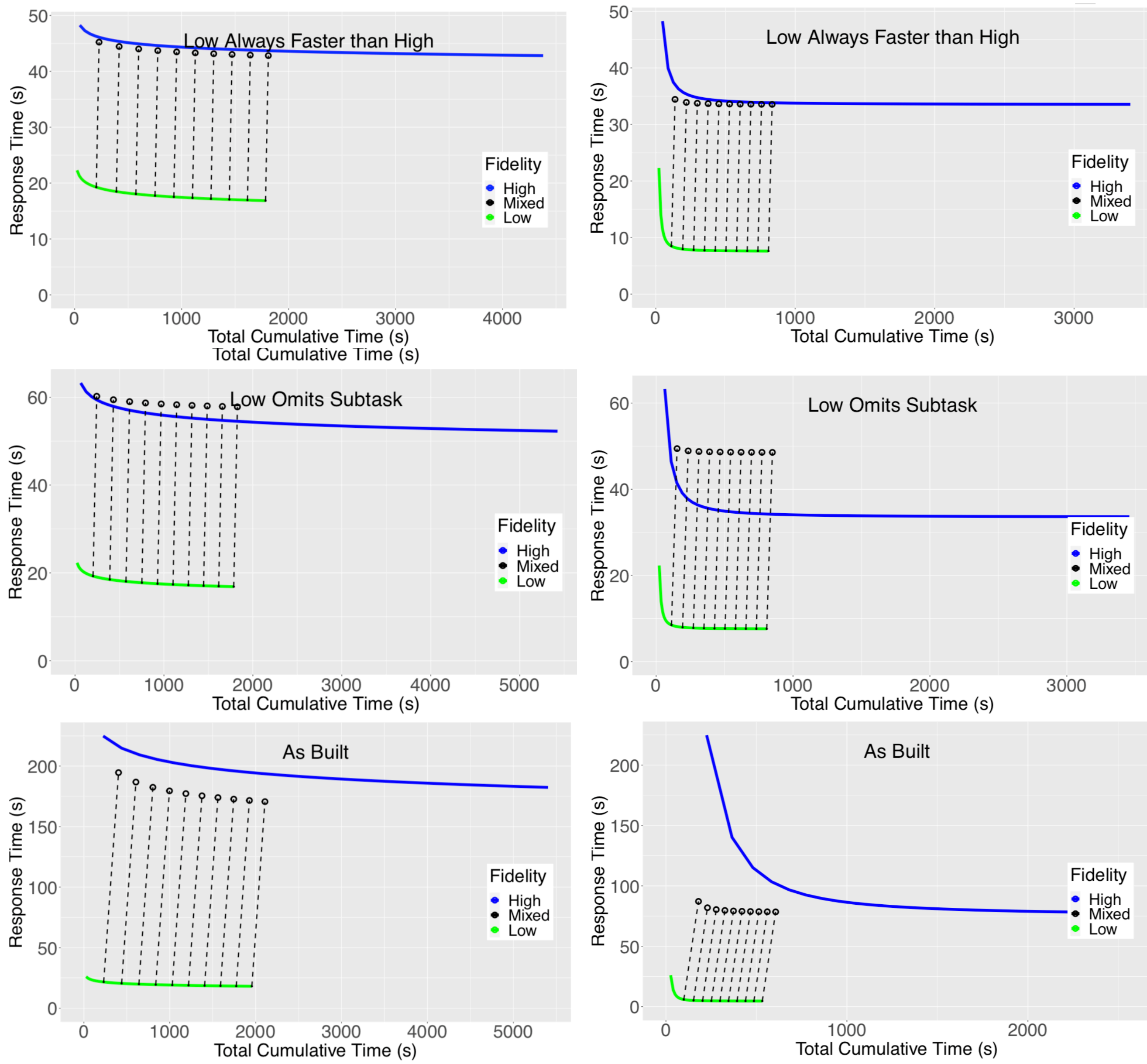

Figure 9. The three predictions at the extremes of the learning rates (0.1, left, and 1.2, right).

Screenshots taken from strudel-sim 26mar22.pdf

Discussion and Conclusion

We have shown through four analyses on a reusable task how the effect of fidelity on learning could be predicted using an approach combining task analysis and the learning curve. The results suggest that lower fidelity simulation trainers can lead to faster training and explain why. The analyses also show that the details of what is trained often matter—what is trained, what is taught in each simulation, what is not taught, what is already learned, how long each subtask takes, and how much training occurs on each training simulation. These analyses show that there are large effects of fidelity on learning. As a side product, we have developed a reusable task for studying learning, retention, and the effect of simulation fidelity.

This approach does not compute which approach to use, but it clarifies the amount of learning with each simulation and can help inform the discussion about how much fidelity to include. We note several contributions and then limitations and future work.

An approach for exploring the effect of fidelity on learning

This analysis can help explain why there are still discussions about whether to choose high- or low-fidelity simulations. The analysis is sensitive to different assumptions about time costs of the two training simulations and what is included in each simulation. The analysis shows that several factors influence the amounts of learning: previous training on a task, setup costs for a training task and other system response times (here, inserting a fault), what can be transferred, what can be trained, and how much training is required. Each will influence the learning curves and differences between levels of fidelity. This analysis points out that it is worthwhile to explicitly note and document what tasks are being trained in each simulation, how many repetitions learners are getting, and how long this will take.

If two people are discussing a training simulation, there is much room during discussions for their hidden assumptions to differ. This approach instead offers a calculus, a way for calculating and choosing how and why to use different levels of simulations. It can provide support for how much more training can be obtained from simulations of varying levels of fidelity.

This analysis approach could also be used to avoid the awkward situation of spending resources to make a simulation/training system more faithful to the external environment by including behavior that is not learned itself but because of the time it takes would none-the-less lead to learning less—tasks that do not get faster and do not get learned are cases where fidelity could be lowered. For example, in contrast to training aircraft, flight simulation trainers do not have to include pre-start checks, taxiing to the runway, flying, or landing—they can teleport the virtual aircraft to the runway for takeoff to practice different kinds of takeoffs with little delays between takeoff attempts.

What is also clear is that time to train is an important measure. When there is a lot of training time (i.e., a large number of training trials is available), a low-fidelity trainer does not offer as much benefit as when training time on the full system is limited. If a low-fidelity trainer is available, it might not just save time but also save money (or lives or equipment if the situation being trained is dangerous). Lower fidelity training simulations, if they cost less, can also lead to large learning gains even when transitioned to more complex tasks, and this has been seen before (Alluisi, 1991; Caro, Isley, & Jolley, 1973; Lesgold et al., 1992). It would be interesting to put those situations through these analyses.

In this maintenance task (i.e., Analysis 1 and 3), there does not appear to be a cost to starting low and going to the high-fidelity training simulation. Thus, these analyses show that initially learning on a low-fidelity simulation can save substantial time and resources, unless there are essential skills that are not in the low-fidelity simulator. Putting on a grounding strap, for example, if it was learnable and not taught in the low-fidelity simulator (i.e., Analysis 2), could have an important role in this analysis. It is also possible that the learnable skill is quickly learned. In this case, the learning curve may change to show benefits of the low-fidelity simulation.

Useful new graph, learning vs time instead of trials

We have introduced a new graph of response time vs. time on task. Figure 1 shows the traditional learning curve plotted by trial. Figure 2 shows the learning curve plotted by cumulative time, time on task. At least for analyzing training simulations, this graph can be useful when understanding how much fidelity to include in a simulation because training trials may take different amounts of time on different simulations. Additional work might include cost or risk to run those training trials as independent variables.

More repetitions are important early

The predictive analyses show that if you have only a short period to train and have a complete enough trainer, it may be better to have learners practice on a low-fidelity simulation than a high-fidelity simulation. Figure 4 shows that the low fidelity when transitioned back to high was faster than training only on high fidelity because the learner had more practice on what could be learned. There is a greater effect of more repetitions in the same period of time when there are not a lot of time to practice.

In contrast, at larger—perhaps extreme—amounts of practice (a large number of repetitions), learners on the low-fidelity simulation may not gain as much relative to the high fidelity simulation as they do at low practice time (less repetitions). The low-fidelity training is still faster, but the effect is smaller. In some situations this will still greatly matter (e.g., where differences in response time are important, such as adversarial tasks), and in some situations this will not matter (e.g., where doing the task correctly is most important, and speed is less important). The relative benefit will be based on which tasks are included in which trainer and relative ratio of practice repetitions.

Limitations

There are a few limitations to this analysis that in many cases will be related to future work to resolve these issues. The largest limitation is that this approach does not yet provide definitive answers about how much fidelity to include in a simulation. These analyses do not account for other differences in training systems such as cost; risk to the learner, environment, and equipment; and time to travel to the system. These are important considerations, and will have an important impact on training system choice.

The model of learning we used was simplistic. It assumes a single learning rate for all learners, subtasks, and for simulations. There is some evidence that learning rate can vary by subtask within a task (Kim & Ritter, 2016), by learner (Kim & Ritter, in prep.), and that learners will use different strategies to find faults (Friedrich & Ritter, 2020; Siegler, 1987). The analysis approach could be expanded to include an appropriate learning rate for each of these aspects (including a greater understanding of learning rates across a wider range of tasks), along with a way to summarize and display the results. The analyses of the effect of learning rates suggests that this will not generally alter the qualitative results, but may alter the relative differences. These differences may be important in some cases. Also, in this analysis the lower-fidelity trainer took less time per task. This analysis shows that the trainer with the lower task time will be favored, not that it specifically be lower fidelity. But, typically, reducing task time will improve training by increasing repetitions, and typically this is seen in lower fidelity simulations. Also, time on task is generally helpful, but there are exceptions to this rule that can be found where the practice is not deliberate (e.g., Ericsson, Krampe, & Tesch-Roemer, 1993) or the learner needs help finding better strategies.

The learning rates might particularly be different between the two systems because of their complexity. Sweller (1988, 2007) suggests that high working memory demands when learning will decrease learning speed. Thus, high fidelity simulations, which are more complicated, may also reduce the learning speed. On the other hand, some learners may like the high-fidelity simulators more, and take them more seriously or put more effort into them. Learners and instructors do, at times, confuse learning and liking (e.g., Moreno & Mayer, 2002). The balance of these two factors is currently hard to predict.

There may also be differences between novices and experts that will be difficult to represent in this approach based on strategy differences. A low-fidelity simulation may help novices become experts if novices would otherwise focus on the wrong aspects of the full task that would be available in a high-fidelity simulation, or hinder their development if cues that are needed are not provided. Similarly, for experts, the low-fidelity simulation may remove cues they are used to relying upon (such as audio cues) that are not required for the task but have become associated with the process, and thus the low-fidelity training may hinder expert performance. This aspect must be kept in mind during the analyses.

The curves make several assumptions that could be explored further. Perhaps the largest assumption was about which parts of the subtasks were essentially already learned (extensive practice at least), and which parts of the subtasks would be learned. The analyses probably would not change much if these are adjusted or if the learning equations were extended to represent this concept not as fixed and learnable, but as previous practice on each subtask.

This analysis was of a single task with a uniform learning rate. Training is made up of numerous tasks, and they probably have different learning rates. For our example system, we analyzed one fault. The fault we analyzed is probably fairly representative of the possible 35 faults. A more accurate analysis would include all of the tasks, and in our case, all of the faults, as well as situations where there are multiple faults. Other system's learning tasks will be different than ours. In both cases, a more complete analysis may require a tool to help compute these costs.

Future Work

The most tangible future work is that we are running a study of the training simulations in Analysis 3. The results of that study will allow us to test the predictions in Figure 8.

In many cases, the most impact from this approach is to allow the ability to perform what-if analyses. In some cases, the learning rates will have to be measured or estimated from similar learners, tasks, simulations, and strategies.

It would be useful to make this analysis approach even easier to use. It could then be used to analyze more realistic, complex tasks, for example, as IMPRINT does for basic task analysis (Booher & Minninger, 2003). This analysis could include the costs of building the additional low-fidelity simulation. This tool could even go so far as to predict the cost of each component in the low-fidelity simulation (e.g., building it out could cost a little more but lead to greater learning savings, system saving, or system effectiveness). This approach can also be informed by tools to model users in simulations automatically (John & Jastrzembski, 2010; Pew & Mavor, 2007; Ritter, 2019; Wallach, Fackert, & Albach, 2019), and could be potentially included in them. One would like this analysis to be in a tool, more like IMPRINT than Excel, and perhaps able to store, compare, and plot the results.

APPLE (Vogel-Walcutt, 2010) was designed and prototyped to support analyzing this type of design question. The FAST algorithm is supporting system designs (i.e., more than just the software) that are more resistant to fatigue (Hursh et al., 2004; van Dongen, 2004), and the Air Force is creating similar tools to predict the effect of practice schedules on learning (Gluck, Jastrzembski, & Krusmark, 2019; Jastrzembski, Addis, Krusmark, & Gluck, 2010; Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018).

This work may ultimately lead to a better method to determine optimal simulation training time based on predicting performance improvement using this analysis method, supported in a tool that makes it easier to perform.

References

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

Alessi, S. M., & Trollip, S. R. (1991). *Computer-based instruction: Methods and development*. Englewood Cliffs, NJ: Prentice-Hall.

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.

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

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press.

Batran, K. E., & Dunlop, M. D. (2014). Enhancing KLM (Keystroke-Level Model) to fit touch screen mobile devices. In *MobileCHI*, 283–286.

Booher, H. R., & Minninger, J. (2003). Human systems integration in Army systems acquisition. In H. R. Booher (Ed.), *Handbook of human systems integration* (pp. 663-698). Hoboken, NJ: John Wiley.

Burns, M., Ritter, F. E., & Zhang, X. (2016). Using naturalistic typing to update architecture typing constants. In *Proceedings of ICCM, 14th International Conference on Cognitive Modeling (ICCM 2016)*, 169-175. University Park, PA: Penn State.

Card, S. K., Moran, T., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Erlbaum.

Card, S. K., Moran, T. P., & Newell, A. (1980). The keystroke-level model for user performance time with interactive systems. *Communications of the ACM, 23*(7), 396-410.

Caro, P. W., Isley, R. N., & Jolley, P. B. (1973). *Research on synthetic training: Device evaluation and training program development* (Technical Report 73-20). Fort Rucker, AL: Human Resources Research Organization (HumRRO), Division No. 6 (Aviation).

Charvat, G. L. (2011). MIT IAP 2011 laptop based radar: Block diagram, schematics, bill of material, and fabrication instructions. from https://ocw.mit.edu/resources/res-ll-003-build-a-small-radar-system-capable-of-sensing-range-doppler-and-synthetic-aperture-radar-imaging-january-iap-2011/

Crossman, E. R. F. W. (1959). A theory of the acquisition of speed-skill. *Ergonomics, 2*, 153-166.

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.

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9*(1), 1-8.

Ericsson, K. A., Krampe, R. T., & Tesch-Roemer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363-406.

Friedrich, M. B., & Ritter, F. E. (2020). Understanding strategy differences in a diagrammatic reasoning task. *Cognitive Systems Research, 59*, 133-150.

Gluck, K. A., Jastrzembski, T. S., & Krusmark, M. A. (2019). Prospective comments on performance prediction for aviation psychology. In M. A. Vidulich & P. S. Tsang (Eds.), *Improving aviation performance through applying engineering psychology* (pp. 80-99): CRC Press.

Gong, R., & Kieras, D. E. (1994). A validation of the GOMS model methodology in the development of a specialized, commercial software application. In *Proceedings of CHI, 1994*, 351-357. New York, NY: ACM.

Havinghurst, L. C., Fields, L. E., & Fields, C. L. (2003). High versus low fidelity simulations: Does the type of format affect candidates' performance or perceptions? In *27th Annual IPMAAC Conference on Personnel Assessment: Exploring New Horizons in Assessment*.

Hursh, S. R., Redmond, D. P., Johnson, M. L., Thorne, D. R., Belenky, G., Balkin, T. J., et al. (2004). Fatigue models for applied research in warfighting. *Aviation, Space, and Environmental Medicine, 73*(3), A44-A53.

Jastrzembski, T. S., Addis, K., Krusmark, M., & Gluck, K. A. (2010). Prediction intervals for performance prediction. In D. D. Salvucci & G. Gunzelmann (Eds.), *Proceedings of ICCM - 2010- Tenth International Conference on Cognitive Modeling* (pp. 109-114).

John, B. E., & Jastrzembski, T. S. (2010). Exploration of costs and benefits of predictive human performance modeling for design. In *Proceedings of the 10th International Conference on Cognitive Modeling*, 115-120. Philadelphia, PA: Drexel University.

John, B. E., & Kieras, D. E. (1996). The GOMS family of user interface analysis techniques: Comparison and contrast. *ACM Transactions on Computer-Human Interaction, 3*(4), 320-351.

John, B. E., Prevas, K., Salvucci, D. D., & Koedinger, K. (2004). Predictive human performance modeling made easy. In *Proceedings of CHI 2004 (Vienna, Austria, April 2004)*, 455-462. New York, NY: ACM.

Katsanos, C., Karousos, N., Tselios, N., Xenos, M., & Avouris, N. (2013). KLM Form Analyzer: Automated evaluation of web form filling tasks using human performance models. In *Human-Computer Interaction – INTERACT 2013. INTERACT 2013. Lecture Notes in Computer Science, vol 8118*, 530-537. Berlin, Heidelberg: Springer.

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning how to operator a device. *Cognitive Science, 8*, 255-273.

Kim, J. W., & Ritter, F. E. (2016). Microgenetic analysis of learning a task: Its implications to cognitive modeling. In *Proceedings of the 14th International Conference on Cognitive Modeling (ICCM 2016)*, 21-26. University Park, PA: ACS Lab, Penn State.

Kim, J. W., & Ritter, F. E. (in prep., September 2021). A microgenetic analysis of subtask learning using a Bayesian hierarchical modeling approach: Subtasks are learned at different rates.

Lesgold, A. M., Lajoie, S., Bunzon, M., & Eggan, E. (1992). SHERLOCK: A coached practice environment for an electronics troubleshooting job. In J. Larkin, R. Chabay & C. Scheftic (Eds.), *Computer assisted instruction and intelligent tutoring systems: Establishing communication and collaboration*. Hillsdale, NJ: Erlbaum.

Liu, D., Macchiarella, N. D., & Vincenzi, D. A. (2009). Simulation fidelity. In *Human factors in simulation and training* (pp. 61-73): Taylor & Francis.

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.

Moreno, R., & Mayer, R. E. (2002). Learning science in virtual reality environments: Role of methods and media. *Journal of Educational Psychology, 94*.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-51). Hillsdale, NJ: Erlbaum.

Noble, C. (2002). The relationship between fidelity and learning in aviation training and assessment. *Journal of Air Transportation, 7*, 34–54.

Pew, R. W., & Mavor, A. S. (Eds.). (2007). *Human-system integration in the system development process: A new look*. Washington, DC: National Academy Press. books.nap.edu/catalog/11893, checked May 2019.

Prophet, W. W., & Boyd, H. A. (1970). *Device-task fidelity and transfer of training: Aircraft cockpit procedures training* (Technical Report 70-10). Fort Rucker, AL: Human Resources Research Organization (HumRRO), Division No. 6 (Aviation).

Ritter, F. E. (2019). Modeling human cognitive behavior for system design. In S. Scataglini & G. Paul (Eds.), *DHM and posturography* (pp. 517-525). London: Academic Press.

Ritter, F. E., & Bibby, P. A. (2008). Modeling how, when, and what is learned in a simple fault-finding task. *Cognitive Science, 32*, 862-892.

Ritter, F. E., & McDermott, A. F. (2020). The effect of task fidelity on learning curves. In *Proceedings of the 18th International Conference on Cognitive Modeling (ICCM 2020)*, 229-235.

Ritter, F. E., Schoelles, M. J., Quigley, K. S., & Klein, L. C. (2011). Determining the number of model runs: Treating cognitive models as theories by not sampling their behavior. In L. Rothrock & S. Narayanan (Eds.), *Human-in-the-loop simulations: Methods and practice* (pp. 97-116). London: Springer-Verlag.

Ritter, F. E., & Schooler, L. J. (2001). The learning curve. In W. Kintch, N. Smelser & P. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 13, pp. 8602-8605). Amsterdam: Pergamon.

Ritter, F. E., Tehranchi, F., Brener, M., & Wang, S. (2019). Testing a complex training task. In *Proceedings of the 17th International Conference on Cognitive Modeling (ICCM 2019)*, 184-185.

Ritter, F. E., Tehranchi, F., & Oury, J. D. (2019). ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 10*(3), Paper e1488.

Rosenbloom, P. S., & Newell, A. (1987). Learning by chunking, a production system model of practice. In D. Klahr, P. Langley & R. Neches (Eds.), *Production system models of learning and development* (pp. 221-286). Cambridge, MA: MIT Press.

Seibel, R. (1963). Discrimination reaction time for a 1,023-alternative task. *Journal of Experimental Psychology, 66*(3), 215-226.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology, 115*, 250-264.

Snoddy, G. S. (1926). Learning and stability. *Journal of Applied Psychology, 10*, 1-36.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257-285.

Sweller, J. (2007). All is in order. In F. E. Ritter, J. Nerb, E. Lehtinen & T. O'Shea (Eds.), *In order to learn: How the sequences of topics affect learning* (pp. 215-224). New York, NY: Oxford University Press.

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.

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.

van Dongen, H. P. A. (2004). Comparison of mathematical model predictions to experimental data of fatigue and performance. *Aviatation, Space, and Environmental Medicine 75*(3, Section II), A15–36.

Vogel-Walcutt, J. (2010). *APPLE, Algorithms Physiologically-derived to Promote Learning Efficiency*. Orlando, FL: U. of Central Florida.

Wallach, D. P., Fackert, S., & Albach, V. (2019). Predictive prototyping for real-world applications: A model-based evaluation approach based on the ACT-R cognitive architecture. In *DIS '19: Proceedings of the 2019 on Designing Interactive Systems Conference*, 1495–1502.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the theoretical adequacy and applied potential of computational models of the spacing effect. *Cognitive Science, 42*, 644-691.