Assessment of a novel virtual environment for examining human cognitive-motor performance during execution of action sequences

Alexandra A. Shaver[1,*], Neehar Peri[2,*], Remy Mezebish[2], George Matthew[2], Alyza Berson[1], Christopher Gaskins[3]; Gregory P. Davis[2], Garrett E. Katz[4]; Immanuel Samuel[5], Matthew J. Reinhard[5], Michelle E. Costanzo[5], James A. Reggia[2,3,6,7], James Purtilo[2,§], Rodolphe J. Gentili[1,3,7,§]

[1] Department of Kinesiology, University of Maryland, College Park, MD, USA.
[2] Department of Computer Science, University of Maryland, College Park, MD, USA.
[3] Neuroscience & Cognitive Science Program, University of Maryland, College Park MD, USA.
[4] Electrical Engineering & Computer Science, Syracuse University, Syracuse, NY, USA
[5] War Related Illness and Injury Study Center, Washington DC VA Medical Center, DC, USA
[6] Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA
[7] Maryland Robotics Center, University of Maryland, College Park, MD, USA.

[*] Co-first authors
[§] Corresponding authors: purtilo@umd.edu; rodolphe@umd.edu

Abstract. The examination of neural resource allocation during complex action sequence execution is critical to understanding human behavior. While physical systems are usually used for such assessment, virtual/remote systems offer other approaches with potential benefits such as remote training/evaluation. Here we describe a virtual environment (VLEARN) operated via the internet that has been developed to study the cognitive-motor mechanisms underlying the execution of goal-oriented action sequences in remote and laboratory settings. This study aimed to i) examine the feasibility of evaluating human cognitive-motor behavior when individuals operate VLEARN to complete various tasks; and ii) assess VLEARN by comparing its usability and the resulting performance, mental workload, and mental/physical fatigue during virtual and physical task execution. Results revealed that our approach allowed human cognitive-motor behavior assessment as the tasks completed physically and virtually via VLEARN had similar success rates. Also, there was a relationship between the complexity of the virtual control systems and the dependency on those to complete tasks. Namely, relative to controls with more functionalities, when VLEARN enabled simpler controls, above average usability and similar levels of cognitive-motor performance for both physical and virtual task execution were observed. Thus, a simplification of some aspects of the VLEARN control interface should enhance its usability. Our approach is promising for examining human cognitive-motor behavior and informing multiple applications (e.g., telehealth, remote training).

Keywords: Virtual environment, Action sequences, Mental workload, Cognitive-motor performance, Human-machine interface, rehabilitation.

1 Introduction

The ability to efficiently recruit neural resources to face varying task demands is critical in driving the underlying mental workload and performance dynamics ultimately enabling adaptive cognitive-motor behavior [1-4]. For instance, an increase in task demands would result in greater engagement of the corresponding neural resources, ultimately causing an elevation in mental workload. Objective indicators of mental workload may be helpful to several applications such as evaluating heterogenous neurological conditions with poorly understood etiology such as Anomalous Health Incident sequalae and post concussive syndrome experienced by military populations, since they capture both behavioral and cognitive performance. The concept of mental workload has been largely studied via various tasks (e.g., single reaching movements, action sequences, dual-tasking) in both physical and/or virtual environments. Despite this large body of work, mental workload is not well understood in the context of performance of complex action sequence tasks which typically: i) generate high cognitive-motor demands in novices (e.g., high-level planning; working memory; attention; [5]); ii) require a substantial amount of practice to be mastered, iii) involve several degrees of freedom while requiring substantial hand-eye coordination, and iv) are a good vehicle to study human behavior in more real-world conditions [6].

Recently, a limited number of studies have examined the changes in mental workload when individuals execute or practice such complex actions sequences [7-9]. These studies generally examined this notion when individuals performed the task using physical systems, except for the well-established Tower of Hanoi task. However, it is also important to further examine performance and mental workload concurrently during the execution of various types of complex tasks in virtual environments operated remotely rather than in a controlled lab setting or with physical equipment. Virtual environments provide individuals the opportunity to be assessed or trained remotely from a more convenient and possibly safer location when in-person assessment or training can be very challenging or dangerous. As such, the current study is important not only to further understand the cognitive-motor mechanisms that support virtually executing complex action sequences to solve a problem but also to inform applications related to telemedicine, telehealth, and training/re-training of civilians and military personnel all over the world. A need for effective, scalable and efficient remote options is crucial during special circumstances such as the COVID-19 pandemic which has forced many institutions globally to delay and/or halt in-person human data collection and further complicated interventions, evaluation and training during an uncertain time for many individuals (e.g., social distancing, wearing masks), problems that a tool, such as the one proposed here, could by-pass via remote operations. A possible first step to enabling the examination of cognitive-motor processes remotely is to develop a new experimental medium that is flexible, cost-effective, easy to use, and mimics physical systems with the level of fidelity needed to provide accurate and meaningful experimental data. As such, in an attempt to fulfill these requirements, our research group has been developing a new virtual environment (named virtualized learning or VLEARN[1])

[1] https://github.com/gmatthew1141/VLEARN

that is accessible remotely via the internet and allows for observing human behavior and neural mechanisms when individuals perform and learn action sequences to successfully execute complex cognitive-motor tasks. So far, VLEARN can simulate multiple scenarios intended to assess cognitive-motor performance and learning during execution of complex action sequences. However, before employing this new virtual platform to conduct human studies that include experimental manipulations, its usability and more generally its effects on cognitive-motor behavior need to be examined.

Thus, this study aims to i) determine the feasibility of experimentally assessing human cognitive-motor behavior (performance, mental workload and fatigue) when individuals operate this novel virtual environment, via the internet, to complete various tasks by executing specific action sequences, and ii) if so, to assess the usability of this novel remote virtual platform and determine which features are appropriately designed and which need improvements. In particular, VLEARN's assessment was conducted by comparing the level of usability, performance, mental workload and fatigue obtained when individuals operated matching virtual and physical systems.

We hypothesized that if the proposed approach is feasible, VLEARN will enable individuals to complete the various virtual tasks while investigators successfully collect metrics related to performance, mental workload and fatigue with data quality similar to using the physical systems. Alternatively, any major limitations of the proposed method (e.g., technical glitches; excessive computational delays, etc.) that compromise the data integrity (e.g., participants drop-out, incomplete data set for the task and/or metrics) would suggest that such an approach is currently not feasible. Also, under the assumption that the proposed approach is feasible, we hypothesized that if the fidelity of the virtual environment is appropriate, measurements indexing the usability, performance, mental workload and fatigue should not differ compared to measurements collected when individuals use the corresponding physical systems.

2 Material and methods

2.1 The virtual environment

2.1.1 General presentation

The VLEARN application serves as a participant and experiment development and management tool, containing both a dashboard and a trial completion page. The dashboard provides a hub where participants can log in and view their assigned experimental tasks to perform, while allowing administrators to create new participant accounts, create tasks/trials, and manage experimental data. Through the trial completion page, participants interact with the virtual environment to complete a task which was designed by the administrator beforehand. On this page, participants have a control readout, as well as a window, from which they complete tasks within a virtual 3D environment. Although VLEARN allows the administrator to design various scenarios, currently three tasks of interest have been developed: i) the well-known Tower of Hanoi task (ToH; [7-11]); ii) a disk hard drive dock maintenance (DM) task where drives were manipulated similarly to prior studies [12-15] and iii) a pipe system maintenance (PM)

task (see Section 2.2 for details). These tasks were created using the Unity game engine, and currently support multi-user collaboration for up to four users, allowing multiple administrators/participants to join in task completion. While not tested here, the RESTful API also provides a system to create autonomous virtual agents to collaboratively execute tasks with a human (Fig. 1).

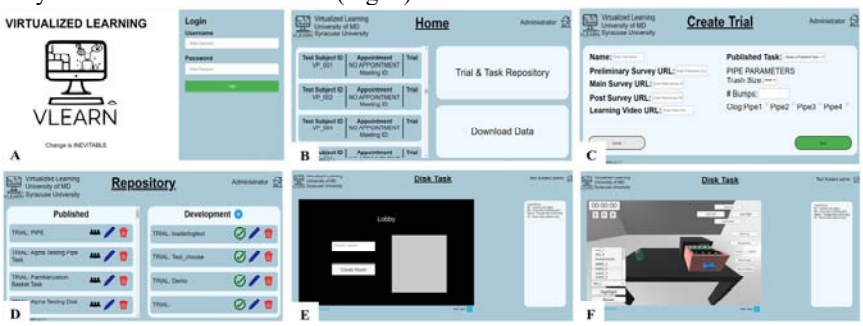

Fig. 1. Administrator's view of VLEARN. (A) Login page for administrator (and participants); (B) Administrator's homepage for visualization of participant list, datasets, trial/task repository. Administrator can: (C) select details of a task to assign to participants; (D) publish trials/tasks that can be (un)assigned, edited or deleted (left) and trials/tasks still being created/edited (right); (F) manipulate the environment before each trial and highlight elements of a task (here DM task).

VLEARN builds on a previously created virtual environment called the Simulator for Maryland Imitation Learning Environment (SMILE) developed at the University of Maryland - College Park [16,17]. SMILE is a Java based simulator for studying imitation learning. An experimenter uses SMILE to create animated demonstrations that can then be observed by robots as they learn to imitate what is being demonstrated. SMILE hypothesizes that robots can learn more effectively by ignoring the demonstrator's motions and instead only observing the behaviors of the object in the demonstration environment. SMILE allows video playback, text logging, and scene creation using XML. In addition, it allows for a simulated robot to interact with the Java environment through MATLAB (for details see [16,17]). Although interesting, SMILE serves a different purpose than the VLEARN software, which allows us to study high-level learning behaviors in humans. Moreover, our proposed solution implements multi-agent interaction, and allows for us to control agents using a flexible web API.

The long-term goal of VLEARN is to facilitate high fidelity research into human learning by reducing confounding factors, improving the movement and interaction system, and providing extensive experimental manipulation options with enhanced data collection and logging capabilities. To this end, we modeled the Unity assets to closely approximate a real environment. Specifically, we ensured that interactions with objects in a scene were realistic, while also considering the limited degrees of freedom provided by mouse and keyboard input. We discretized the movement system such that an agent can only travel to fixed positions in the virtual world using teleporter pads. Moreover, we translated almost all keyboard interaction into clickable buttons on the software interface to facilitate manipulation of the virtual environment. This is important to ensure

that even participants who are not used to manipulating virtual environments (e.g., computer graphics, software design; gamers) can still operate VLEARN fairly easily. This ensures that learning effects are mitigated, thus avoiding introducing experimental bias. It also ensures that a number of potential individuals can participate simultaneously in cognitive-motor studies using this virtual system. Lastly, we expanded the logging capabilities to facilitate multi-agent interactions. We defined an object-oriented relationship between all objects and agents in a scene, facilitating important research insights about the order of object interactions.

2.1.2 Software architecture

The VLEARN front-end was developed in Node.js, while the virtual world was implemented in Unity and rendered through WebGL. The dashboard connects the administrator and the study participant to the simulation or the participant information. The simulations are stored on a web server with the corresponding XML files to generate the task. VLEARN synchronously logs each participant's interactions with the environment. All experimental data are stored in a MongoDB database, and each participant is only identified by a unique hash in accordance with Institutional Review Board standards. The administrator prepares trials by uploading an XML file to the web server.

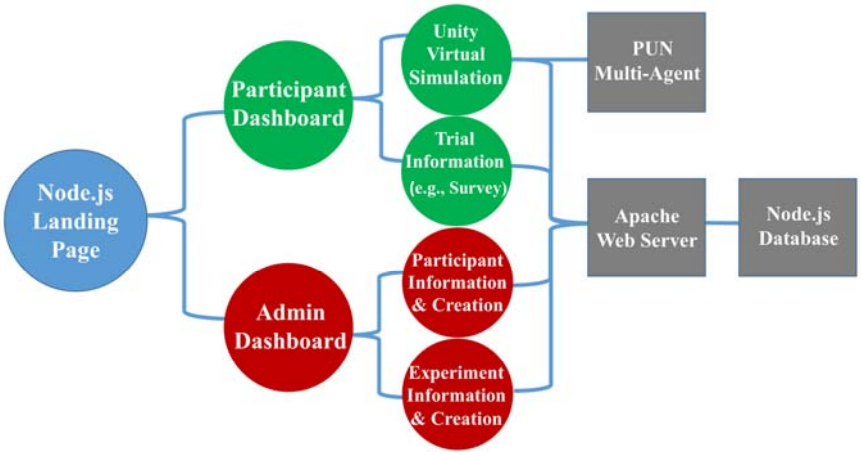

Fig. 2. High-level architecture of VLEARN. The circles and squares represent the front- and back-end while administrator and participant actions are denoted by red and green circles, respectively.

These XML files contain the information needed to generate each task and serve as a convenient way to save the starting layout of a task (see Fig. 2 for the high-level architecture of VLEARN). A RESTful API in Node.js was also implemented to allow for external logging of interactions with participants and objects (e.g., toggle, disk, magnetic pointer) through events (e.g., “press”, “hold”, “drop”, “hide”). Events contain environment or agent observations when querying for information. All events contain an agent identifier, Cartesian coordinates and Euler angle to show position and rotation

of interactable objects in a scene. Some events additionally contain an object identifier which uniquely identifies the actions taken related to a given object in the log. As such, events can be tracked, saved, and replayed as a sequence of events or interactions. For example, in DM task when the participant inserts a drive into the slot, the corresponding event would contain the participant's identifier, current position and the drive's identifier. Additionally, the RESTful API can be queried for only interactions involving a specific object or user. These event lists could be used in the place of manually entered behavioral data for quick analysis of sequence efficiency and eventually for real-time feedback. These events originate from Unity's Photon Unity Network (PUN V2), a real-time multiplayer game development framework. The PUN cloud service enables up to 20 concurrent users maximum on worldwide servers.

2.2 Experimental evaluation

2.2.1 Participants

Twelve healthy individuals participated in this study (2 men and 10 women; age range 19 - 33 years). No history of neurological impairment or use of medication known to alter the central nervous system was reported. Participants had a normal or corrected-to-normal vision and were free of drug and alcohol use at the time of the study. Prior to starting the study all individuals provided their written informed consent which was approved by the University of Maryland-College Park Institutional Research Board.

2.2.2 Experimental tasks

To assess the proposed virtual environment, three tasks were considered: i) a modified version of the well-known Tower of Hanoi (ToH) task which consists of moving disks on three pegs (e.g., [7-11]); ii) a disk hard drive dock maintenance (DM) task where faulty hard drives need to be removed and replaced by functioning ones [12-15] and iii) a pipe system maintenance (PM) task where clogged pipes had to be cleaned by removing an obstructive object. These three tasks were completed over two testing sessions where participants manipulated a physical system as well as a matching virtual system implemented through VLEARN (see Fig. 3) via the internet. Thus, VLEARN was assessed by examining to what extent its usability and the resulting human performance during execution of these tasks differed from using the physical systems. These three tasks were employed since their completion i) required performing action sequences involving cognitive-motor processes and ii) offered a fairly good range of functionalities to control the interface (i.e., increased functionalities from ToH to DM to PM). However, a larger set of functionalities does not necessarily mean that the task demands were higher. Importantly, the aim of this study was not to manipulate task demands to probe the engagement of cognitive-motor resources (e.g., attention, high-level planning) by having participants complete tasks using scenarios of increasing complexity. Instead, as a first step, this work mainly aimed to study the feasibility of remotely assessing cognitive-motor behavior with the proposed approach and the usability of VLEARN when participants operated it to perform these three tasks.

2.2.2.1 Tower of Hanoi task

Typically, the ToH task consists of several disks stacked in ascending order of diameter on one of three identical, evenly-spaced pegs. The physical ToH system was composed of a wooden board with three pegs and wooden disks. In the virtual system, this task was executed using a classic point-and-click control system to manipulate the disks and as such served as a standard approach for potential subsequent comparisons (see Fig. 3; first column). The objective of the original ToH task is to move all disks from the leftmost to the rightmost peg while following three rules: a) only one disk can be displaced at a time, b) a disk may not be placed on the table or held while another disk is being moved, and c) a larger disk cannot be stacked on top of a smaller disk. As mentioned earlier, since this work aimed to assess the proposed virtual environment, participants were asked to perform a modified version of the ToH with three disks and three pegs using the physical and virtual system. Namely, participants were asked to move the disks from the leftmost peg to the middle peg and finally to the rightmost peg or back to the leftmost peg. Trials were deemed successful when the task goal was completed while the two first rules mentioned above (i.e., only one disk can be displaced at a time; a disk may not be placed on the table or held while another disk is being moved) were respected.

2.2.2.2 Disk drive dock maintenance task

The DM task has been used in prior research to examine high-level plan generation in a humanoid robot and humans during action sequence imitation [13,14]. This task involves a mock-up hard drive docking station with a drawer that, when opened, allows participants to manipulate four hard drives placed in individual slots, each being associated with a LED indicator and a toggle switch. LED indicators were either red, green, or off designating that the associated drive was broken, working properly, or had been turned off, respectively. The physical system was a custom-made mock-up controlled by an Arduino processor [13-15] and used as a model for the virtual system (see Fig. 3; second column). The goal of the task was to safely replace the faulty drive with a new drive. Trials were considered successful if the goal was attained while following the rule that the LED had to be turned off when a drive is added or removed. Although multiple possibilities could be considered to challenge individuals, for the reasons previously mentioned, participants only had to replace one faulty drive.

2.2.2.3 Pipe maintenance task

Although the PM task was initially designed to examine cognitive-motor processes that involve both cognitive (e.g., high-level planning) and motor (e.g., fine motor precision) demands, here, akin to the two tasks mentioned above, fairly simple task sequences involving all key components were used to assess the usability of the proposed virtual environment compared to its physical counterpart. In both environments, this task involved a mock-up pipe station with a main valve which had to be closed before any of the other four PVC pipes could be safely manipulated. The main valve and the four pipes were each associated with a toggle switch and LED indicator that were red,

green, or off designating whether the water had stopped (i.e., closed main valve or clogged pipe), was working properly, or had been turned off, respectively. The pipes could be opened or closed by replacing or removing the PVC adaptor (top used to cover pipe opening). A magnetic tool was used to extract the obstructive object from the affected pipe. A red LED and small buzzer were triggered to alert participants that the object had touched the edge of the physical pipe. Similarly, during virtual trials the object would turn red if it came into contact with the pipe during extraction. Virtual extraction was accomplished by controlling the virtual tool with a computer mouse or trackpad (Fig. 3; third column) depending on what the participant used at home. This task aimed to safely clear a clogged pipe using a tool to extract the object without letting it touch the edge of the pipe. Trials were considered successful when the task goal was reached while two rules were respected: a) the main valve had to be closed (red) before any pipes can be opened; b) the corresponding LED had to be turned off before a pipe can be opened.

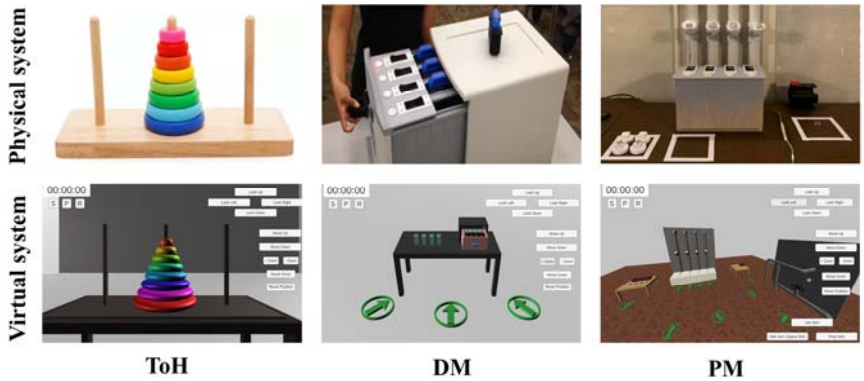

Fig. 3. The physical and virtual systems employed to execute the ToH, the DM and PM tasks to assess VLEARN usability and human cognitive-motor performance when performing with this simulator. ToH: Tower of Hanoi; DM: Disk drive dock maintenance; PM: Pipe maintenance.

2.2.3 Experimental procedures

Participants were required to perform the three (ToH, DM and PM) tasks over two testing sessions using in-person physical systems and the virtual environment, VLEARN, controlled through the internet. For the latter, participants were given an individual username and password to access the VLEARN website. Order of the testing sessions, modality and tasks were counterbalanced. First, for each task (i.e., ToH, DM and PM) and both execution modalities (i.e., physical and virtual) a familiarization phase was conducted to ensure that the participants understood the tasks and how to use the current system. Specifically, the rules and goal of the task were explained, and participants were provided with up to five minutes to manipulate the system they would be using for the upcoming trials. To mitigate practice effects, the action sequences used during the testing session were not used during this phase. With the PM task, participants were not allowed to attempt any actual extraction but instead only touched the edge of the pipe with the pointer. This allowed participants to explore the collision

feedback (i.e., LED, red object color change, buzzer) with both the physical and virtual systems without practicing the actual extraction. Once this familiarization period was completed, for each task and execution modality participants had to perform four blocks of four trials resulting in a total of 16 trials per task. To vary the conditions and ensure use of all system components, the task was slightly different for each block (i.e., the peg to transfer the disks, the disk drive to replace and the pipe to clean differed between blocks, respectively). At the beginning of each block, a video demonstration of the action sequence (i.e., the reference sequence) to perform was presented to the participants. Participants were permitted to ask about the rules of the task and for the next step of the sequence at any time during a trial and see again the video demonstration between trials at their request. This approach was employed to ensure that the usability and the cognitive-motor states examined here were primarily related to operating the virtual or physical system, not the engagement of cognitive-motor resources (e.g., working memory; attention; high-level planning processes) due to demands related to over-complicated action sequences. Individuals were allowed to start their trial whenever they wanted after a verbal 'Go' signal and trials were stopped as soon as the task was completed or the time limit of two or five minutes was reached, whichever came first for the physical and virtual trials, respectively. These limits were set such that both virtual and physical trials potentially allowed to collect the same number of trials for direct comparison while their respective sessions lasted a maximum of about two hours[2]. Any session going beyond the session time limit was stopped. Trials were considered successful if participants completed the action sequence reaching the goal within the time limit while following the corresponding rules. Trials with burdensome or excessive technical glitches (e.g., frozen screen, system component not working properly, internet connection issues) were halted and restarted. For each trial, the performance of the participant was video recorded for subsequent data processing.

After 16 trials with the physical or virtual system, individuals were asked to complete the System Usability Scale (SUS) to determine the perceived usability of the systems used for this task. While other options exist, the SUS is a widely accepted measure of perceived usability that is versatile, cost and time effective, as well as, robust even when small sample sizes are used [18,19]. Also, it has been used successfully to estimate the usability of many software systems, devices, services and is related to internet self-efficacy [18,19] . The SUS consists of 10 questions which alternate between positive and negative statements about the system. Answers are recorded using a five-point scale [20].

After each block of four trials, participants completed questionnaires to assess their level of perceived workload and fatigue. The NASA TLX is a well-established multidimensional questionnaire used to report different aspects of perceived workload during cognitive-motor performance [4,21]. NASA TLX scores are generally consistent with more objective measurements, such as those obtained via neuroimaging (e.g., [3]). Workload is assessed along six subscales: mental, physical, temporal, perceived performance, effort, and frustration (ranging from 0 to 100 in increments of 5) (for details,

[2] The 2 mins time limit for the physical trials was set from prior work which clearly established that it was largely enough for task completion and thus did not bias the present study.

see [21]). Although all the subscales of the NASA TLX were examined here, the mental demand dimension was of primary interest since it has been shown as the most representative of the mental workload (e.g., [3,4]). A visual analog scale was used to measure participants' levels of mental and physical fatigue. This scale ranged from 0-100 in increments of 5 where 0, 50 and 100 indicated that individuals were not fatigued at all, moderately fatigued and very fatigued, respectively. For trials executed with the physical and virtual system, these measurements were collected using online surveys [22]. Participants' performance during physical and virtual trials was examined through video analysis. The video recordings allowed us to compute i) the Levenshtein's Distance (LD; [8,14]) which indicates to what extent the executed sequence differs from the demonstrated (reference) sequence and ii) sequence completion time (SCT; [8]) which was the time between starting the first and completing the last sequence action.

2.2.4 Data processing

2.2.4.1 Survey data

The raw SUS scores were normalized and then combined resulting in a single score between 0-100 for each participant, task and execution modality [23,24]. Similarly, each subscale of the NASA TLX and fatigue scores were separately averaged, resulting in scores between 0-100 for each participant, task and execution modality [21].

2.2.4.2 Performance data

The SCT represented the time elapsed by the participant to complete the demonstrated action sequence for a given task and execution modality. Then, the average SCT was computed for each participant and conditions to be subjected to statistical analysis. In addition, the LD was computed for each participant and condition. The LD measures the distance between two sequences, which represents the minimum number of operations (here insertions, deletions, and substitutions were considered; see below) needed to match one sequence to the *reference sequence* [8,14,25]. Computing LD is achieved by defining an alphabet of symbols representative of all possible sequence components, sequences, and operations. In general, a sequence alphabet can be defined as a finite set $\{A_1, A_2, \ldots A_i, \ldots, A_{n-1}, A_n\}$ where A_j is the j^{th} atomic symbol among all N possible symbols. For instance, the alphabet for the PM task was {Open main valve; Close main valve; Press toggle i; Remove cleanout adapter i; Put down cleanout adapter i; Pick up spare cleanout; Discard in bin; Pick up tool; Extract object from pipe; Discard object in bin; Put down tool} (where i = {1,2,3,4} (for the alphabet for the ToH and DM tasks, see [14]). As such, the operators modify one action at a time resulting in a different sequence. In the current context, each atomic "symbol" represents an elementary action in a given action sequence — an atomic action. A motor sequence would then be defined as a finite, ordered list of zero or more atomic actions from the alphabet with or without repeating atomic actions. For instance, one of the demonstrated action sequences for the PM task was < Open main valve, Press toggle 1, Remove cleanout adapter 1, Pick tool up, Extract object, Discard object, Put down tool, Replace cleanout

adapter 1, Press toggle 1, Close main valve >. To compare the reference (or demonstrated sequence) and the sequence executed by the participant, the following three classical LD operators were considered: (i) *insertion* of one action, (ii) *deletion* of one action, and (iii) *substitution* of one action for another one (i.e., a replacement). More specifically, insertions refer to the addition of an action anywhere in the sequence that increases the sequence length by one compared to the reference sequence. A deletion eliminates an action at any location in the sequence which decreases its length by one. Substitutions describe the replacement of an existing action in the sequence with a different action without changing the action sequence length. Thus, this processing allowed to obtain the LD, the number of insertion (NI), number of deletion (ND) and the number of substitution (NS) (for details see [8,14]). Computationally, the LD was computed using the well-established dynamic programming approach Wagner-Fischer algorithm (for details see [8,14,26]). Then, the average LD, NI, ND and NS were subjected to statistical analysis.

2.2.5 Statistical Analysis

First, an analysis to assess the success rate for the three tasks executed physically and virtually was conducted. Then, for each task separately, the mean survey scores (total SUS score, each NASA TLX subscale scores, the physical and mental fatigue survey scores) as well as the mean SCT, LD, NI, ND, and NS obtained when individuals used the physical and virtual systems were compared using paired *t-tests* or *Wilcoxon signed-ranked tests* depending on whether the assumption of normality (assessed by a Lilliefors test) was violated or not. In addition, the Cohen's d effect sizes were computed and reported. A one sample t-test was used to compare the mean SUS scores to the well-established industry threshold value of 68 (i.e., scores smaller and greater than this cut-off represent below and above and average usability) [23,24]. The false discovery rate was employed to account for the multiple comparisons conducted to compare the measurements listed above (i.e., survey scores; SCT, LD, NI, ND, and NS) obtained for both physical and virtual systems. All criterion alpha levels were set to $p < 0.05$.

3 Results

3.1 Usability

The results of the qualitative analysis revealed that all participants fully completed the three tasks either physically or virtually. Thus, no participant drop-out from the study or any session. This resulted in the same number of trials for both the physical and virtual systems. Considering technical issues and rule breaking, the vast majority of the trials were successful for the three tasks when physically (97.48 %) or virtually (94.01 %) performed. Specifically, most of the trials were free of any technical glitches for the physical (98.44 %) and virtual (93.75 %) system. Further examination revealed that this difference in success rate was comparable between tasks. Also, for the three tasks, a large majority of the physical (96.53 %) and virtual (94.27 %) trials executed did not have any rule breaks. Additional analyses revealed that when virtually executed

the PM task completion was less successful than when physically completed (Physical: 94.79 %; Virtual: 89.58 %). However, both the ToH and DM tasks presented a similar and greater success (ToH – Physical: 100 %, Virtual: 100 %; DM – Physical: 94.79 %, Virtual: 93.23 %) regardless of the system (i.e., physical or virtual) used.

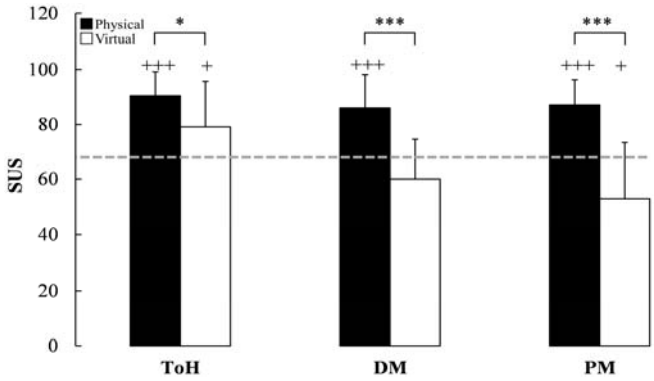

Fig. 4. Usability scores when individuals perform the ToH, DM and PM tasks with the physical and virtual systems. The dashed gray line represents the 68 threshold usability level (see text for details). ToH: Tower of Hanoi; DM: Disk drive dock maintenance; PM: Pipe maintenance. The stars (*) and crosses (+) represent the significance level for the Physical vs. virtual contrast and the Physical or virtual vs, average acceptability threshold contrast, respectively. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; +: $p < 0.05$; ++: $p < 0.01$; +++: $p < 0.001$.

The SUS scores for each of the three tasks were significantly higher when executed with the physical compared to the virtual systems (ToH: $t(11) = 2.580$, $p = 0.036$, $d = 0.745$; DM: $t(11) = 5.461$, $p < 0.001$, $d = 1.576$; PM: $t(11) = 4.893$, $p < 0.001$, $d = 1.412$). Also, while the SUS scores obtained for the physical system were all above the usability threshold (ToH: $t(11) = 8.537$, $p < 0.001$, $d = 2.465$; DM: $t(11) = 5.000$, $p < 0.001$, $d = 1.443$; PM: $t(11) = 4.893$, $p < 0.001$, $d = 2.057$) the results were less consistent for the virtual system. Specifically, the SUS scores obtained with the virtual system when executing the ToH ($t(11) = 2.376$, $p = 0.041$, $d = 0.686$) and the PM ($t(11) = -2.530$, $p = 0.036$, $d = 0.730$) tasks were above and below this threshold, respectively. Finally, the average SUS score for the DM was below the usability threshold although statistically not different from it ($t(11) = -1.886$, $p = 0.086$, $d = 0.544$) (see Fig. 4).

3.2 Mental workload

The same statistical analysis revealed that the physical and virtual execution of the ToH task did not affect any workload dimensions of the NASA TLX ($p > 0.137$, $0.001 < d < 0.544$). Also, no difference in the perceived temporal demand between the physical and virtual execution of any of the three tasks was detected ($p > 0.117$, $0.095 < d < 0.634$). However, the perceived mental demand increased when the DM ($z = -2.511$, $p = 0.024$, $d = 0.858$) and PM ($t(11) = -3.930$, $p = 0.019$, $d = 1.135$) tasks were executed

with the virtual compared to the physical system. Similarly, perceived effort and frustration were greater when participants operated the virtual system to complete the DM (Effort: z = -2.590, p = 0.022, d = 0.816; Frustration: z = -2.805, p = 0.019, d = 0.917) and PM (Effort: t(11) = -2.786, p = 0.019, d = 1.105; Frustration: z = -3.062, p = 0.019, d = 0.899) tasks relative to the physical systems.

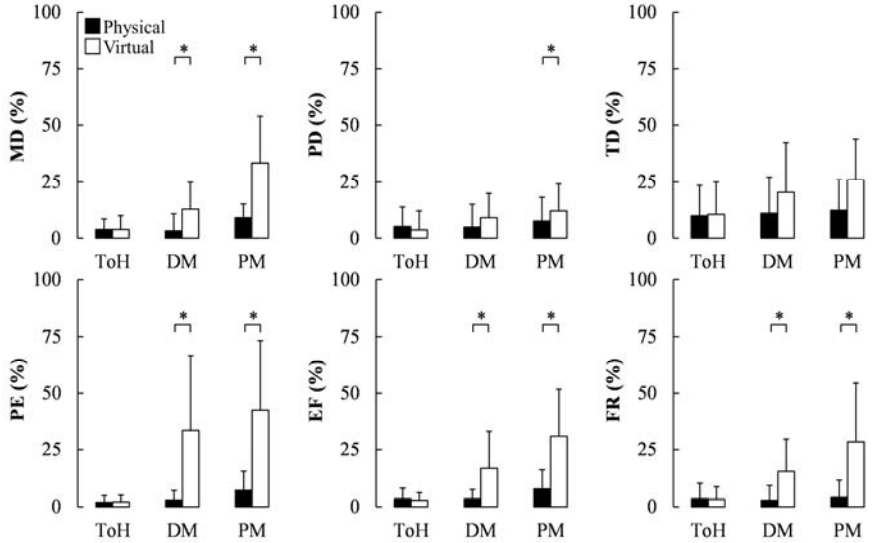

Fig. 5. Differences in perceived mental demand (top left panel) and the other five dimensions (PD, TD, PE, EF, FR) obtained by means of the NASA TLX questionnaire during the performance of the ToH, DM and PM tasks using the physical (black bars) and virtual (white bars) systems. MD: Mental demand; PD: Physical demand; TD: Temporal demand; PE: Performance; EF: Effort; FR: Frustration. ToH: Tower of Hanoi; DM: Disk drive dock maintenance; PM: Pipe maintenance. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

In addition, the performance was perceived as better when executing the DM (z = -2.669, p = 0.020, d = 0.939) and PM (t(11) = -3.688, p = 0.019, d = 1.065) tasks with the physical relative to the virtual system. Finally, the analysis also revealed that the physical demand was perceived as higher when the PM task was executed in the virtual compared to the physical environment (z = -2.688, p = 0.020, d = 1.035) whereas this was not observed for the DM task (z = -1.174, p = 0.333, d = 0.278) (see Fig. 5).

3.3 Performance

Statistical analysis of SCT revealed no differences between the ToH executed with the physical and virtual systems (z = -0.533, p = 0.594, d= 0.237). However, an elevation of SCT was observed when participants executed the DM task with the virtual compared to the physical system (z = -3.059, p = 0.002, d = 1.795). Similarly, compared to the physical system, the execution of the PM task with the virtual system resulted in a significantly longer SCT (t(11) = -11.600, p < 0.001, d = 3.349). For the three tasks,

no difference in LD or its operators (NI, ND and NS) were revealed for action sequences executed with the physical and virtual system ($p > 0.469$) (see Fig. 6; first column; the NI, ND and NS are not represented).

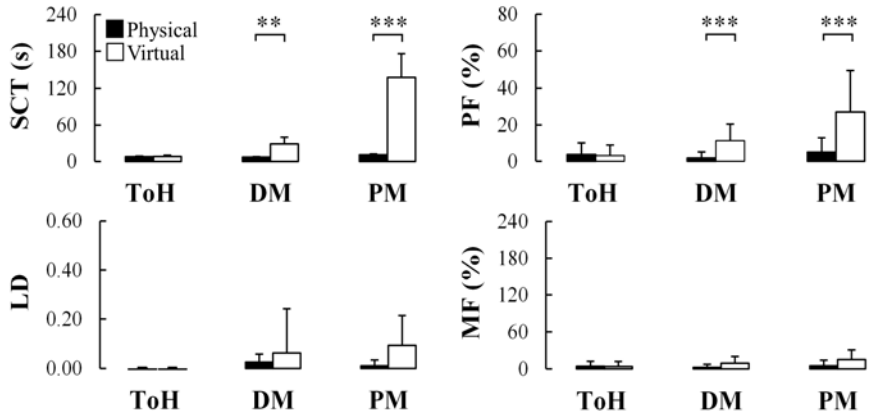

Fig. 6. Changes in performance (SCT, LD; first column) and perceived fatigue (physical, mental; second column) during execution of the ToH, DM, PM tasks with the physical (black bars) and virtual (white bars) systems. ToH: Tower of Hanoi; DM: Disk drive dock maintenance; PM: Pipe maintenance. PF: Physical fatigue; MF: Mental fatigue. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

3.4 Fatigue

No differences in physical fatigue were observed between the virtual and physical systems for the ToH task ($z = 1.035$, $p = 0.319$, $d = 0.318$). However, the completion of the DM and PM tasks with the virtual systems resulted in greater physical fatigue in participants relative to the physical systems (DM: $z = -2.670$, $p = 0.040$, $d = 1.012$; PM: $z = -2.473$, $p = 0.040$, $d = 0.870$). In addition, the mental fatigue elicited by each of the three tasks was not significantly different between physical and virtual systems (ToH: $z = 0.997$, $p = 0.319$, $d = 0.289$; DM: $z = -1.992$, $p = 0.093$, $d = 0.570$; PM: $z = -1.844$, $p = 0.098$, $d = 0.540$) (see Fig. 6; second column).

4 Discussion

Overall, findings revealed that the proposed approach appropriately assessed human cognitive-motor behavior when individuals executed various virtual tasks involving action sequences remotely. The results also revealed that when individuals executed the virtual and physical ToH tasks, similar levels of usability, performance, mental workload and mental/physical fatigue were obtained. Conversely, relative to physical execution, the virtual completion of the PM task and, to a slightly lesser extent, the DM task resulted in below average levels of usability as well as performance decrements along with elevations in mental workload and physical fatigue.

4.1 Systems-dependent differences in usability, performance, mental workload and fatigue

First, for all three tasks, it appeared that the entire study could be successfully conducted remotely using VLEARN and the internet to complete tasks with a similar level of success compared to physically executing these tasks. Although slightly better for the physical system, few technical issues among the three tasks were observed when individuals performed with VLEARN. In addition, both the physical and virtual trials had a very similar task completion success rate for both the ToH and the DM tasks. The success rate for PM task was lower for virtual compared to physical execution. Second, for all three tasks, the usability of the physical system reached an "excellent" rating according to the acceptability range paired with adjectives and letter-grade scales proposed by Bangor and colleagues [18-20]. Although not surprising, this is important since these physical set-ups served as a reference to evaluate the usability of VLEARN and more generally its effects on human performance when executing the three tasks considered here. The virtual environment for the ToH task received the highest SUS score of all virtual systems. Importantly, it was the only virtual system which received a score above the industry-driven and widely acknowledged threshold (i.e., 68 points) which corresponds to a "good" rating on the aforementioned scale [18-20]. Although the ToH task executed physically elicited a higher usability relative to its virtual execution, this difference was much smaller compared to those obtained for the DM and PM tasks as indicated by smaller effect sizes. The virtual and physical completion of the ToH led to similar performance as suggested by comparable imitation quality of the demonstration (LD $\approx$ 0) and duration of sequence completion (similar SCT). It must be noted that here the former was expected for all three tasks and both systems due to the simplification of sequences for experimental purposes (see section 2.2). Thus, similar SCT suggest that the velocity at which the actions of the sequence were performed when operating the physical and virtual systems were similar. Also, the same between-systems comparison led to similar levels of mental workload (and subscale scores of the NASA TLX) which, when combined with performance results, suggest comparable cognitive-motor efficiency along with a similar physical and mental fatigue [3,4,27].

However, the virtual execution of the DM and PM tasks resulted in a level of usability below the industry standard score of 68 with scores of 60 and 53.13 which both corresponded to "OK" and fell in the "marginally low" range of acceptable usability with letter-grades of D and F, respectively [18-20]. However, while the usability score for the virtual DM task was below average, it was not statistically different from the standard score of 68 whereas the PM task was well below this standard. Along these lines, compared to physical execution, the SCT for both the DM and PM tasks with VLEARN was longer whereas no major discrepancies between the demonstrated and imitated sequences was observed (the latter was expected for the reasons mentioned in section 2.2). Thus, these results (similar LDs and greater SCT) suggest that slower execution of the actions composing the sequence when operating the virtual system (relative to the physical system) were likely not due to mistakes while forming the sequences (e.g., adding extraneous actions) and/or excessive pause between actions due to mistake or hesitation (as observed in the video analysis).

Furthermore, relative to physical task execution, higher mental workload (as well as perceived performance failure, effort and frustration) were obtained when the DM and PM tasks were executed virtually. In particular, a reduction in performance (i.e., SCT) along with this elevation of the mental workload collectively suggest a reduction of the cognitive-motor efficiency [3,4,27] when the DM and PM tasks were completed virtually relative to physically. It is important to note that although a decrement of usability and performance along with an elevation of the mental workload were observed for these two virtual systems, the changes were more prominent for the PM task (as indicated by greater effect sizes). Importantly, although the execution of the DM and PM tasks with the virtual relative to the physical system led to differences in mental workload it did not translate to mental fatigue which was comparable for both systems. Finally, the execution of the PM task with VLEARN was perceived as more physically demanding than when executed with the actual set-up whereas this was not observed for the DM task. However, when both tasks were virtually executed an elevation in physical fatigue was observed. It must be noticed that, as expected, temporal demand was not significantly different between any of the physical and virtual task systems since the emphasis was placed on using the core components of each system to complete the imitation task correctly rather than quickly.

4.2 The effect of the controls on the usability, performance, mental workload and fatigue

Generally, the virtual and physical execution of the ToH task led to acceptable levels of usability and cognitive-motor performance without eliciting elevated mental or physical fatigue. Conversely, compared to its physical execution, the virtual completion of the DM and PM tasks led to below average levels of usability along with degraded performance and increased mental workload implying decreased cognitive-motor efficiency which translated in an elevation of the physical fatigue. Several reasons discussed below could explain these results. First, these differences are likely due to the fact that when individuals used VLEARN to execute the ToH task, they could employ a classical point-and-click technique whereas the DM and PM required actively manipulating 3D objects in 3D space with their mouse/trackpads which ultimately imposed a certain accuracy requirement and thus was likely more challenging. For instance, in the ToH task, to move a disk from one peg to another, individuals had just to click the disk to select it then click the peg to which they wanted to move the disk. However, replacing a faulty disk in the DM task required multiple steps, for example, participants had to click the disk and actively move it (e.g., pick up, drop, proper placement of the cursor) with their mouse/trackpad while using keyboard controls to switch between movement planes until the disk was above the empty slot before finally using a right-click to drop the disk into the slot. As such, when this task was executed in the virtual environment, removing and replacing the drive proved more challenging than manipulating its physical counterpart. Similarly, to complete the PM task, individuals had to combine mouse/trackpad with keyboard and on-screen controls to adjust angles before tool use, place, extract and discard object which was likely more demanding.

A second element to explain these results was that the PM and, to a lesser extent, the DM tasks involved a greater number of controls and components as well as an increased dependency on those controls to manipulate relevant objects in the environment. For instance, the layout of the virtual PM task required individuals to use transport pads (selected by mouse-click) to actually navigate through the 3D environment to complete this task because not all system components were accessible from individual transport pads. As a result, when executing the PM task with VLEARN, the participants had to move to relevant transport pads (see green circles with arrows in Fig. 3) and manage switching between on-screen, keyboard, and cursor control throughout each trial. However, while executing the PM task with the physical system, participants had all system components within reach and field of vision while being able to rely on haptic and visual feedback during task execution.

Therefore, the use of the simplified controls to virtually complete the ToH task allowed action sequences to be completed without any additional challenges compared to the physical system resulting in comparable performance (particularly the SCT). In addition, although it was suggested that the execution of a physical 3D relative to a virtual 2D ToH task differently engage cognitive-motor resources [11], these simplified controls did not necessarily magnify these differences as the perceived mental workload elicited with both physical and virtual systems were comparable for this task. Such similar levels of performance and mental workload resulted in comparable cognitive-motor efficiency as well as levels of physical and mental fatigue when the ToH task was executed with both the virtual and physical systems.

Conversely, the virtual execution of the DM and for PM tasks required additional controls involving active object manipulations which were more demanding potentially requiring further engagement of cognitive (e.g., attentional) and motor (e.g., fine coordination) resources. These constraints were particularly challenging because the virtual controls inherently lack the natural feedback (e.g., cutaneous; proprioceptive) present when physical systems are used. Challenges related to using the more complex virtual controls to perform the DM and PM tasks were likely magnified by the use of a trackpad (for 10 of 12 participants) instead of a traditional computer mouse in addition to the on-screen and keyboard controls. As a result, the execution of the DM and PM tasks with VLEARN led to lower levels of usability as well as a degraded performance (SCT) combined with a greater mental workload resulting a lower cognitive-motor efficiency compared to those observed when these tasks were physically completed [3,4,27]. Interestingly, these changes were more prominent in the PM compared to the DM task; the latter being associated with a SUS score below, but not statistically different from, the acceptance-threshold as well as smaller effect sizes when contrasting the virtual and physical cognitive-motor performance. Possibly, the navigation element of the virtual PM task may have magnified the discrepancies in usability, performance and mental workload compared to its completion with the physical system. Such differences are important since compared to the DM task, the success rate for PM task was lower for its virtual than physical execution. Namely, for the former less than half of the trials reached the time limit, thus such higher rule breaking was also related to the execution action sequence rules per-se (see Section 2.2). Possibly, a greater deployment of attentional resources (and thus higher mental workload) may have been needed to deal with

the interface controls, leaving less of these resources for action monitoring contributing thus to rule breaking. This is consistent with the idea that both attentional control and action monitoring are closely related [28]. Finally, the use of these demanding controls to perform both of these tasks virtually may have ultimately led to greater physical fatigue compared to their execution with the physical set-up.

4.3 Limitations, conclusions, and future work

Overall, this work suggests that our approach allows to experimentally assess human cognitive-motor behavior (performance, mental workload, fatigue) when individuals operate VLEARN via the internet as suggested by similar success rates for both the physical and virtual execution of the three action sequence tasks considered here. Similar levels of usability, performance, mental workload and fatigue were observed when individuals operated the physical systems or a virtual system with simple controls (e.g., point-and-click method used for the ToH task). This suggests that under such conditions VLEARN reproduced its real counterpart with fairly good fidelity. However, when VLEARN used more complex control systems (e.g., those used for the DM and PM tasks), usability and cognitive-motor behavior degraded in particular for the PM task which contained the most elements. Thus, these complex control systems (which are critical for tasks with many components to manipulate) should be revised. Otherwise, excessive complexity of the controls can become a cofounding factor when assessing human cognitive-motor behavior with experimental manipulations (e.g., task demands). The simplification of the controls may be easier to implement with tasks having a limited number of elements (e.g., the DM task) relative to those with many components (e.g., the PM task) for which a more immersive system may be needed. Although simplified controls (e.g., point-and-click) may somewhat limit the study of finer motor manipulations, this is already well adapted for examining high-level planning processes engaged to generate action sequences under different levels of challenge.

This study had limitations. First picking up and dropping objects was notably harder for the participants who used a trackpad instead of a traditional computer mouse. Although individuals used personal computers with their trackpad or computer mouse, having them trying different control options during the familiarization phase may have allowed them to choose the option that best matched to their experience and ultimately provided enhanced results. This should be considered in future studies. Also, although this was a performance and not a learning study, an exploratory analysis of the blocks revealed that there were limited practice effects such that performance, mental workload and fatigue were stable during the last two blocks for both the physical and virtual systems. Thus, a future study to assess learning and retention could be conducted although ultimately the design of the controls used in VLEARN should minimize learning to be able to operate this virtual platform.

This study provided valuable information for revising and extending VLEARN. Overall, our approach allowed us to examine cognitive-motor performance and mental workload during remote execution of various complex action sequences via VLEARN in healthy and patient populations. Such work could inform various applications such

as telehealth evaluations for Veterans having complex symptom presentations. In particular, the addition of brain monitoring along with task demand manipulation to investigate high-level planning processes would enable the objective measurement of brain dynamics and performance outcomes that may be influenced by military exposures.

Although the present work could be extended in different directions, immediate future efforts will first aim to update and extend VLEARN by enhancing its control interface and possibly incorporating novel hardware (e.g., joysticks, immersive VR technology) to improve its usability. Such an approach would allow us to remotely study human cognitive-motor behavior during various action sequence tasks which can be manipulated experimentally (e.g., high versus low cognitive demand). A second immediate future step would be to deploy this virtual system to remotely assess cognitive-motor performance combining behavior and electroencephalography when individuals execute action sequences to complete complex tasks. Future work that is currently underway aims to allow multiple human or robotic agents to interact within the VLEARN environment to evaluate human-human, human-robot, and robot-robot teaming when collaboratively performing or learning complex cognitive-motor tasks.

Acknowledgment

This work was supported by The Office of Naval Research (N00014-19-1-2044).

References

1. Wickens, C.D. Multiple resources and mental workload. Hum Factors 50(3) 449–455 (2008)
2. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A. State of science: mental workload in ergonomics. Ergonomics 58(1) 1–17 (2015).
3. Shaw, E.P., Rietschel, J.C., Hendershot, B.D., Pruziner, A.L., Miller, M.W., Hatfield, B.D., Gentili, R.J. Measurement of attentional reserve and mental effort for cognitive workload assessment under various task demands during dual-task walking. Biol Psychol 134, 39–51 (2018).
4. Shuggi, I.M., Oh, H., Wu, H., Ayoub, M.J., Moreno, A., Shaw, E.P., Shewokis, P.A., Gentili, R.J. Motor Performance, mental workload and self-efficacy dynamics during learning of reaching movements throughout multiple practice sessions. Neuroscience 423, 232-248, (2019).
5. Welsh, M.C., Huizinga, M. Tower of Hanoi disk-transfer task: influences of strategy knowledge and learning on performance. Learn Individ Differ 15(4), 283–298 (2005).
6. Wulf, G., Shea, C.H. Principles derived from the study of simple skills do not generalize to complex skill learning. Psychon. Bull. Rev., 9(2), 185–211 (2002).
7. Hardy, D.J., Wright, M.J. Assessing workload in neuropsychology: An illustration with the Tower of Hanoi test. J Clin Exp Neuropsychol, 40(10) 1022-1029 (2018).
8. Hauge, T.C., Katz, G.E., Davis, G.P., Jaquess, K.J., Reinhard, M.J., Costanzo, M.E., Reggia J.A., Gentili, R.J. A novel application of Levenshtein distance for assessment of high-level motor planning underlying performance during learning of complex motor sequences. J. Mot. Learn. Dev. 8(1), 67-86 (2019).
9. Radüntz T. The effect of planning, strategy learning, and working memory capacity on mental workload. Sci Rep. 10(1), 7096 (2020).

10. Vakil, E., Lev-Ran Galon, C. Baseline performance and learning rate of conceptual and perceptual skill-learning tasks: The effect of moderate to severe traumatic brain injury. J Clin Exp Neuropsychol 36(5), 447–454 (2014).
11. Milla, K., Bakhshipour, E., Bodt, B., Getchell, N. Does movement matter? Prefrontal cortex activity during 2D vs. 3D performance of the Tower of Hanoi puzzle. Front. Hum. Neurosci., 13, 156 (2019).
12. Katz, G.E., Huang, D.W., Gentili, R.J., Reggia, J.A. Imitation learning as cause-effect reasoning. In: Steunebrink, B., Wang, P., Goertzel, B. (eds.) Artificial general intelligence. AGI 2016. Lecture notes in computer science, vol 9782. Springer, Cham (2016).
13. Katz, G., Huang, D.W., Hauge, T., Gentili, R., Reggia, J. A novel parsimonious cause-effect reasoning algorithm for robot imitation and plan recognition. IEEE Trans Cognit Dev Syst PP(99):1–17 (2017)
14. Hauge, T. C., Katz, G. E., Davis, G. P., Huang, D. W., Reggia, J. A., Gentili, R. J. High-level motor planning assessment during performance of complex action sequences in humans and a humanoid robot. Int. J. Soc. Robot, 13, 981–998 (2021).
15. Shaver, A., Shuggi, I., Katz, G., Davis, G., Reggia, J., Gentili, R. Effects of practicing structured and unstructured complex motor sequences on performance and mental workload. In: North American Society for the Psychology of Sport and Physical Activity Virtual Conference, J Sport Exerc Psychol, vol. 42 (S1), S56-S56, Human Kinetic Publisher Inc (2020).
16. Huang DW, Katz GE, Langsfeld JD, Gentili RJ, Reggia JA (2015) A virtual demonstrator environment for robot imitation learning. In: IEEE international conference on technologies for practical robot applications (TePRA), Woburn, MA, USA, pp 1–6
17. Huang D.W., Katz G.E., Langsfeld J.D., Oh H., Gentili R.J., Reggia J.A. An object-centric paradigm for robot programming by demonstration. In: Schmorrow, D.D., Fidopiastis, C.M. (eds.) Foundations of Augmented Cognition. AC 2015. LNCS, vol 9183. Springer, Cham (2015).
18. Bangor, A., Kortum, P. T., Miller, J. T. An empirical evaluation of the system usability scale. Intl. Int J Hum-Comput Int, 24(6), 574-594 (2008).
19. Kortum, P.T., Bangor, A. Usability ratings for everyday products measured with the system usability scale. Int J Hum-Comput Int 29(2), 67-76 (2013).
20. Bangor, A., Kortum, P., Miller, J. Determining what individual SUS scores mean: adding an adjective rating scale. J. Usability Stud. 4(3), 114-123 (2009).
21. Hart, S.G., Staveland, L.E. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Advances in Psychology 52, 139–183 (1988).
22. Childs, A. (2020). Qualtrics.
23. Sauro, J., Lewis, J.R. Quantifying the user experience: Practical statistics for user re-search (2nd ed.). Morgan Kaufmann, Cambridge, (2016).
24. Barnum, C.N. Usability testing essentials (2nd ed.), Morgan Kaufmann, Cambridge, (2021).
25. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics doklady 10(8), 707–710 (1966)
26. Wagner, R.A., Fischer, M.J. (1974). The string-to-string correction problem. J ACM 21(1), 168–173.
27. Jaquess, K.J., Gentili, R.J., Lo, L.C., Oh, H., Zhang, J., Rietschel, J.C., Miller, M.W., Tan, Y.Y., Hatfield, B.D. Empirical evidence for the relationship between cognitive workload and attentional reserve. Int J Psychophysiol. 121:46-55 (2017).
28. Mahon, A., Bendžiūtė, S., Hesse, C. Hun, A.R. Shared attention for action selection and action monitoring in goal-directed reaching. Psychol. Res., 84, 313–326 (2020).