HUMAN PERFORMANCE MODELING OF UPPER LIMB PROSTHETIC DEVICES

A Dissertation

by

JUNHO PARK

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Maryam Zahabi
Committee Members,	Mark Benden
	He (Helen) Huang
	Farzan Sasangohar
	Xudong Zhang
Head of Department,	Lewis Ntaimo

May 2023

Major Subject: Industrial Engineering

ABSTRACT

Limb amputation can cause severe functional disability for the performance of activities of daily living (ADLs). Amputee patients use prosthetic devices (PDs) to perform ADLs. PDs require a substantial amount of cognitive resources, and some users reject their devices due to poor usability. However, very few studies have investigated usability issues, and they mainly used subjective methods such as questionnaires. In addition, no prior studies classified cognitive workload of using PDs in early stage of the design process. To fill out these research gaps, the objectives of this work were to: 1) Provide an objective usability evaluation of prosthetic devices, 2) Develop a human performance modeling tool to assess the usability and cognitive workload of upper limb PDs (i.e., HPM-UP: Human Performance Model for Upper limb Prostheses) and (3) Validate the model with experimental data.

Chapters 1 and 2 provide a review of literature on usability evaluation of prosthetic devices and human performance modeling approaches. In Chapter 3, the computational formulations for each dimension of HPM-UP was developed based on the literature and existing theories: (1) adaptive learning curve formulation was used for calculating learnability; (2) based on the learnability formula, error rate was calculated using a natural exponential function, (3) memorability was calculated based on the ACT-R declarative module, (4) efficiency was formulated based on user task performance, (5) satisfaction

was formulated based on the expectation confirmation theory, and (6) cognitive workload classification model was developed with the Naïve Bayes algorithm.

Chapters 4 and 5 focused on validating the HPM-UP. A human subject experiment (Experiment 1) was conducted with 30 able-bodied participants using three types of PDs. Hypotheses were formulated for each dimension of HPM-UP to test if there was any significant difference among the human-subject data, HPM-UP estimates, and the benchmark model estimates. A second human subject study (Experiment 2) was conducted with 20 able-bodied participants to validate the HPM-UP in a virtual environment. The findings of both experiments suggested that there were no significant differences between the human subject data and HPM-UP estimates. However, there were significant differences between human subject data and benchmark model. Also, there were significant differences between the HPM-UP estimates and the benchmark model outcomes.

HPM-UP can be run using a graphical user interface (GUI) and do not require hard-coding to run the model. It is a first comprehensive usability evaluation package developed in an R Shiny package format and released on GitHub, which can be used by other researchers, designers, or clinicians. The outcomes of this research are expected to be useful for both researchers and practitioners as this is the first computational modeling approach to assess the usability of prosthetic devices early in the design cycle. In addition, the results are expected to provide a basis to enhance the design of prosthetic devices to reduce cognitive workload and improve device usability.

DEDICATION

I express my infinite gratitude for the sincere love, patience, and understanding my wife Jiyeon and daughter Danbi showed me during my journey for Ph.D. Special thanks to my parents and all family members for their encouragement with everlasting prayer. In particular, I have no words to express my gratitude for the great courage, perseverance, and struggle as a human being, which my father, who is physically handicapped, has shown me throughout his life. Many thanks to my friends for their encouragement. I give all this achievement, honor, and glory to the Lord.

ACKNOWLEDGEMENT

I want to express the most profound appreciation to my committee chair, Dr. Maryam Zahabi, and my committee members, Dr. Benden, Dr. Huang, Dr. Sasangohar, and Dr. Zhang for their guidance and support throughout this research. Thanks also to my colleagues, faculty, and staff for making my time at Texas A&M University a great experience.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Zahabi (chair), Dr. Zhang, and Dr. Sasangohar from Wm Michael Barnes ’64 Department of Industrial & Systems Engineering, Dr. Benden from School of Public Health, and Dr. Huang from Joint Department of Biomedical Engineering in North Carolina State University.

Funding Sources

Funding for this research was provided by the National Science Foundation (NSF) (No. IIS-1856676). Its contents are solely the authors' responsibility and do not necessarily represent the views of the NSF.

NOMENCLATURE

5ST	5 Second Test
ACT-R	Adaptive Control of Thought - Rational
ADL	Activities of daily living
AI	Artificial Intelligence
ANOVA	Analysis of Variance
AR	Augmented Reality
AUC	Area under the ROC curve
B&B	Box and Block
CC	Continuous Control
CRT	Clothespin Relocation Task
CMN-GOMS	Card, Moran, and Newell - GOMS
CPM	Cognitive performance model/modeling
CPM-GOMS	Cognitive-Perceptual-Motor GOMS
CSD-OPUS	Client Satisfaction with Device module of the Orthotics and Prosthetic Users' Survey
CSV	Comma Separated Value
CV	Cross Validation
CW	Cognitive Workload
DARPA	Defense Advanced Research Projects Agency
DC	Direct Control
DCQ	Device Calibration Quality
ECRL	Extensor Carpi Radialis Longus
ECT	Expectation Confirmation Theory
ED	Extensor Digitorum
E-GOMSL	Enhanced GOMSL
EMG	Electromyography
EPIC	Executive-Process Interactive Control
ERP	Event-Related Potential
ETD	Electric Terminal Device
FCR	Flexor Carpi Radialis
FD	Flexor Digitorum
FI	First Impression
FMG	Force myography

fNIRS	functional Near-Infrared Spectroscopy
FT/FA	Frontal Theta/Frontal Alpha
FT/PA	Frontal Theta/Parietal Alpha
GOMS	Goals, Operators, Methods, and Selection rules
GOMSL	GOMS Language
HCI	Human-Computer Interaction
HMI	Head-Mounted Interface
HPM	Human Performance Model/Modeling
HPM-UP	Human Performance Model for Upper-limb Prostheses
HSI	Human-System Interaction
IMPRINT	Improved Performance Research Integration Tool
IMU	Inertial Measurement Unit
IRB	Institutional Review Board
ISO	International Standard Organization
JHFT	Jebsen Hand Function Test
KLM	Keystroke-Level Model
LDA	Linear discriminant analysis
LPP	Late positive potential
LTM	Long-Term Memory
MAV	Mean Absolute Value
MHP	Model Human Processor
ML	Machine Learning
MTM	Motion-Time Measurement
NASA-TLX	NASA Task Load Index
NGOMSL	Natural GOMSL
NB	Naïve Bayes
NSF	National Science Foundation
ODCQ	Objective Device Calibration Quality
PD	Prosthetic Devices
PPT	Purdue Pegboard Test
PR	Patter Recognition
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta Analyses
QN-MHP	Queueing Network – Model Human Processor
QUEST 2.0	Quebec User Evaluation of Satisfaction with assistive Technology
Q-TFA	Questionnaires for persons with a TransFemoral Amputation
RAM	Random Access Memory
RF	Random Forest
RFE	Recursive Feature Elimination

RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristics
SANLaB-CM	Stochastic Activity Network Laboratory for Cognitive Modeling
SDCQ	Subjective Device Calibration Quality
SDK	Software Development Kit
SFS	Sequential Feature Selection
SHAP	Southampton Hand Assessment Procedure
SOAR	State, Operator, And Result
SUS	System Usability Scale
SVC	Support Vector Classifier
TAPES-R	Trinity Amputation and Prosthesis Experience Scale – Revised
TCP	Transmission Control Protocol
TCT	Task Completion Time
UEFI	Upper Extremity Functional Index
USE	Usefulness, Satisfaction, and Ease of Use
VR	Virtual Reality

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION[*]

1.1. Background

Amputee patients experience severe functional disability in activities of daily living (ADL) due to the lack of prosthetic device usability. More than 2 million amputees live in the U.S., and about 185,000 amputations occur each year; this number is expected to be doubled by 2050 due to the increasing rates of contributing diseases (Amputee Coalition, 2021). Amputees use prosthetic devices regularly to perform ADL. These activities may not be possible without prosthetic devices or require additional effort and time (Gaskins et al., 2018; Lusardi et al., 2013). However, existing devices are often reported to be challenging to use, leading to reduced utilization and device rejection (Engdahl et al., 2015; Kannenberg and Zacharias, 2014). A study assessing the usability of different prosthetic devices found that 53% of passive hand users, 50% of body-powered hook users, and 39% of myoelectric hand users rejected prosthetic arms. The main reasons for rejection were poor dexterity, glove durability, and lack of sensory feedback (Biddiss et al., 2007; Bowker, 2004; Montagnani et al., 2015).

Using upper limb prostheses requires substantial cognitive resources (Geurts and Mulder, 1994; Geurts et al., 1991; Heller et al., 2000; Hofstad et al., 2009; Williams et al., 2006). Cognitive resources are used to compensate for the loss of motor control and mitigate the damage of somatosensory feedback from the amputated limb (Childress, 1980; Heller et al., 2000; Herberts and Körner, 1979; Krewer et al., 2007; Williams et al., 2006; Witteveen et al., 2012). Therefore, using prostheses can cause a lack of cognitive capacity to conduct other mental activities (Heller

[*] Part of this chapter is reprinted with permission from J. Park and M. Zahabi, "Cognitive Workload Assessment of Prosthetic Devices: A Review of Literature and Meta-Analysis," in *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 2, pp. 181-195, April 2022, doi: 10.1109/THMS.2022.3143998. Copyright 2023 by IEEE.

et al., 2000; Williams et al., 2006). The high mental workload can also reduce the primary task performance (Duysens et al., 2012). For example, a patient may find it difficult to avoid obstacles or walk uneven terrain. In case of upper limb amputation, most of the current control strategies use limited information (i.e., shoulder movements or recorded electromyography (EMG) signals) for activating several degrees of freedom (DoF) of the prosthetic devices, which is non-intuitive and unnatural, and can result in high cognitive workload (CW) (Cordella et al., 2016). By assessing CW, the analyst can explain the underlying attentional resources engaged during task execution and support the evaluation/development of prosthetic devices (Gaskins et al., 2018).

1.2. Literature Review

1.2.1. Usability Evaluation of Prosthetic Devices

1.2.1.1. Usability Dimensions

The International Standard Organization (ISO) defines usability as "*The extent to which specified users can use a product to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use*." (ISO, 2019) The word "usability" also refers to how easy and pleasant the features are to use. Nielsen (2012) proposed five dimensions of usability, including:

- Learnability: How easy is it for users to accomplish basic tasks the first time they encounter the design?
- Efficiency: Once users have learned the design, how quickly can they perform tasks?
- Memorability: When users return to the design after a period of not using it, how easily can they re-establish proficiency?

- Errors: How many errors do users make, how severe are these errors, and how easily can they recover from the errors?
- Satisfaction: How pleasant is it to use the design?

Learnability is an indicator that shows how users reach optimal performance in interacting with a system (Joyce, 2019). Alternatively, it can be interpreted as how easy it is for users to accomplish a task the first time they encounter the interface and how many repetitions it takes to become efficient at that task. Learnability can be measured using learning time or the number of trials in a training session. However, different learning times are acceptable, depending on the type or purpose of the system.

Efficiency refers to how fast users can perform tasks once they have learned to use the system and is typically measured using the task completion time (TCT) (Dix et al., 2000). There are similarities between learnability and efficiency in that both dimensions use time as their measurement. However, there are some differences between them (Joyce, 2019). First, learnability is related to the first use, while efficiency focuses on the performance after users get used to the system. Second, learnability aims to assess if the system is learnable for target group users with learning curve estimation. Once the level of plateau is found, an analyst can evaluate learnability with the curve's slope, or the number of trials needed to pass a training session. However, when calculating efficiency, the analyst assumes that the users have already learned the system.

Memorability means the system should be easy to remember so that the users can return to the system after some time without learning everything all over again. This dimension requires the users to have some experience with the system; however, there is a gap between the last time they used the system and the time they are using it again. Thus, having users recognize the system rather than recalling it is recommended (Nielsen, 2005).

An *error* can be defined as a function performed by a user that does not lead to the aimed result. There are three categories under the error dimension with regards to usability. The first category is error frequency (Dix et al., 2000), which can be calculated as the number of errors in one trial and is used to measure the error ratio. The second category is error severity, which refers to the time duration between when the error happens and when that error ends (Albert and Tullis, 2013; Kim, 2005). The last category is error recovery, which can be part of error severity because it is related to how a user can effectively and efficiently return from the wrong track to the regular track to complete the task.

Finally, *satisfaction* refers to how pleasant the system is to use. Also, it refers to the level of comfort and acceptability of the system to its users and other people affected by its use (Dix et al., 2000). Thus, this usability dimension affects the motivation of use and is usually measured by rating scales such as the 7-point scale in Usefulness, Satisfaction, and Ease of Use (USE) survey (Lund, 2001). There are seven questions under the "Satisfaction" category of USE survey. Users need to evaluate a product or service using a score from 1 (strongly disagree) to 7 (strongly agree).

1.2.1.2. Usability of upper limb prosthetic devices

Usability of prosthetic devices has been measured by several subjective questionnaires such as the Client Satisfaction with Device module of the Orthotics and Prosthetic Users' Survey (CSD-OPUS) (Bravini et al., 2014). This is a self-report instrument for evaluating the outcomes (satisfaction) of prosthetics and orthotics. Another questionnaire used in this domain is Quebec User Evaluation of Satisfaction with assistive Technology (QUEST 2.0) (Demers et al., 2002). This questionnaire is designed for a person's evaluation of those distinct dimensions of the assistive device that are influenced by one's expectations, perceptions, attitudes, and personal values. The third questionnaire is USE, which measures subjective usability and has been applied for

evaluating prosthetic devices and other products (e.g., wearable devices, smartphones, website). Finally, the System Usability Scale (SUS), which is also designed for products or services (Brooke, 1996), has been used to assess the usability of prosthetic devices. SUS evaluates various products and services, including hardware, software, mobile devices, and websites.

Usability dimensions of each questionnaire are summarized in Table 1. These dimensions are categorized in terms of device appearance and subjective attributes. It was found that SUS and USE surveys are more focused on subjective attributes such as comfort, ease of use, satisfaction, and willingness to use as they are heavily used in assessing the usability of websites or software. Meanwhile, CSD-OPUS and QUEST 2.0 surveys incorporate physical attributes such as weight or aesthetic aspects of prosthetic devices.

Table 1. Comparison of usability surveys

Usability dimension		**CSD-OPUS**	**QUEST 2.0**	**SUS**	**USE**
Device appearance	Dimension		✓		
	Weight	✓	✓		
	Durability	✓	✓		
	Aesthetic	✓			
Subjective attributes	Comfort	✓	✓	✓	✓
	Ease of use		✓	✓	✓
	Satisfaction		✓		✓
	Willingness to use		✓	✓	✓

1.2.2. Cognitive Workload Assessment of Prosthetic Devices

CW of prosthetic devices can be measured using physiological measures, subjective measures, performance measures, and cognitive performance models (CPM) (Figure 1) (Park and Zahabi, 2022a). Physiological measures include various types of brain activity measures such as functional near-infrared spectroscopy (fNIRS), P200 (which represents some aspect of higher-order perceptual processing, modulated by attention), P300 (an event-related potential (ERP) component elicited in the process of decision making), late positive potential (LPP; an event-

related potential that reflects facilitated attention to emotional stimuli), and frontal theta/parietal alpha (FT/PA) (Deeny et al., 2014b; Leeb et al., 2015; Rezazadeh et al., 2011; Rupp et al., 2013; Shaw et al., 2018; Shaw et al., 2019a; Shaw et al., 2019b; Zhang et al., 2015; Zhang et al., 2018). A few studies used cardiac (Crea et al., 2017; Gonzalez et al., 2012a; Knaepen et al., 2015), respiratory (Gonzalez et al., 2012a), skin, and eye-tracking measurements (Parr et al., 2019). Skin measurements included skin conductance and temperature (Crea et al., 2017; Gonzalez et al., 2012a; Gonzalez et al., 2012b). Eye-tracking measures included the blink rate and pupil diameter (White et al., 2017; Zahabi et al., 2019b; Zhang et al., 2016b). Among all the CW measures, NASA Task Load Index (NASA-TLX) was the most frequently used method (Carlson et al., 2013; Davidson, 2017; Khalid, 2014; Pruziner et al., 2019; Ruiz Ramírez, 2016; Saraiji et al., 2018; Volkmar et al., 2019). The main reasons for the frequent use of NASA-TLX include its capability to assess CW in motor tasks (Berntsson, 2019; Hart, 2006; Hart and Staveland, 1988), the measure being non-intrusiveness, and its consideration of overall workload as well as the magnitude of each factor (Arenas, 2015; Bark et al., 2014).

Primary task measures were mainly used when the participants performed ADLs and were defined in terms of TCT and the number of transported items (e.g., clothespins) (Hargrove et al., 2018; Hargrove et al., 2017; Kuiken et al., 2015; Kuiken et al., 2016; Olsen et al., 2019; Raveh et al., 2018). CW was also measured using secondary task performance measures when the participants were asked to perform verbal, semantic, or numerical cognitive tasks along with the ADLs or other primary tasks during the experiment (e.g., participants counted backward from 100 to 1 with three steps while they were moving an object with their prosthetic device (Resnik et al., 2018)).

One study used the CPM approach to determine the CW of prosthetic devices (Zahabi et al., 2019b). The finding of this study comparing DC and PR control modes suggested that CPM approaches such as GOMS language (GOMSL) models can be used to predict cognitive demands of using upper-limb prostheses (Zahabi et al., 2019b).

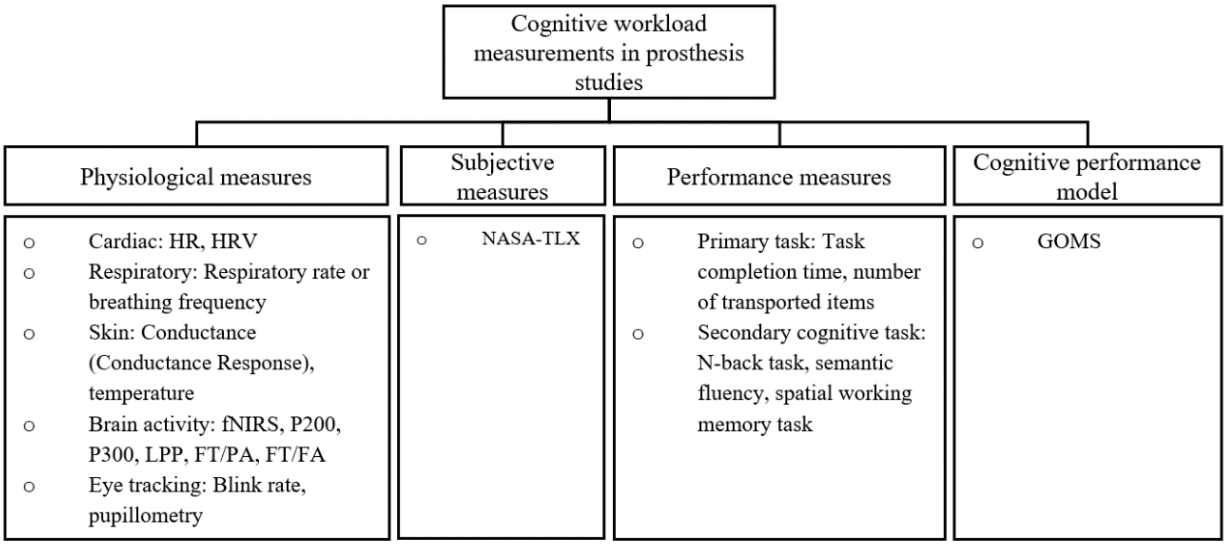


Figure 1. Cognitive workload measurements in prosthesis studies (Reprinted from Park, J., & Zahabi, M. (2022a). © 2023 IEEE)

A detailed comparison of these techniques based on sensitivity, intrusiveness, cost, and accuracy can be found in Park and Zahabi (2020). Physiological measures allow understanding of psychological processes through their effect on the body, rather than through task performance or perceptual ratings (Cain, 2007). Therefore, the principal advantage of physiological measures is that these measures are continuous and objective. However, some signals can be contaminated by head or body movements (e.g., neuroimaging or EEG measures), especially in experiments using prosthetic devices or electrode caps (Zahabi et al., 2019b).

Most studies used NASA-TLX to measure CW since the method is unobtrusive and can be easily collected after the experiment sessions. Subjective measurement techniques such as NASA-TLX quantify humans' understanding and judgments of their experienced demand. While these

methods have high face validity, their interpretation and ability to predict performance are uncertain (Cain, 2007). These measures also provide discrete rather than continuous values, and prior studies have found a dissociation between subjective and performance measures (Yeh and Wickens, 1988). Furthermore, subjective measures are limited due to recall bias and substantial individual differences (Hart, 2006).

Performance measures are classified into two major categories, including primary and secondary task measures. Primary task measures evaluate the operator's performance on the task of interest. Examples of primary task measures of workload include speed, accuracy, reaction or response time, and error rate. Secondary task measures provide an index of the remaining operator capacity while performing primary tasks and are more diagnostic than primary task measures (Cain, 2007). Examples of secondary tasks include n-back, verbal shadowing, and pursuit tracking tasks. Performance measures have advantages in that they evaluate participants' performance on the task of interest directly, and this is useful where the demands exceed operators' capacity such that performance degrades from baseline or ideal level (Cain, 2007). However, the complexity of the secondary tasks or environmental conditions can reduce walking performance or the primary task (Morgan et al., 2014). In addition, if the amputees are acclimated to the prosthesis and the environment is stable, cognitive load can be limited. Therefore, physiological measurements can be used instead of task performance measurements to capture subtle changes in CW under these conditions. Performance measures often lack scientific rigor, making interpretation of the results difficult. Unknown or uncontrolled factors may affect results rather than the intended manipulations in the study. Also, due to the protective (or compensatory) effect of increased effort in the task, measuring performance might not be sufficient to evaluate the participant's state. For

example, the performance does not reflect information about the costs involved in the adaptive response to stress (Cain, 2007).

One possible alternative to assess CW is the CPM method. Although CPM was used only in a case study with one amputee participant under DC and PR conditions, the method can be applied to other configurations and experimental conditions, considering its capability to predict task performance and calculate memory chunks (Zahabi et al., 2019b). The models can calculate task performance, the number of cognitive/perceptual/motor operators, and memory chunks to identify bottlenecks in the task. These models have been widely applied in other domains such as human-computer interaction research, aviation, health care, usability testing, and cybersecurity (Din, 2015; Prada and Boehm-Davis, 2004; Rosyidah et al., 2019; Stanley et al., 2019; Zahabi and Lyman, 2019; Zahabi and McCollum, 2019; Zahabi et al., 2019a). However, it is essential to note that CPM approaches assume expert performance, and therefore, the methods might have limited application to novice prosthetic users. Table 2 summarizes the comparison among different CW assessment techniques.

Table 2. Comparison of CW assessment techniques (Reprinted from Park, J., & Zahabi, M. (2022a). © 2023 IEEE)

Technique	**Pros**	**Cons**
Physiological measures	Continuous & objective (Cain, 2007; Zahabi et al., 2019b)	Intrusiveness, susceptible to temperature and humidity (Charles and Nixon, 2019)
Subjective measures	High face validity (Zahabi et al., 2019b)	Discrete, ability to predict task performance is uncertain (Cain, 2007; Yeh and Wickens, 1988), recall bias and individual differences (Hart, 2006)
Task performance measures	Useful to test changes of CW using direct modification on the task (Zahabi et al., 2019b)	Lack of interpretability (Cain, 2007), lack of scientific rigor, and plausible compensatory effect (Zahabi et al., 2019b)
Human performance modeling	High interpretability, non-intrusive, high versatility (can be edited and embedded in various situations) (Zhang and Wu, 2017)	Need time and effort to learn the modeling techniques, need to be validated with human subject data

1.2.3. Human Performance Models

Human performance models (HPM) were reviewed as a basis for the proposed method and due to the following reasons. First, conducting human subject experiments with prosthetic users is difficult due to the intrusiveness of study equipment (e.g., cables and sensors). This could be a reason why most previous studies relied on questionnaires. Second, HPM has the potential to assess the usability of prosthetic devices without human-subject experiments. Some prior studies illustrated HPM's capability to predict device usability and cognitive workload (Zahabi et al., 2019b). Third, HPM can compensate for the limitations of physiological and subjective measurement techniques. HPM is a modeling approach that can be generated by observing the user performing some tasks without any interruption and therefore is a non-intrusive approach. Therefore, in the following sub-sections, a review of HPM approaches is provided.

1.2.3.1. Human performance modeling approaches – GOMS family

There are five major approaches in HPM (Gil, 2010; Kotseruba and Tsotsos, 2016; Van Rijn et al., 2011; Yuan et al., 2020), including Goals, Operators, Methods, and Selection rules (GOMS) (Card et al., 1983), Adaptive Control of Thought (ACT-R) (Anderson et al., 1997), Executive-Process Interactive Control (EPIC) (Kieras and Meyer, 1995), Queueing Network – Model Human Processor (QN-MHP) (Liu et al., 2006), and State, Operator, And Result (SOAR) (Laird et al., 1991). Among these methods, GOMS language was the only method applied in the prosthetic device domain.

GOMS is a human information processor model for human-system interaction that explains a user's cognitive process using four components: Goals, Operators, Methods, and Selection rules (Card et al., 1983). GOMS works based on the Model Human Processor (MHP) theory (Card et al., 1986b). *Goals* are symbolic structures that establish a state to be achieved and determine a set

of possible methods by which it may be accomplished. *Operators* are fundamental perceptual, motor, or cognitive acts whose execution is needed to change any aspect of the user's mental state or affect the task environment. *Methods* describe a procedure for accomplishing a goal. Finally, *Selection Rules* are needed when a goal is attempted, but more than one method is available to the user to accomplish it. There are numerous extensions from GOMS, including Keystroke-Level Model (KLM) (Card et al., 1980), Cognitive-Perceptual-Motor GOMS (CPM-GOMS) (John, 1990), GOMS Language (GOMSL) (Kieras, 2006; Kieras, 1988), Natural GOMSL (NGOMSL) (Kieras, 1994), and Enhanced GOMSL (E-GOMSL) (Gil, 2010). Using an interactive computer system, KLM predicts how long it takes an expert user to accomplish a routine task without errors. CPM-GOMS is an advanced method in that it can model parallel processes. NGOMSL is a high-level (natural language) syntax for GOMS representation, whereas GOMSL is an executable form of NGOMSL and a computationally realized version of MHP.

Methods in the GOMS family are fast and straightforward to use as compared to other modeling approaches. User-defined operators can be added. However, the models do not include errors and their resolution is low (i.e., describes interaction broadly). Also, GOMS family does not involve detailed calculation of memory chunks.

1.2.3.2. Other modeling approaches

ACT-R provides models of elementary and irreducible cognitive and perceptual operations that enable human information processing. In theory, each task that humans can perform consists of a series of these discrete operations (Anderson et al., 1997). ACT-R's primary time unit is 50ms which can describe human information processing in a fine-grained resolution. In addition, ACT-R can generate essential outcomes such as time to perform a task and accuracy. There are several advantages of using ACT-R, such as modeling of parallel activities (Yuan et al., 2020), memory

lapse and loss (Leiden et al., 2001), reinforcement learning (Van Rijn et al., 2011), memory retrieval (Samsonovich, 2015), and emotions (Ritter et al., 2019). However, this method has several limitations including: (1) it takes a long time to model (at least several days to weeks of using the system and takes months to years to become an expert in its use) (Salvucci and Lee, 2003); (2) ACT-R is suitable to model tasks under ten seconds (Ritter, 2009); (3) it is a rule-based system which requires analysts' manual input (Jang et al., 2011).

QN-MHP is a computational cognitive architecture that integrates the mathematical framework of queueing network theory with the Model Human Processor (Liu et al., 2006). Based on a network structure of twenty process units, different cortical areas of the human brain and corresponding functional modules of human information acquisition, processing, and implementation are simulated. Because of this "brain-like" structure, QN-MHP can visualize internal information flows during the simulation of related activities. However, its inability to generate or model complex cognition such as language comprehension or problem-solving requires creating new rules by the model itself rather than relying only on the rules preprogrammed by the model developer (Liu et al., 2009).

EPIC is a general framework, represented as a simulation modeling environment, in which models of human performance in specific tasks may be constructed (Kieras and Meyer, 1995). EPIC focuses on perception and motion (Taatgen and Anderson, 2010; Yuan et al., 2020). The detailed description on perception and motion was influential for ACT-R and SOAR that incorporated perceptual and motor components into the models (Ritter et al., 2019). Similar to ACT-R and SOAR, EPIC encompasses a production-rule system (a "cognitive processor") that provides procedural knowledge. There are also "perceptual processors" that process different sensory (tactile, visual, and auditory) information. The outputs of the perceptual processors are

sent to the working memory. In addition, there are two types of working memory (unrelated to the sensory-motor information): one storing the current goals and steps to reach them ("control store"), and a "general" working memory for miscellaneous information.

SOAR is a functional approach to understand what cognitive mechanisms underlie intelligent human behavior (Laird et al., 1991). Also, it is an architecture for human cognition expressed in the form of a production system. SOAR can represent extensive and complex rule sets (Kieras, 2005). Its primary use is in artificial intelligence (AI) but also in cognitive modeling. In addition, it has been combined with EPIC's perceptual-motor processors. The method focuses on problem-solving mechanisms and can develop AI agents that solve problems based on different types of knowledge, whether programmed or learned by the system. Table 3 provides a comparison of these models.

Table 3. HPM model comparison

CPM / **Factors**	**CMN-GOMS**	**NGOMSL**	**CPM-GOMS**	**ACT-R**	**EPIC**	**Soar**	**QN-MHP**
Resolution	Low	Low	Moderate	High	High	High	Low
Parallel processing	×	×	✓	×	✓	✓	✓
User-defined operator	✓	✓	✓	×	✓	✓	✓
Modeling difficulty	Low	Low	Low ~ Medium	High	High	High	Medium
Open source	✓	✓	✓	✓	✓	✓	×
Applied in prosthetic device domain	×	✓	×	×	×	×	×

1.3. Research Gaps

This study aims to fill several research gaps in the literature. The first research gap is regarding the limitations of the nature of subjective usability and CW assessments. Current

approaches for measuring usability of prosthetic devices with questionnaires are simple, easy to administer, and have high face validity. In addition, they do not interfere with the study since the data are collected after the experiment sessions. However, their interpretation and ability to predict performance are uncertain (Cain, 2007). These measures also provide discrete rather than continuous values, and prior studies have found a dissociation between subjective and objective performance measures (Yeh and Wickens, 1988). Furthermore, subjective measures are limited due to recall bias and substantial individual differences (Hart, 2006).

The second gap is the lack of using an advanced HPM approach to assess the usability and CW of prostheses. As mentioned in Section 1.2.2., only GOMSL has been used to measure TCT and the number of cognitive, perceptual, and motor operators. However, the model still lacks consideration of parallel operations and memory processes. Other advanced modeling techniques might be more suited to this application, such as ACT-R and CPM-GOMS.

The third gap is regarding the usability of the HPM tool itself. Although there are a number of software applications for HPM, many of them are difficult to use for analysts. Unlike general programming languages such as C, C++, or Java, modelers in the HPM domain need to study a specific programming grammar and should be familiar with human factors concepts and theories (e.g., perception, cognition). Although there are some graphical user interface (GUI)-based modeling tools (e.g., CogTool), they are not directly applicable for prosthetic device analysis but for applications such as mobile devices or surface transportation (John, 2005; Salvucci et al., 2005). Therefore, there is a need to have a user-friendly interface to use HPM for prosthetic device usability analysis.

1.4. Research Objectives

Based on the identified gaps from the literature review, the main objective of this research was to develop a model with an advanced HPM approach to assess the usability of prosthetic devices via a user-friendly interface. The model was validated with two human subject experiments (one with a physical prosthetic device and another using virtual reality simulations) and the performance of the model was compared with a benchmark model. Figure 2 illustrates the research objectives stemmed from the research gaps.

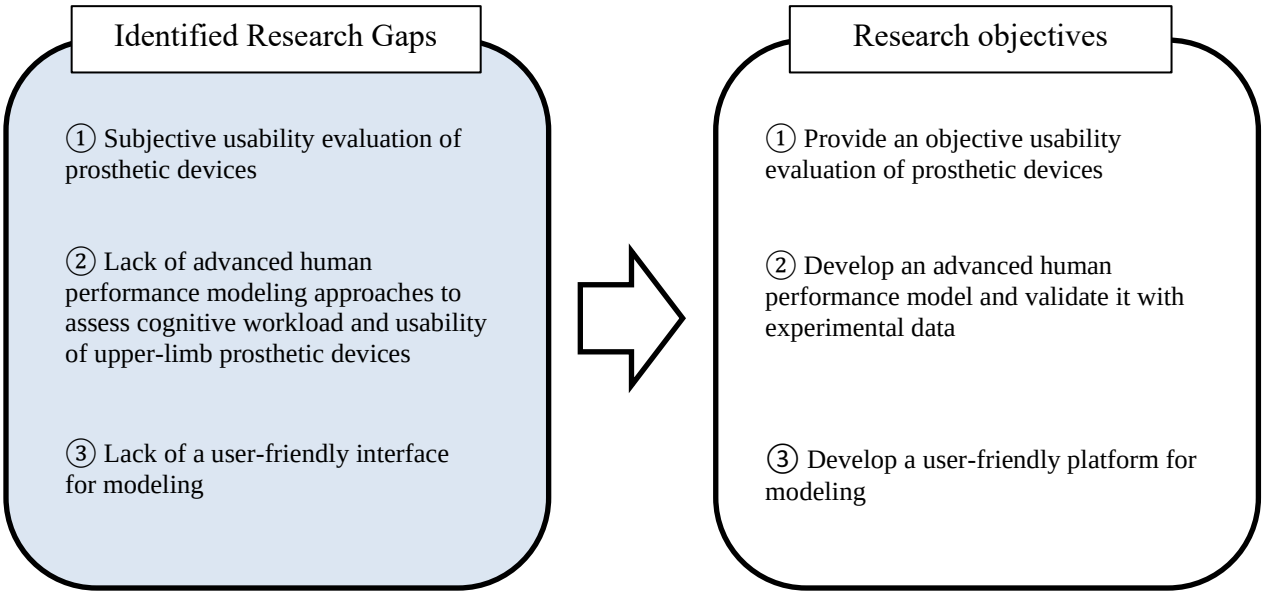


Figure 2. Identified research gaps and the plan to address those gaps

2. RESEARCH SCOPE

2.1. HPM-UP Development

The Human Performance Model for Upper-limb Prosthesis (HPM-UP) was developed based on a hybrid modeling approach including top-down (theories in human factors, ergonomics, and psychology) and bottom-up (data-driven) approaches. The HPM-UP uses CPM-GOMS and ACT-R declarative memory modules. For example, participants might frequently adjust the hook while they are looking at it. In this case, perception and motor operators work simultaneously, which is an example of parallel activities that CPM-GOMS can model. In addition to the CPM-GOMS logic, declarative memory module functions from ACT-R 7.0 were used for simulating the number of memory chunks (Bothell, 2017; Leiden and Best, 2005). For example, while performing ADLs, participants had to remember a particular device configuration to adjust the hook and complete the tasks.

R software package 4.0.5 was used for model development. R is a worldwide package for research, especially for statistical analysis. The HPM-UP package can be downloaded for free and therefore, is accessible for researchers and usability analysts. In addition, the package includes a GUI for analysts to use. The Details of the model development process are described in Chapter 3.

2.2. Usability Dimensions in HPM-UP

The HPM-UP evaluates six usability dimensions. The dimensions came from Nielsen's usability principles (Nielsen, 2012) as they have been frequently used in evaluating the usability of other products and services. In addition, "cognitive workload" was explicitly included in the

HPM-UP as a sixth dimension. For the HPM-UP, there was a need to convert the general definitions of Nielsen's definition to computational variables related to prosthetic device application. Therefore, the modified usability dimensions were defined as follows:

- *Learnability*: The number of training trials required to pass the training criteria
- *Error (Error Rate)*: The error rates in performing a task with a prosthetic device
- *Memory Load*: The number of memory chunks stored in working memory when performing a task with a prosthetic device
- *Efficiency*: Task completion time of one ADL cycle (e.g., moving one pin from one bar to another bar)
- *Satisfaction*: The relationship among perceived performance, expectation, and desire
- *Cognitive Workload*: Classified cognitive workload level (e.g., "High", "Moderate", "Low")

2.3. Model Validation

To validate the HPM-UP, two human subject experiments were conducted. A recent literature review found that there are two types of output controls for the upper limb prosthetic studies including: (1) physical devices, and (2) virtual environment (Park and Zahabi, 2022a). Therefore, the first experiment used a physical prosthetic device (Chapter 4), and the second experiment was conducted using a virtual reality (VR) simulation (Chapter 5).

3. MODEL DEVELOPMENT

3.1. Overview

Although several existing questionnaires such as the SUS and USE survey can be used to assess the usability of prosthetic devices, and CW of these device can be measured subjectively using validated subjective ratings such as NASA-TLX, these methods are mainly used toward the later stages of the design and development process when there is a functional prosthetic device and there is a need for conducting a human subject study which can be costly and time consuming. Furthermore, there could be self-report biases with survey responses. The main motivations for the HPM-UP approach are to overcome these limitations with subjective usability and CW measurement techniques in prosthetic assessments and to predict usability and CW of prostheses in early stages of the design process before conducting human subject experiments. This can save time and energy and reduce the workload of experimenters, clinicians, and amputee patients. HPM-UP can also provide usability and CW estimates for future prostheses with novel control schemes before entering the physical device development phase. At this stage, conducting human subject experiments may not be feasible. The logic behind the HPM-UP methods is based on theories and is transparent so that other researchers can update the parameters or configurations of prostheses with collaborations with clinicians or device designers to better fit their needs.

Figure 3 illustrates an overview of how HPM-UP works. Methods including CPM-GOMS, error probability modeling technique, memory function (from ACT-R), learning curve, machine learning algorithms, and satisfaction formulas have been developed and integrated into the HPM-UP. Users of HPM-UP can run the model using a scenario (i.e., by conducting a task analysis) and

selecting the parameters (e.g., prosthetic device control). Then, the HPM-UP generates six usability dimensions. Each usability dimension is described in detail in the following sub-sections.

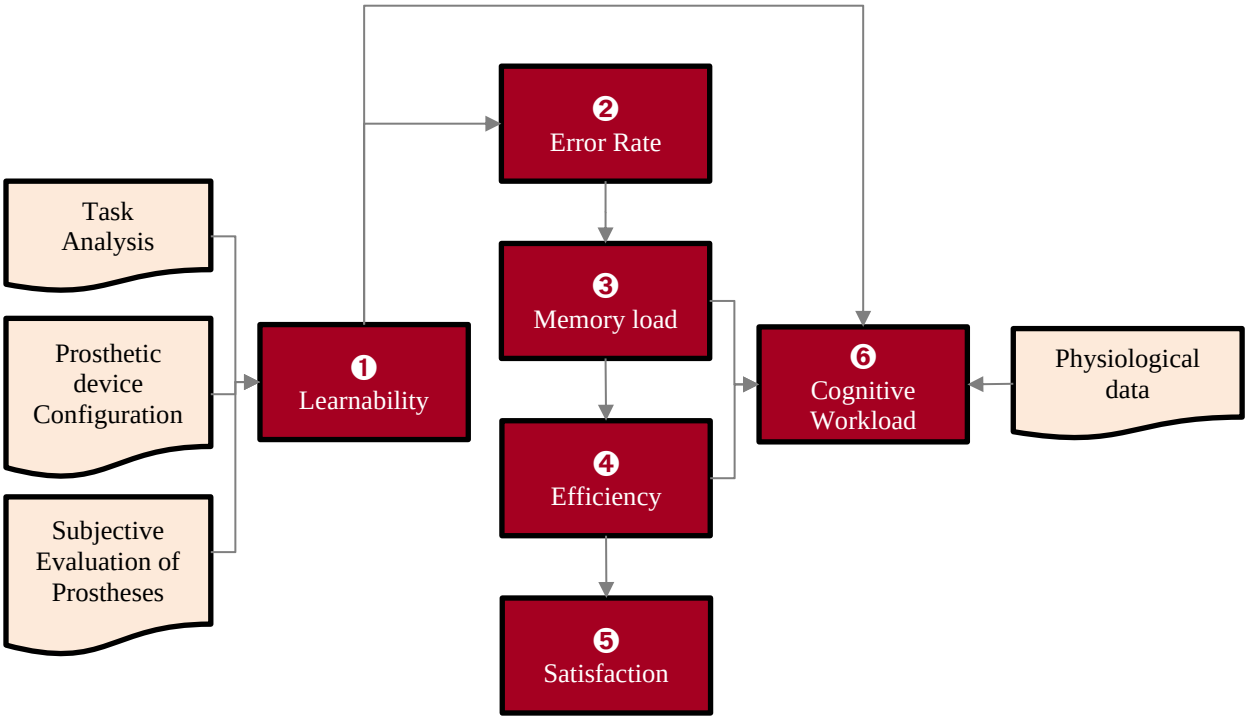


Figure 3. HPM-UP overview

3.2. Usability Dimensions

3.2.1. Learnability

3.2.1.1. Learning curve equation

The learnability dimension in HPM-UP is defined as the number of training trials to pass the training threshold level. This definition came from the original definition of learnability: "*Learnability considers how easy it is for users to accomplish a task the first time they encounter the interface and how many repetitions it takes for them to become efficient at that task*" (Albert and Tullis, 2013; Newell and Rosenbloom, 1981). Thus, learnability in HPM-UP can be defined as the number of trials required to reach a plateau.

The original learnability equation is based on the learning curve's unit theory (Camm, 1985; Mislick and Nussbaum, 2015; Zhang et al., 2016a) which is defined as Equation (1).

$$Y_x = Ax^b \quad (1)$$

In equation (1), Y_x is the cost of unit x, A is the theoretical cost of unit 1, x is the unit number, and b is a constant representing a slope. In HPM-UP, x is the number of training trials which is the outcome of learnability. A is the task completion time for the first trial. Y_x can be replaced with the task completion time in each trial and can be gathered from training trials. The approach used to calculate A and b will be discussed in the following subsections.

3.2.1.2. Patterns of task completion times during training trials

The training data from 10 participants were used to investigate the patterns in TCT (Figure 4). Then, average of the past three datapoints were plotted in Figure 4. In this figure, the X-axis indicates the required training trial numbers. For example, the task completion time in trial 5 refers to the average task completion time of trial numbers 3, 4, and 5. Thus, a total of five trials were needed for training.

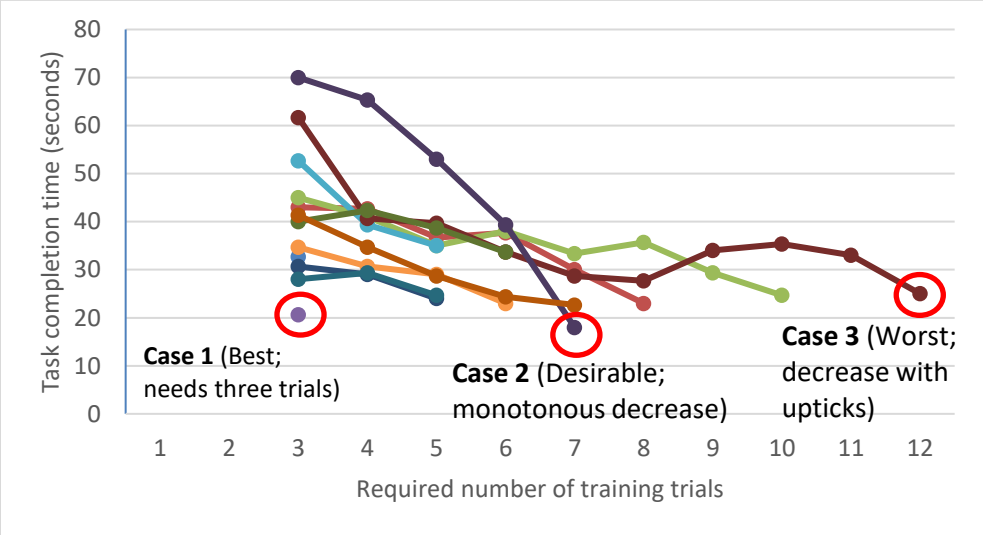


Figure 4. Task completion time pattern in training trials

The TCT in training trials indicates three distinct trends (Figure 4). Case 1 is illustrated as a single point, which means the participant required only three trials to pass the training threshold (best case scenario). Case 2 illustrates a monotonous decrease (i.e., the TCT continuously decreases, which is desirable). Finally, case 3 is the most complex trend (i.e., both increasing and decreasing trend, the worst case).

3.2.1.3. Modified *A* (i.e., task completion time for the first trial)

As shown in Figure 4, TCT of the first trial (i.e., *A*) is different across participants, which might be due to the device calibration quality (DCQ). There are two types of DCQ including: Objective DCQ (ODCQ) and subjective DCQ (SDCQ). ODCQ can be calculated using Equation (2) (Music, 2022).

$$ODCQ = \frac{The\ number\ of\ correctly\ matched\ input\ gestures}{The\ number\ of\ all\ input\ gestures} \qquad (2)$$

ODCQ can be used to estimate *A* for each participant, however, it requires collecting data from all participants and conducting a detailed video analysis, which might not be suitable for the HPM-UP package as the goal of this package is to use it in early stages of the design and development process of prosthetic devices. Additionally, video analysis cannot guarantee if specific activities were intended or not by the participants. For example, although a gesture in the video looks erroneous, that could have been exerted intentionally by the participants.

Therefore, the SDCQ was used to adjust *A*. Having high SDCQ could result in small *A* (i.e., high subjective level of device calibration quality could lead to shorter TCT in the first trial). However, subjective evaluations can have *self-report bias* (VandenBos, 2007). To mitigate the self-report bias, the *first impression* of the prosthesis was also considered. *First impression* (FI) refers to one's initial perception of a person (or object), typically involving a positive or negative

evaluation as well as a sense of physical (or psychological) characteristics (VandenBos, 2007). In this study, although all of the participants passed the training criteria, the number of required training trials were different among participants, which led to having different error rates, or efficiency in the experimental trials. These differences could have been caused from different first impressions of the device.

The FI was quantified based on the findings of previous studies. Human builds impression from a set of personality traits resulting from incomplete information – implicit personality theory (Beauvois, 1982; Schneider, 1973). Second, the built impression formation can be classified or unified – impression formation of personality (Asch, 1946). Third, one salient characteristics in one area (i.e., FI) can affect other dimensions (i.e., halo effect) (Clifford and Walster, 1973; Thorndike, 1920). Fourth, humans tend to seek or interpret any evidence in favor of his or her FI - confirmation bias (Nickerson, 1998; Snyder and Swann, 1978). Thus, if the first impression is positive, the individual will tend to minimize the negative aspects of the surrounding elements and exaggerate the positive aspects. Conversely, the more negative FI, the more the individual will tend to minimize the positive and accentuate the negative aspects. Lastly, the time to form the first impression has been studied and it was found that the impression formation can be done quickly (i.e., at the very early stage of the exposure to the stimulus) (e.g., 100 milliseconds) (Fiske and Neuberg, 1990).

The concept of FI is also applied in the human-computer interaction (HCI) domain. Some studies assessed the first impression of websites using a 7-point Likert scale within 5 seconds of its use (i.e., 5 Second Test; 5ST) (Gronier, 2016). The study suggested that there was no significant difference between some typical usability tests (i.e., questionnaires asked after the entire experiment) and the 5ST approach. Furthermore, based on the confirmation bias, the direction of

the first impression (i.e., positive or negative) can affect the usability of a product (Michalco et al., 2015; Raita and Oulasvirta, 2011), including easiness (Hassenzahl and Monk, 2010), satisfaction (Liu et al., 2010), and reliability (Kim and Fesenmaier, 2008).

Therefore, the *A* factor was adjusted (A') based on the *FI* and *SDCQ* (Equations (3)). If *SDCQ* is 1, which means users think the DCQ is perfect or has no errors, the *FI* is the only concern. However, if *SDCQ* is for example 0.5 (e.g., 5 out of 10 gesture inputs were correct), the *A* factor will be doubled. The *FI* factor is a number between 0 and 2. If the *FI* is less than 1 (i.e., the first impression is negative), *A* will be decreased. If *FI* is greater than 1 (i.e., the first impression is positive), *A* will be increased.

$$A' = A \times \frac{FI}{SDCQ} \quad (3)$$

SDCQ in this study was calculated based on the average of the responses to questions Q3 (easiness in adjusting the device (fixing, fastening)), Q6 (easiness of using the device), and Q8 (effectiveness of using the device (the degree to which the device meets a user's needs)) of the USE questionnaire. *FI* was calculated from the difference between the SDCQ and participant's training performance. This indirect approach was used to avoid self-report bias (VandenBos, 2007) that could have occurred if *FI* was measured directly.

3.2.1.4. Modified slope

The slope is represented by *b* as shown in Equation (4).

$$\text{slope of learning curve} = \frac{cost\ of\ unit\ 2n}{cost\ of\ unit\ n} = \frac{A(2n)^b}{A(n)^b} = 2^b \quad (4)$$

$$\ln(slope) = b \times ln(2)$$

$$\therefore b = \frac{\ln(slope)}{\ln 2}$$

Previous studies found that the slope could be estimated from the reference value for each industrial domain (Mislick and Nussbaum, 2015) such as:

- 85% Aircraft industry
- 80~85% Shipbuilding
- 75~85% Electrical
- 90~95% Electronics
- 90~95% Machining
- 88~92% Welding

Also, it can be estimated based on the degree of automation.

- 70% = entirely manual operations
- 80% = 75% manual + 25% automated
- 85% = 50% manual + 50% automated
- 90% = 25% manual + 75% automated

Since the tasks in this study were performed by prostheses, ideally, the slope could be 0.7 (i.e., entirely manual operation) when the calibration quality is perfect (Mislick and Nussbaum, 2015) (i.e., the prosthesis could always be controlled based on user's input). However, since calibration quality could be varied, there is a need to adjust the slope. If SDCQ becomes 0, the slope changes to 0.90 which means there is almost no learning occurred between the trials (Mislick and Nussbaum, 2015). Therefore, Equation (5) shows the linear relationship between the SDCQ and slope.

$$slope = -0.2 \times \mathrm{SDCQ} + 0.9 \quad (5)$$

Equation (5) does not include a potential effect of physical and/or mental workload on device learnability. Therefore, two parameters were added to differentiate the slope based on the level of workload as shown Equation (6).

$$slope = -0.2 \times R_1 \times \mathrm{SDCQ} + (0.9 + R_2) \quad (6)$$

The initial values for $R_1$and R_2 parameters were determined based on the pilot study with 10 participants. R_2 refers to the baseline physical and mental workload of participants before the experiment. R_1 was determined based on the changes of physical and mental workload from the baseline.

3.2.1.5. Revised learning curve

Based on the changes in slope and *A* factor described earlier, the learning curve equation was revised to estimate learnability as shown in Equation (7).

$$L = A'x^b = A\frac{FI}{SDCQ}x^{\frac{\ln(-0.2R_1SDCQ+(0.9+R2))}{ln2}} \quad (7)$$

3.2.2. Error Rate

3.2.2.1. Limitations of previous studies

Including errors in the HPM-UP for analysis of prosthetic devices is necessary since unlike the direct human interaction with a mouse or keyboard controlled with high precision, a prosthesis device uses an EMG signal that does not always correctly project user's intentions to the outcome (i.e., hook movement). That is, there is a mediator (prostheses) between the human and the task, which can affect the task performance. There can be unintended or due to wrong hook movements

(e.g., due to the EMG signals not correctly classified or muscle fatigue). To have a more precise task performance calculation, it was necessary to account for errors in HPM-UP.

Some prior studies tried to estimate errors in using protheses (Hargrove et al., 2007; Lock et al., 2005; Scheme and Englehart, 2011). For example, Scheme and Englehart (2011) suggested to use active and total error to evaluate pattern recognition LDA classifiers instead of using offline classification accuracy, as prior work has shown that offline classification accuracy, while easy to calculate, has a weak correlation to prosthesis usability (Lock et al., 2005). Active error is calculated by the ratio of the number of incorrect active decisions and the number of total active decisions (Music, 2022). Total error rate is the ratio between the number of incorrect decisions (i.e., mismatches between the input gesture and hook movement) and the number of total decisions. Active error rate reflects the negative effects of corrective actions to the total error rate (Scheme and Englehart, 2011). These corrective actions must occur after a prosthesis user has made an inadvertent action that is not required to complete the task and therefore, can increase frustration (Hargrove et al., 2007). Therefore, active error rate can provide a useful metric for measuring this frustration and can be informative when assessing prosthesis usability.

However, this approach to estimate errors has some limitations. First, it is challenging to clearly figure out whether a specific gesture is an active or inactive decision from the human subject experiment. Even by analyzing the log data which were generated from software such as MATLAB or videos after our pilot tests, it was not possible to differentiate which hook movement came from active decisions. For example, sometimes participants needed to intentionally test the synchronization between their hand gestures and hook movement, therefore, they performed some gestures. These actions should not be counted as errors. In addition, there were some circumstances that the hook moved correctly without the related hand gestures. Unlike the training, in the

experimental trials, participants were able to discover their individual strategies to complete tasks efficiently (i.e., minimal hand or muscle movements). These also could not be considered in counting the active error rate or total error rate. The second limitation of the previous approach to calculate errors was that it only considered the pattern recognition (PR) configuration, which again might limit the generalizability of the approach to another prosthetic device configuration such as DC or CC.

3.2.2.2. Error rate based on learnability

To address the identified limitations, the error rate in HPM-UP was formulated differently from the Scheme and Englehart (2011) study . The error rate in HPM-UP depends on learnability (e.g., participants who reached the training criteria faster exhibited few numbers of errors in experimental trials)

Error rate was estimated based on learnability (i.e., Err(L)) because of the causal relationship between two dimensions. If $\mathrm{L} = 1$, (i.e., learnability is 100%), the estimated error rate in the experimental trials will be 0 (i.e., *Err*(1) = 0). If $\mathrm{L} = 0$, *Err*(0) will be 1 which means that participants will make errors during the experimental trials since they failed to learn how to use the device during the training. The error rate follows the natural exponential function as shown in Equation (8). This exponential curve was fitted based on the results of pilot testing using the calculated learnability and error rate observed during the experiment.

$$\mathrm{Err(L)} = max\left\{\frac{1}{1-e}(e^{L} - e), 0\right\} \quad (8)$$

3.2.3. Memory Load

Declarative memories are the kind of memories that can be declared (e.g., the name of one's fifth grade math teacher). This section describes efforts of humans when they retrieve

information from declarative module (or a particular region of the brain), including activation, memory load, and recall probability. The equations are based on the declarative memory structure implemented in ACT-R (Dehban et al., 2016).

Activation is a degree of association between previous experiences and current context which describes whether a chunk will be helpful at any given moment (Bothell, 2017). Chunks are the elements of declarative knowledge in the ACT-R theory and are used to communicate information among modules through the buffer (Bothell, 2020). The activation of a memory trace is calculated using the Equation (9) based on Altmann and Schunn (2019) and Estes (2015):

$$Activation = \ln\left(\frac{n}{\sqrt{T}}\right) \quad (9)$$

where *n* is the number of times that chunk is rehearsed, and *T* is the total time the trace is held in memory (or age of the item). The number of rehearsals refers to the familiarity of a chunk. The default value of this parameter was set as 3, a plausible level of rehearsal that exhibited the best overall fit (Estes, 2015). However, for the information from long-term memory (LTM), or recall information, the number of rehearsals was set to 10 to indicate that a chunk from LTM is difficult to be forgotten (Estes, 2021). This number can be updated while the model is running. For example, if the number of rehearsals increases, the activation also increases, leading to higher recall probability.

In order to mimic the division of activation across all working memory chunks, the activation was reduced as a function of the number of chunks in the problem span based on the logic in Cogulator software (Estes, 2021). Therefore, the divided activation in HPM-UP was calculated as Equation (10) (Estes, 2015).

$$Divided\ activation = Activation + \frac{1}{stack\ depth} - 1 \quad (10)$$

"Stack depth" is the location or address of the chunk in the memory. For example, if a chunk was added as a third chunk in the memory, stack depth becomes 3. It also refers to the number of chunks that activation must be divided into. The idea of limited activation source pools and their distribution among all the chunks held in working memory has been previously documented in the literature (Anderson et al., 1996). Equations (10) and (11) allows HPM-UP to model a relationship between the number of chunks to be memorized and decay of chunks over time. Memory load was defined as the overall occupancy of chunks in the entire task and was calculated using Equation (11), which divides the summation of the duration of all chunks by the total task duration (Estes, 2015; Estes, 2021).

$$Memory\ load = \frac{\sum_{i=1}^{7}(chunk\ duration)_i}{total\ task\ duration} \quad (11)$$

3.2.3.1. Chunk structure and lifecycle

As shown in Figure 5, there are seven slots for chunks in Cogulator (Miller, 1956). The first chunk of information goes to the bottom, and the next chunk can be stacked on the top of the first chunk. Once all slots are filled with chunks and a new chunk of information arrives, the chunk with the lowest activation decays and the chunks' positions after the decayed chunk are updated. Then, the new chunk is added to the last slot (i.e., top stack).

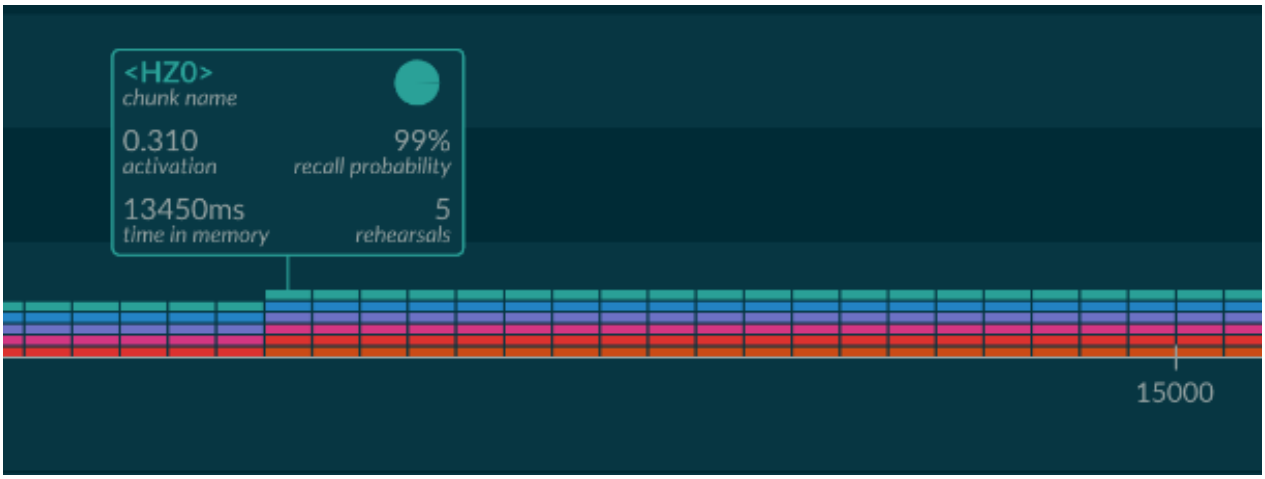


Figure 5. Memory chunk structure in Cogulator

The HPM-UP shows the lifecycle of a chunk in a table format in its output window. The table includes the chunk's name, the time when it was first stored in the short-term memory, the time when it decayed, and its duration (Figure 6).

	chunk_num	chunk_name	stack_depth	pushed_time.Global.	elapsed_time.Local.	rehearsal	activation	prob_recall
1	1	<gesture - close>	1	784.000	119912.5	18	0.49699079	0.9994388
2	2	<mode - 1>	2	834.000	119862.5	29	0.47412339	0.9993709
3	3	<gesture - rotation>	3	4939.100	115757.4	18	-0.15204302	0.9857940
4	4	<mode - 2>	4	5773.734	114922.7	21	-0.07760751	0.9901654
5	5	<gesture - open>	5	7623.734	113072.7	18	-0.27364382	0.9742135
6	6	0	0	0.000	0.0	3	0.00000000	0.0000000
7	7	0	0	0.000	0.0	3	0.00000000	0.0000000

Figure 6. Memory chunk structure in HPM-UP

3.2.3.2. Recall probability

Based on the divided activation calculated from Equation (10), recallability was calculated as Equation (12) (Dehban et al., 2015; Estes, 2021).

$$P_i = \frac{1}{1 + e^{\frac{\tau - A_i}{s}}} \qquad (12)$$

In this equation, τ is a threshold to forget the chunk (-1), *s* is the noise or the variance from one scenario (refers to perceptual/cognitive/motor operators, methods, and selection rules used by

an individual to accomplish a specific goal) to another (Bothell, 2017), which is set to 0.8 based on Estes and Masalonis (2003), and A_i is the divided activation value from equation (10). Finally, P_i refers to the recall probability of the i^{th} chunk. The default value of threshold and noise came from the Cogulator software as the model assumes the user is an expert.

3.2.4. Efficiency

Efficiency was calculated using Equation 13 and based on the task completion time and considering error rate. HPM-UP calculates the efficiency of an expert (without errors) based on the task completion time estimates from the CPM-GOMS method and inflates the task time based on error rate to estimate efficiency for all users.

$$Eff(Err) = Eff\{Err(L)\} = (Expert's\ efficiency) \times \{1 + Err(L)\} \quad (13)$$

3.2.4.1. Perceptual and cognitive operators

The duration of general operators follows the established time durations in Cogulator (Estes, 2017). Table 4 includes some of the perceptual and cognitive operators with their established time durations.

Table 4. Perceptual and cognitive operators

Operator	Duration (ms)	Operator	Duration (ms)
Look	550	Recall	550
Attend	50	Think	1250
Store	50	Verify	1250

3.2.4.2. Specific motor operators for prosthetic devices

Unlike perceptual and cognitive operators, specific motor operators for modeling human interaction with prosthetic devices do not exist. Thus, the HPM-UP incorporated operators including “Reach,” “Grasp,” “Move,” and “Turn” from the literature and estimated their time from the Motion-Time Measurement (MTM) technique. The MTM is an analysis procedure that

analyzes any manual operation into basic motions required to perform it. Then, it assigns a predetermined time standard to each motion, determined by the influencing factors under a specific task. Among various MTM approaches, MTM-1 was selected as it has the most fine-grained level of description for human movements (Maynard et al., 1948). QN-MHP also used MTM-based equations for calculating movement time (Feyen, 2003).

"Reach" operator was defined as the time to reach an object in a fixed location. In performing ADLs, the user requires to pick up or hold an item in a fixed location. The distance between the hook and the object was measured as 12 inches in our experiment. Thus, it can be matched to "R12A" in Table 5 (i.e., R: reaching movement, 12: distance in inches, A: Reach to object in a fixed location, or to object in other hand or on which other hand rests). In Table 5, the time measurement unit (TMU) for this operator is equal to 0.036 seconds, or 36 milliseconds. Thus, the time to reach an object in HPM-UP was calculated as Equation (14).

$$\text{Time for Reach} = \text{R12A} = 9.6 \times 36 \text{ ms} = 345.6 \text{ ms} \quad (14)$$

Table 5. MTM-1 - REACH

Distance Moved Inches	Time TMU				Hand In Motion			CASE AND DESCRIPTION
	A	B	C or D	E	A	B		
3/4 or less	2.0	2.0	2.0	2.0	1.6	1.6	A	Reach to object in fixed location, or to object in other hand or on which other hand rests.
1	2.5	2.5	3.6	2.4	2.3	2.3		
2	4.0	4.0	5.9	3.8	3.5	2.7		
3	5.3	5.3	7.3	5.3	4.5	3.6	B	Reach to single object in location which may vary slightly from cycle to cycle.
4	6.1	6.4	8.4	6.8	4.9	4.3		
5	6.5	7.8	9.4	7.4	5.3	5.0		
6	7.0	8.6	10.1	8.0	5.7	5.7		
7	7.4	9.3	10.8	8.7	6.1	6.5	C	Reach to object jumbled with other objects in a group so that search and select occur.
8	7.9	10.1	11.5	9.3	6.5	7.2		
9	8.3	10.8	12.2	9.9	6.9	7.9		
10	8.7	11.5	12.9	10.5	7.3	8.6		
12	9.6	12.9	14.2	11.8	8.1	10.1		
14	10.5	14.4	15.6	13.0	8.9	11.5	D	Reach to a very small object or where accurate grasp is required.
16	11.4	15.8	17.0	14.2	9.7	12.9		
18	12.3	17.2	18.4	15.5	10.5	14.4		
20	13.1	18.6	19.8	16.7	11.3	15.8		
22	14.0	20.1	21.2	18.0	12.1	17.3		
24	14.9	21.5	22.5	19.2	12.9	18.8	E	Reach to indefinite location to get hand in position for body balance or next motion or out of way.
26	15.8	22.9	23.9	20.4	13.7	20.2		
28	16.7	24.4	25.3	21.7	14.5	21.7		
30	17.5	25.8	26.7	22.9	15.3	23.2		
Additional	0.4	0.7	0.7	0.6				TMU per inch over 30 inches

"Grasp" operator is defined as the time to pick up an item. Since the item in our experimental tasks is clothespin and doorhandle, the object's diameter was between 0.25 and 0.5 inches. Based on Table 6, this operator was matched with "1C2" and its time was calculated based on Equation (15).

$$\text{Time for Grasp} = 1C2 = 8.7 \times 36 \text{ ms} = 313.2 \text{ ms} \quad (15)$$

Table 6. MTM-1 - GRASP

TYPE OF GRASP	Case	Time TMU	DESCRIPTION	
PICK - UP	1A	2.0	Any size object by itsef, easily grasped	
	1B	3.5	Object very small or lying close against a flat surface	
	1C1	7.3	Diameter larger than 1/2"	Interference with Grasp on bottom and one side of nearly cylindrical object.
	1C2	8.7	Diameter 1/4" to 1/2 "	
	1C3	10.8	Diameter less than 1/4"	
REGRASP	2	5.6	Change grasp without relinquishing control.	
TRANSFER	3	5.6	Control transferred from one hand to the other.	
SELECT	4A	7.3	Larger than 1" x 1" x 1"	Object jumbled with other objects so that search and select occur.
	4B	9.1	1/4 " x 1/4 " x 1/8" to 1" x 1" x 1"	
	4C	12.9	Smaller than 1/4" x 1/4" x 1/8"	
CONTACT	5	0	Contact, Sliding, or Hook Grasp.	

The “Move” operator in the CRT requires moving an object to an exact location. The distance between the two locations was 20 centimeters in our experiment. Since the HPM-UP supposes users are wearing a prosthetic device, the device's weight should also be included in the equation. Since the device's weight was close to 4.54lb, a dynamic factor of 1.06 was applied (Table 7). Thus, the time to move the upper limb was calculated as Equation (16).

$$\text{Time for Move (no weight)} = \text{M20C} = 22.1 \times 36 \text{ ms} = 795.6 \text{ ms} \quad (16)$$

$$\text{Time for Move (including weight)} = 795.6 \times 1.06 = 843.3 \text{ ms}$$

Table 7. MTM-1 - MOVE

Distance Moved cm	Time TMU				Wt. Allowance			CASE AND DESCRIPTION
	A	B	C	Hand in Motion B	Wt. (lb) Up to	Dynamic Factor	Static Constant TMU	
3/4 or less	2.0	2.0	2.0	1.7	2.50	1.00	0.00	A Move object to other hand or against stop.
1	2.5	2.9	3.4	2.3				
2	3.6	4.6	5.2	2.9	7.50	1.06	2.20	
3	4.9	5.7	6.7	3.6				
4	6.1	6.9	8.0	4.3	12.50	1.11	3.90	
5	7.3	8.0	9.2	5.0				
6	8.1	8.9	10.3	5.7	17.50	1.17	5.60	B Move object to approximate or indefinite location.
7	8.9	9.7	11.1	6.5				
8	9.7	10.6	11.8	7.2	22.50	1.22	7.40	
9	10.5	11.5	12.7	7.9				
10	11.3	12.2	13.5	8.6	27.50	1.28	9.10	
12	12.9	13.4	15.2	10.0				
14	14.4	14.6	16.9	11.4	32.50	1.33	10.80	
16	16.0	15.8	18.7	12.8				
18	17.6	17.0	20.4	14.2	37.50	1.39	12.50	C Move object to exact location.
20	19.2	18.2	22.1	15.6				
22	20.8	19.4	23.8	17.0	42.50	1.44	14.30	
24	22.4	20.6	25.5	18.4				
26	24.0	21.8	27.3	19.8	47.50	1.50	16.00	
28	25.5	23.1	29.0	21.2				
30	27.1	24.3	30.7	22.7				
Additional	0.8	0.6	0.85		TMU per inch over 30 inches			

The “Turn” operator is used to model hook rotation for 90 degrees. In addition, this operator is used for pronation and supination of the hand in the PR and CC configurations. Although the object (i.e., clothespin) is small (0 to 2lbs), the task takes more time (based on our observations of video recordings) than the estimate in Table 8 because the operator is turning the

hook while sustaining the position of the prosthetic device above the desk with some system delay (120ms). Thus, the time was calculated as Equation (17).

$$\text{Time for Turn} = 5.4 \times 36 \text{ ms} + 120 \text{ ms} = 314.4 \text{ ms} \quad (17)$$

Table 8. MTM-1 - TURN

Weight	Time TMU for Degrees Turned										
	30°	45°	60°	75°	90°	105°	120°	135°	150°	165°	180°
Small - 0 to 2 lbs	2.8	3.5	4.1	4.8	5.4	6.1	6.8	7.4	8.1	8.7	9.4
Medium - 2.1 to 10 lbs	4.4	55	6.5	7.5	8.5	9.6	10.6	11.6	12.7	13.7	14.8
Large - 10.1 to 35 lbs	8.4	10.5	12.3	14.4	16.2	18.3	20.4	22.2	24.3	26.1	28.2

For modeling the DC configuration, the HPM-UP needed to include two additional hand flexion and extension operators that were not in the MTM-1 library. Thus, the time for "flexion" was added to the HPM-UP as 209.5ms (*SD* = 61.4ms) based on Sheng and Wan (2013), and the time for "extension" was added to the model as 201.4ms (*SD* = 51.9ms) (Sheng and Wan, 2013). Table 9 provides a summary of motor operators in HPM-UP.

Table 9. List of motor operators in HPM-UP

Operator	**Duration (ms)**	**Operator**	**Duration (ms)**
Reach	345.6	Grasp	313.2
Move	819.5	Turn (supination or pronation)	314.4
Flexion	209.5± 61.4	Extension	201.4 ± 51.9

3.2.5. Satisfaction

3.2.5.1. Expectation confirmation theory

The theoretical background to formulate *satisfaction* came from the expectation confirmation theory (ECT) which is a cognitive theory that explains satisfaction as a function of *expectations* and *perceived performance* (Oliver, 1977, 1980). Although the theory originally

appeared in psychology and marketing studies, it has been applied in other scientific fields, including consumer research and information systems (Bhattacherjee, 2001).

Once users accumulate experience on a product or service, they can subjectively evaluate their performance with the device (Lowry et al., 2006). The user compares the *desire* and *expectations* against the perceived performance of the product. *Expectation* is a belief or subjective prediction about a product's attributes or performance at some point in the future (Bhattacherjee, 2001). *Perceived performance* is a user's perception of the degree to which a product can fulfill his or her expectation in actual usage. *Desire* is the level of attributes and benefits that leads to attaining the user's desired outcomes (Spreng et al., 1996). The relationship between these three concepts, disconfirmation of beliefs, and satisfaction is shown in Figure 7. ECT posits that satisfaction is directly influenced by disconfirmation of beliefs and perceived performance and is indirectly influenced by both expectations and perceived performance by means of a mediational relationship which passes through the disconfirmation construct.

When a product outperforms user's original expectations, the disconfirmation becomes positive, which leads to increase post-purchase or post-adoption satisfaction. However, when a product underperforms the user's original expectations, the disconfirmation becomes negative, which decreases post-purchase or post-adoption satisfaction (or increase dissatisfaction). Therefore, *satisfaction* can be determined by the amount of difference between *perceived performance* and *expectation*. In addition, there is an inverse relationship between *expectation* and *satisfaction*. If *expectation* increases, the possibility to reduce satisfaction increases. Reversely, if *expectation* decreases, the possibility to reduce satisfaction decreases.

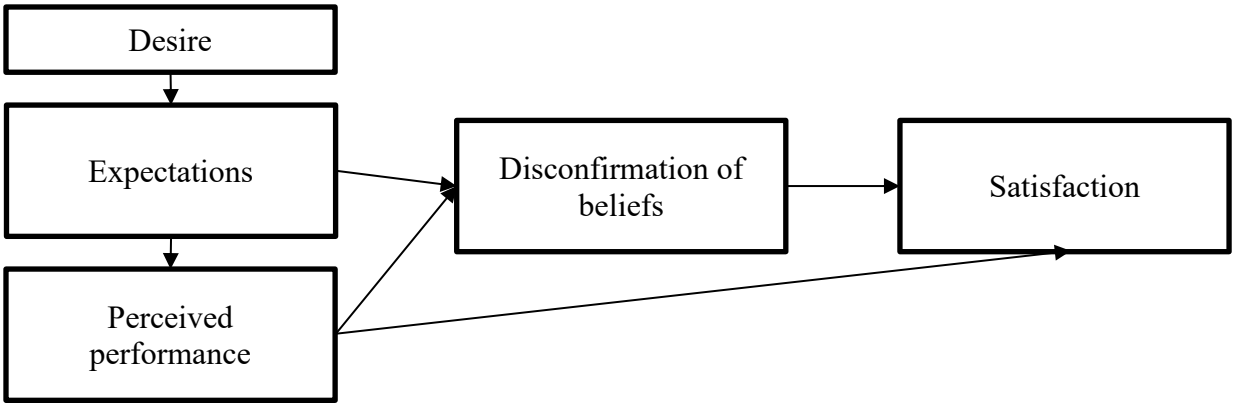


Figure 7. Expectation-confirmation theory

3.2.5.2. Satisfaction in HPM-UP

Expectation (f) in the HPM-UP was defined based on the expected task performance after passing the training session (or expected performance before the experiment trials). *Desire* was determined with q that was used in the calculation of *learnability* dimension. *Perceived performance* was calculated from the *Efficiency* dimension. Based on these concepts, disconfirmation of beliefs was formulated as Equation (18).

$$disconfirmation\ of\ belief_i \quad (18)$$

$$= Z\left\{\overrightarrow{P_i} - f\left(\overrightarrow{L_i}, \overrightarrow{q_i}\right)\right\}$$

$$= Z\left\{Percevied\ performance_i - \frac{Entire\ task\ duration\ (120\ seconds)}{\min threshold_i + Learnability_i \times (\max threshold_i - \min threshold_i)} \times q_i\right\}$$

In this equation, $\overrightarrow{P_i}$ is a matrix or vector of participants' perceived performance, which is calculated from the efficiency module of HPM-UP. *Expectation* (f) is a function of learnability ($\overrightarrow{L_i}$) and desire ($\overrightarrow{q_i}$) because *expectation* can be estimated using the training performance or *learnability* as it is a belief or subjective prediction of performance in the future. For example, if users pass the training criteria only with 3 or 4 trials, they may perform well during the experimental trials. Thus, *learnability* was used as a variable to determine *expectation* (i.e.,

expected performance) before the experimental trials based on the thresholds defined for each configuration (i.e., DC:20-35s, PR: 15-25s, and CC: 16-23s). Based on the definition of desire (i.e., the level of attributes and benefits that leads to attaining the user's desired outcomes), *q* ranges between 0 and 1 and is multiplied by *expectation*.

Effort, which is one of the dimensions in NASA-TLX and is defined as the level of difficulty (mentally and physically) in performing an activity (Hart and Staveland, 1988), could also affect perceived performance. Therefore, the ECT was revised to include the level of effort needed to perform the tasks (Figure 8).

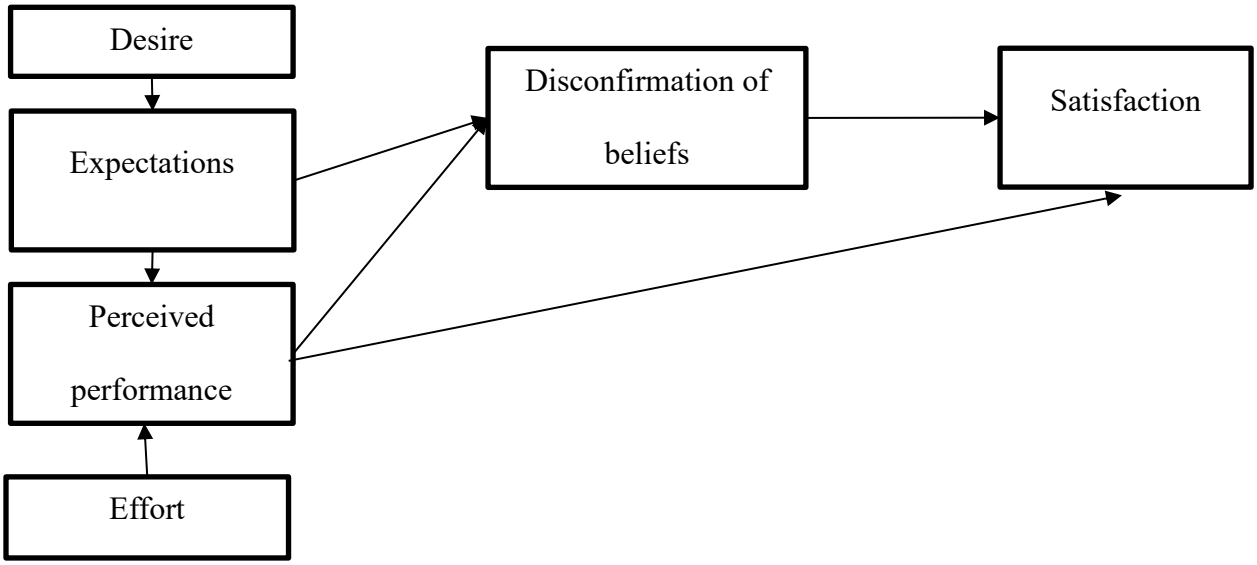


Figure 8. Revised expectation-confirmation theory for this study

Satisfaction was formulated with Equation 19 based on the disconfirmation of beliefs and effort. In addition, a constant value (c) was added which refers to the minimum level of satisfaction.

$$Satisfaction_i = (Disconfirmation\ of\ belief_i) \times (1 - \frac{Effort}{100}) + c \quad (19)$$

3.2.6. Cognitive Workload Classification

Previous software packages and models included cognitive workload as one of the outputs of the model. For example, Cogulator uses S-shaped curve fitting (Estes, 2015) based on memory chunks and QN-MHP uses a computational approach based on the dimensions of NASA-TLX (Jeong and Liu, 2018). HPM-UP can predict cognitive workload using machine learning (ML) algorithms. Variables such as pupil size, task performance, number of cognitive/perceptual/motor operators, and number of memory chunks can be used as inputs for the ML algorithm. The algorithm classifies CW of using prosthetic devices in different classes (e.g., “High” or “Low”)

To classify CW, ML algorithms can be used with several advantages compared to inferential statistics (Park et al., 2022). First, ML algorithms can be used to find relationships among features in high dimensional spaces and deal with non-linear factors and uncertainty without strict assumptions in inferential statistics (Moustafa et al., 2017). Second, the method allows for classification of CW in near real-time (Braarud et al., 2021). With these advantages, several ML algorithms have been used to classify CW of operators in various domains such as construction or aviation.

The most frequently used ML methods for classifying CW were support vector classifier (SVC) (Meyer, 2017), random forest (RF) (Liaw and Wiener, 2002), and Naïve Bayes (NB) algorithms (Majka, 2018). A majority of studies used physiological measurements (e.g., heart rate) as input features to classify CW (Meteier et al., 2021; Walambe et al., 2021) and some used task performance outcomes (e.g., response time to secondary task) (Ding et al., 2020; Li et al., 2020). However, prior studies had several limitations. First, there has not been any investigation on classification of CW for prosthetic devices, although high CW is one of the major challenges with existing prosthetic devices. Second, although several measures such as physiological responses,

task performance, and subjective responses have been used as input features in CW classification algorithms, no study used CPM generated outcomes as input features to classify CW. CPM models and their outcomes can be generated by observation of different tasks and using knowledge elicitation approaches with small sample size and do not require extensive human subject experiments, and therefore can be used in early stages of the design cycle (Park and Zahabi, 2022a). Third, there were limited number studies exploring the effect of a subset of features on the ML outcomes. Some studies tested subsets of features, however, they are limited to only physiological (Ding et al., 2020) or task performance data (Braarud et al., 2021). Therefore, this research aimed to investigate multimodal input features to classify CW in using EMG-based prosthetic devices.

3.2.6.1. Data labeling

Participants' NASA-TLX scores and weights for each dimension were collected based on the procedure described in Hart and Steveland (1988). The weights were captured before the first trial of the experiment by asking the participants to complete the pairwise comparison rating form. After each trial, participants completed the workload ratings for each dimension based on what they experienced during that trial. Using these weights, the weighted average was calculated for each trial to have a single and overall score of NASA-TLX and then the overall scores were clustered into different classes. Since this target variable (i.e., the overall NASA-TLX score (0-100%)) was a continuous variable, there was a need to group the data into different categories before classification.

A clustering analysis was conducted on all participants' NASA-TLX scores to find the optimal number of classes of CW using the NbClust package in R. There are several clustering analysis approaches, and each algorithm generates different results based on specific indices or

methods (e.g., kmeans). We tested all the combinations of clustering methods and indices and found that the most frequent optimal number of classes determined from different methods were two, four, and three clusters respectively. Although we could simply select the most frequent optimal number of classes (which was having two classes of workload), we decided to include the top three selected classes as having more detailed classification (e.g., low, medium, high workload) would provide more precise estimate of workload. However, due to the lack of sufficient number of data points in some of these classes, only two or three classes of CW were used in the analysis.

3.2.6.2. Algorithm Selection

Three algorithms of Random Forest (RF), Support Vector Classifier (SVC), and Naïve Bayes (NB) were selected to classify CW since (1) they were used extensively in recent studies (Braarud et al., 2021; Kaczorowska et al., 2021; Meteier et al., 2021; Shao et al., 2021; Sharma et al., 2021; Walambe et al., 2021), (2) included physiological data (e.g., pupillometry) and task performance (e.g., response time on secondary task) measures as their input features, and (3) exhibited high prediction accuracy (> 80%) in small datasets (Kaczorowska et al., 2021).

3.2.6.3. Optimization and Validation

Given the small dataset (i.e., 90 datapoints for each task = 10 participants per each control scheme $\times$ 3 control schemes $\times$ 3 trials), overfitting was the major concern for establishing the ML structure. Therefore, we first split our dataset into training (70% of the data) and testing (30% of the data) groups. We randomly partitioned the data from 30 participants into the training and testing datasets (i.e., the data points of one participant only appeared either in training or testing dataset). Then, 10-fold CV was employed to optimize the hyperparameters (Götze et al., 2020b). A hyperparameter grid search method was conducted using the sklearn Python library (Pedregosa et al., 2011) and a Pipeline function to streamline testing across three different model types (i.e.,

RF, SVC, and NB). RF has a wide range of applications and can have good performance even with the default hyperparameters (Donges, 2021). Among the different hyperparameters of RF, the three most notable and influential parameters are the number of trees in the model forest, the maximum tolerable depth of each tree, and the number of features necessary at each branching point (Probst, 2019). Limiting the number and depth of trees reduces overfitting of the data; otherwise, though a model may be ideal for the training data if allowed to infinitely grow, out-of-sample performance would be extremely poor. Considering the number of data points at each branching point in the tree is another means of limiting the shunting of model performance towards narrow-minded behavior. In preliminary testing, however, the number of features necessary at each breaking point continuously output its default value of 2, and thus it was not considered in the final grid search.

SVC employs a spatial approach to delineating class margins and has a reputation for being computationally expedient in rudimentary modeling. Many studies with similar dataset challenges have employed SVC to classify data efficiently (Braarud et al., 2021; Raihan-Al-Masud and Mondal, 2020). In these situations, a linear kernel type was used, specifying which subtype of SVC to employ (Meteier et al., 2021). In doing so, the chief remaining hyperparameter was the regularization variable ('c' in Table 10). This parameter calculates the amount of tolerable error the algorithm considers before passing a model as output. Like the tree count for random forest, a regularization constant that is too small could massively overfit the data.

For NB, given our small and unbalanced dataset, a complement NB model was implemented (Rennie et al., 2003). Hyperparameter grid searching was performed only for the "alpha" parameter (Table 10) as it determines the portion of the largest variance of all features that is added to variances for calculating stability (Jain, 2021; Rennie et al., 2003). Controlling the

degree of smoothness permitted by the model in delineating different classes allowed a balance to be obtained between cross validated performances in the grid search k-folding.

Table 10. Classifiers and hyperparameters

Classifier	Hyperparameter	Definition	Range	References
RF	n_estimators	Number of trees in the forest	[start: 100, end: 1000, step size: 100]	Götze et al. (2020a); Götze et al. (2020b)
	max_depth	Maximum number of layers of decisions tolerated	[1, 13, 1]	Mullainathan and Spiess (2017); Nadi and Moradi (2019)
	min_samples_split	Number of samples necessary to be present in the creation of a branching point in the tree (default: 2)	Fixed as default value 2	Götze et al. (2020a); Götze et al. (2020b)
SVC	c	Regularization parameter - i.e., how much error tolerable in producing model	[0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	Raihan-Al-Masud and Mondal (2020)
	kernel	Specifies which kernel to use in the program	Fixed as linear	Braarud et al. (2021); Meteier et al. (2021)
Naive-Bayes	alpha	Additive (Laplace/Lidstone) smoothing parameter	20 points from [1, 10] spaced evenly in log-space	Jain (2021); Rennie et al. (2003)

3.2.6.4. **Feature Selection**

To make modeling more efficient, feature selection methods were used to eliminate less-contributory features from the training data set. Each of the selection methods attempted to increase testing performance. The K-Best method of selection was employed as the representative method of the univariate filter class of selectors (Aggarwal, 2018). For more multivariate methods, the recursive feature selection (RFE) and forward feature selection methods were employed (Ferreira and Figueiredo, 2012; Raihan-Al-Masud and Mondal, 2020). RFE considers multivariate feature contribution as a whole and iteratively eliminates the least contributory features until the desired count is obtained (Guyon et al., 2002). Sequential forward selection (SFS) adds features

by order of significance until the number of features is obtained. RFE and SFS have demonstrated a decent performance in improving model accuracy and efficiency in prior studies (Ferreira and Figueiredo, 2012). Each of the three algorithms was employed for each model type and was executed and tested for specified feature counts 1 to 13 (i.e., the total number of features in the data set).

3.2.6.5. Model Evaluation

The test dataset was used for model evaluation. Classification accuracy, area under the receiver operating characteristic curve (AUC), precision, recall and F1-score were also calculated as measures of model performance (Ding et al., 2020; Skaramagkas et al., 2021). Accuracy is the ratio of correctly classified samples. F1-score is the harmonic mean of recall (i.e., probability of detecting each class) and precision (i.e., reliability of results in each class). The final F1-score was obtained by calculating recall and precision separately for each class and averaging them, weighted by the number of samples in each class. We used F1, recall, and precision because they are useful metrics for both balanced and imbalanced dataset, while accuracy is usually a good metric for a balanced dataset (Jeni et al., 2013). In addition, computation time for grid search was calculated (Intel® Core i7-8700 @ 3.20Ghz). We calculated grid search time because grid search was the most demanding and the dataset was extremely small. To improve the reliability and generalizability of ML results, we ran each of the models with 15 random seeds per suggestion from Colas et al. (2019) and calculated the average prediction performance.

3.3. HPM-UP in Action

3.3.1. Overview

An overview of HPM-UP graphical user interface (GUI) is illustrated in Figure 9.

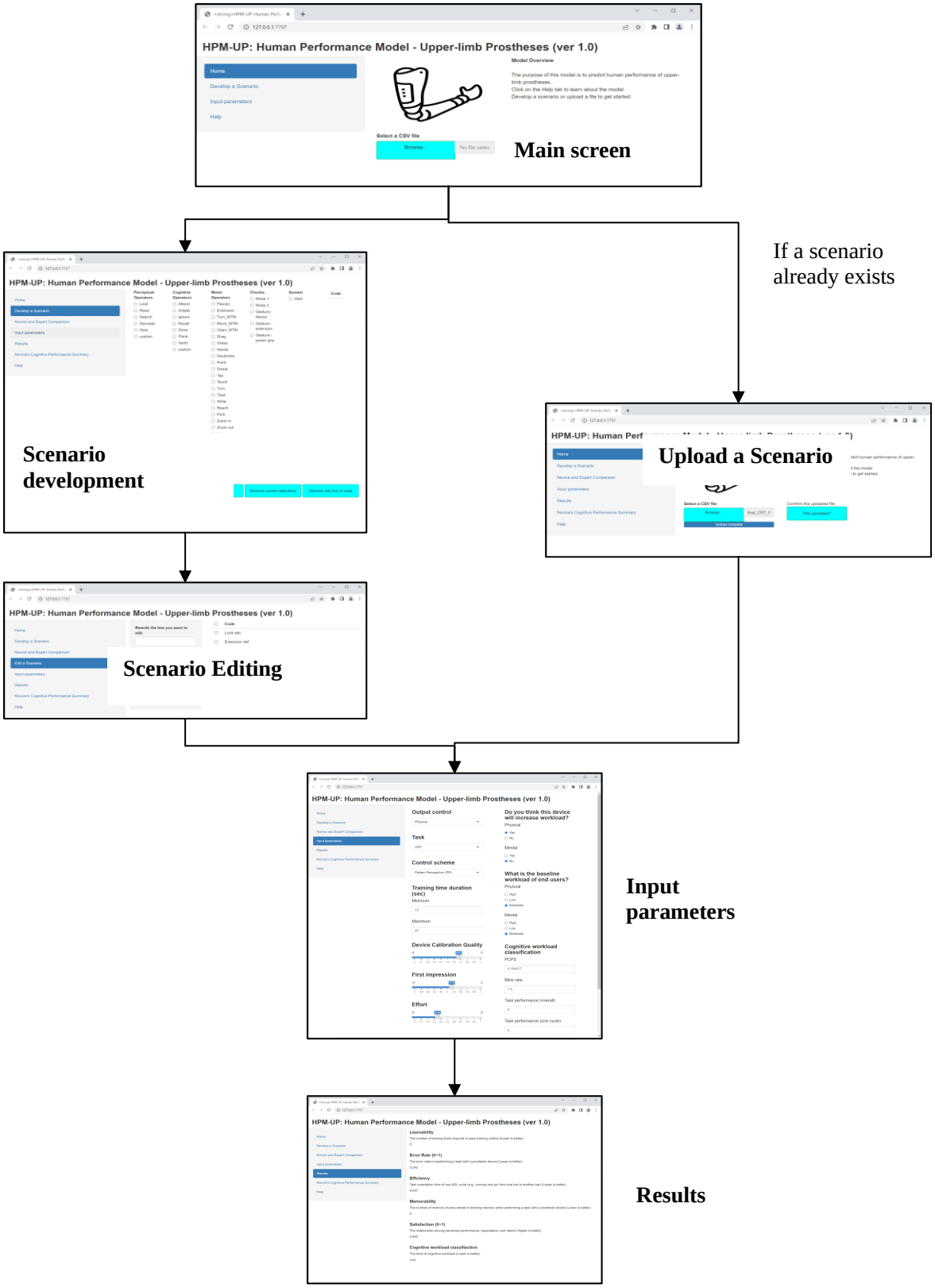


Figure 9. An example of HPM-UP GUI

Analysts can start using HPM-UP by either using the “Develop a Scenario” tab or loading an already developed scenario (Microsoft CSV format) (Figure 9). Then, input parameters should be determined from the “Input parameters” tab. Lastly, the model will assess the usability of the prosthetic device based on dimensions including learnability, error rate, memory load, efficiency, satisfaction, and classified workload as shown in the “Results” tab.

3.3.2. Scenario Development

3.3.2.1. Using the “Develop a Scenario” Tab

If the analysts would like to develop a scenario manually, they can click the “Develop a Scenario” tab (Figure 10).

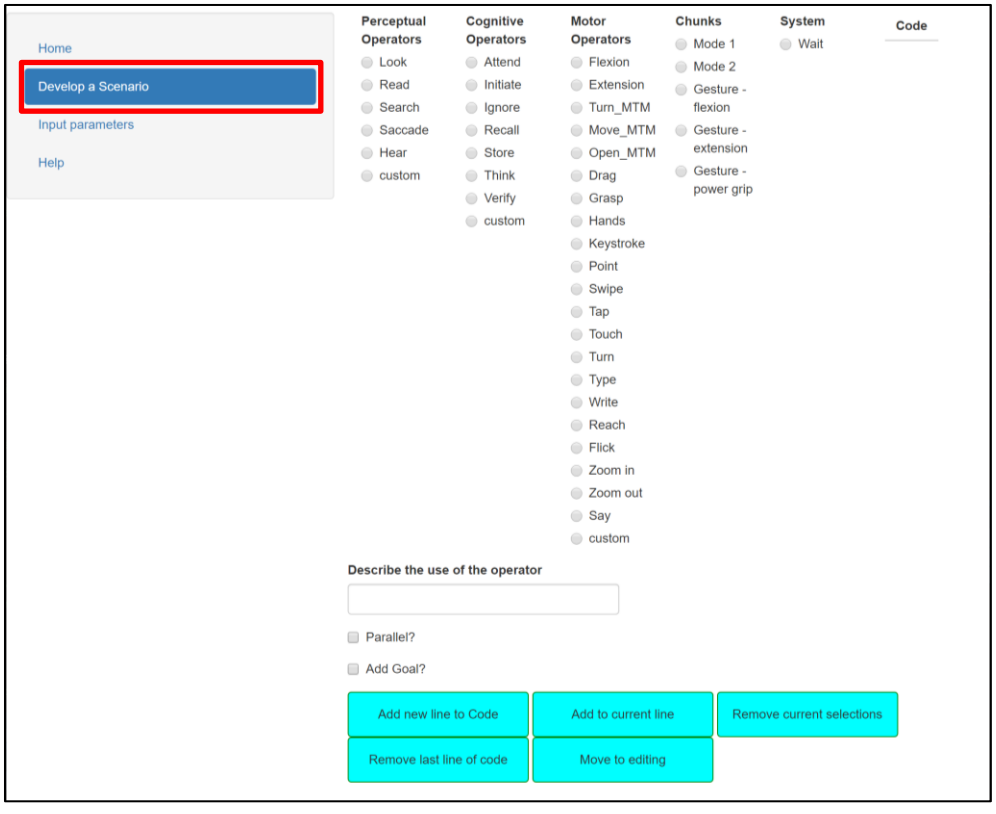


Figure 10. “Develop a Scenario” tab

Analysts can choose one of the appropriate operators or memory chunks from the radio buttons on the screen. If they would like to unselect radio buttons, they can click on the "Remove current selections" button. First, analysts should define a goal for their model (by clicking the "Add Goal?" check box). Once the goal is described in the text box "Describe the use of the operator," the analysts can click "Add new line to Code" to add a line of code to the scenario (Figure 11).

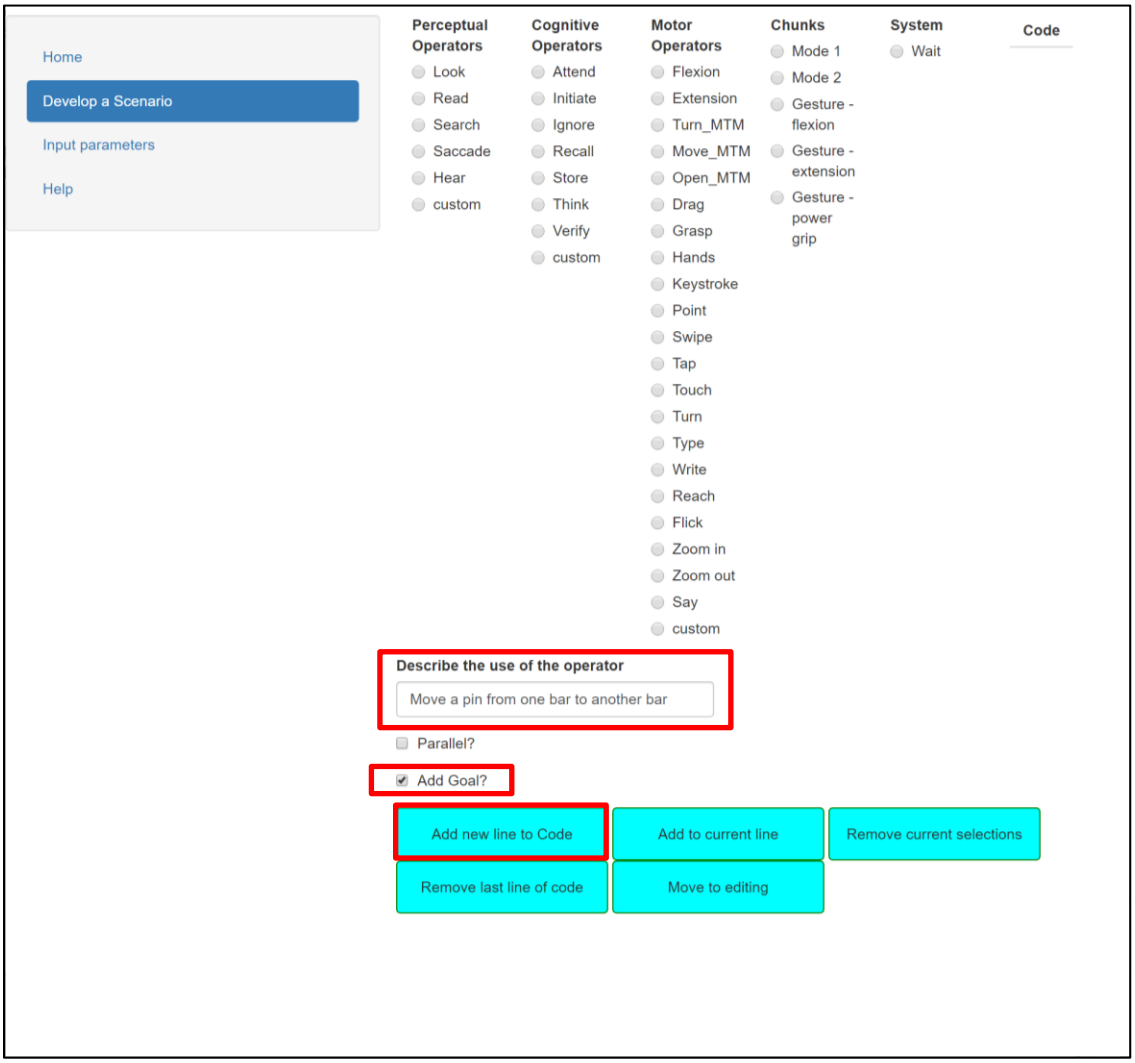


Figure 11. "Develop a Scenario" tab – Define a goal

The analyst can continuously develop the scenario by clicking one of the operators and describing the operator in the related textbox (Figure 12). Whenever analysts click "Add line new to Code," the screen shows the added lines under the "Code" column.

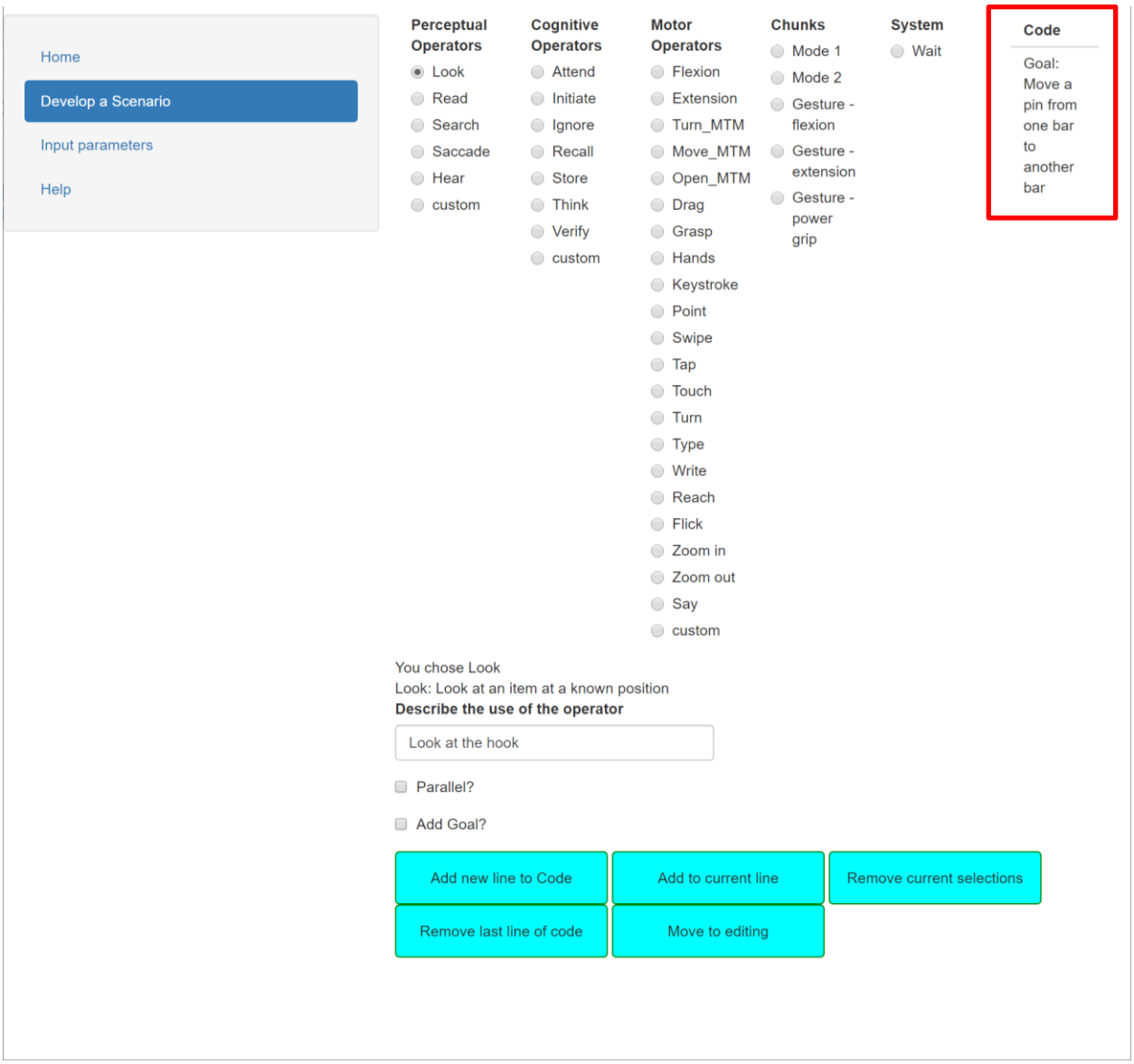


Figure 12. "Develop a Scenario" tab – Add an operator to the scenario

Parallel activities can be added by clicking the "Parallel?" check box when adding a line to the scenario (Figure 13).

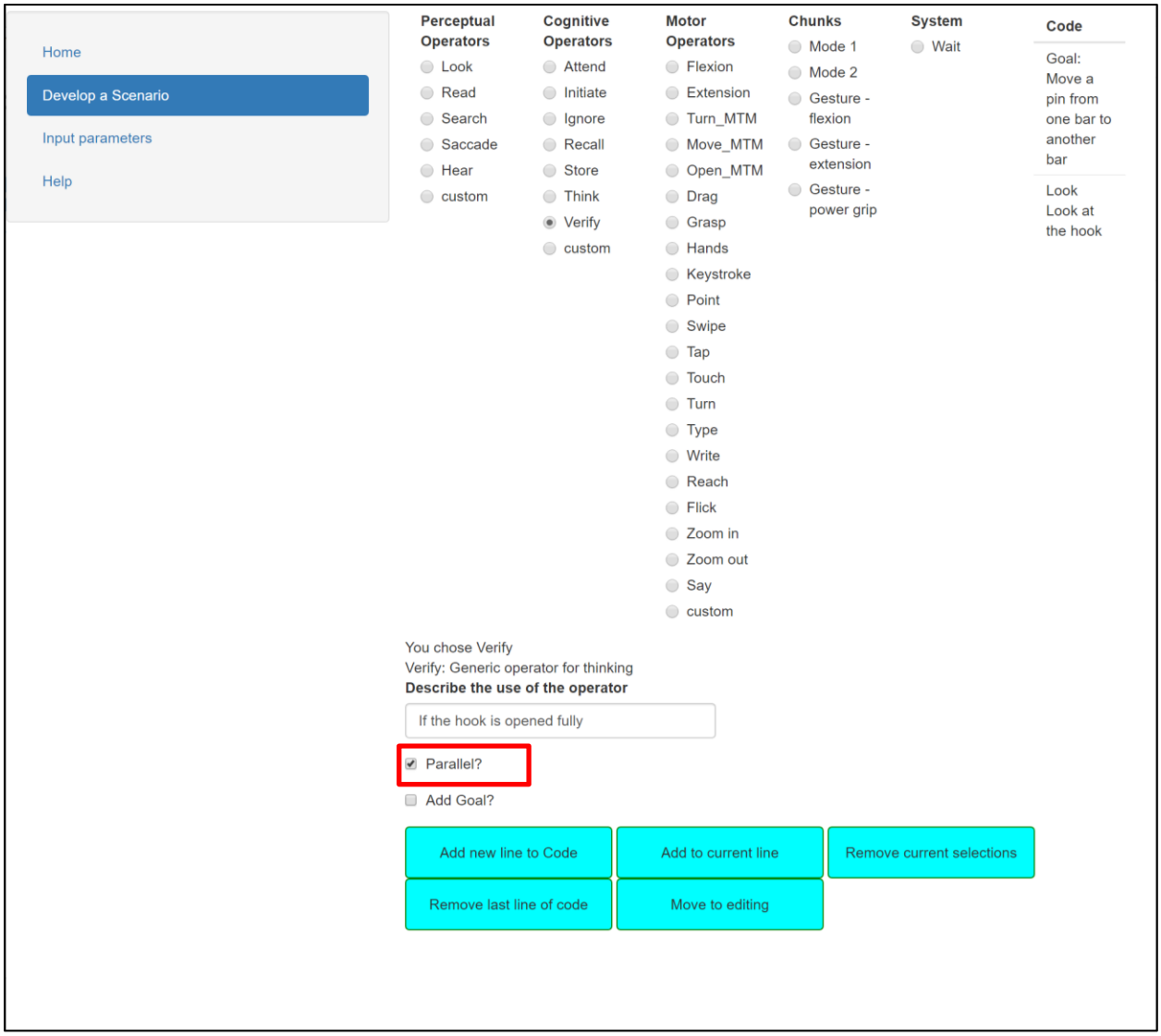


Figure 13. "Develop a Scenario" tab – Defining a parallel activity

The parallel operators will be added to the code with a line starting with "Also:" (Figure 14).

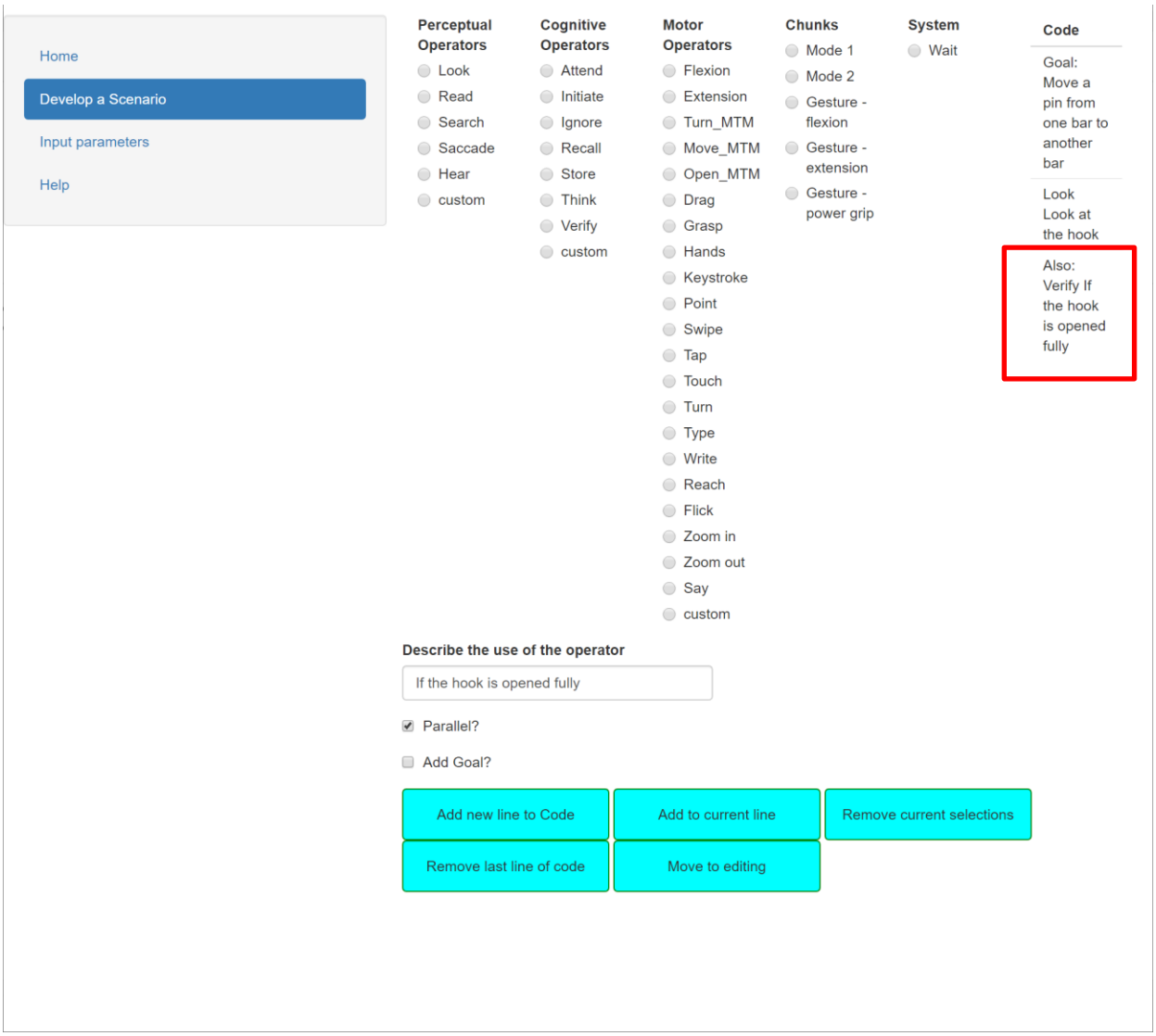


Figure 14. "Develop a Scenario" tab – Added parallel operator

A chunk can be added to the scenario once analysts choose an operator, click one of the chunks under the "Chunks" column, describe the operator, and click "Add new line to Code" (Figure 15).

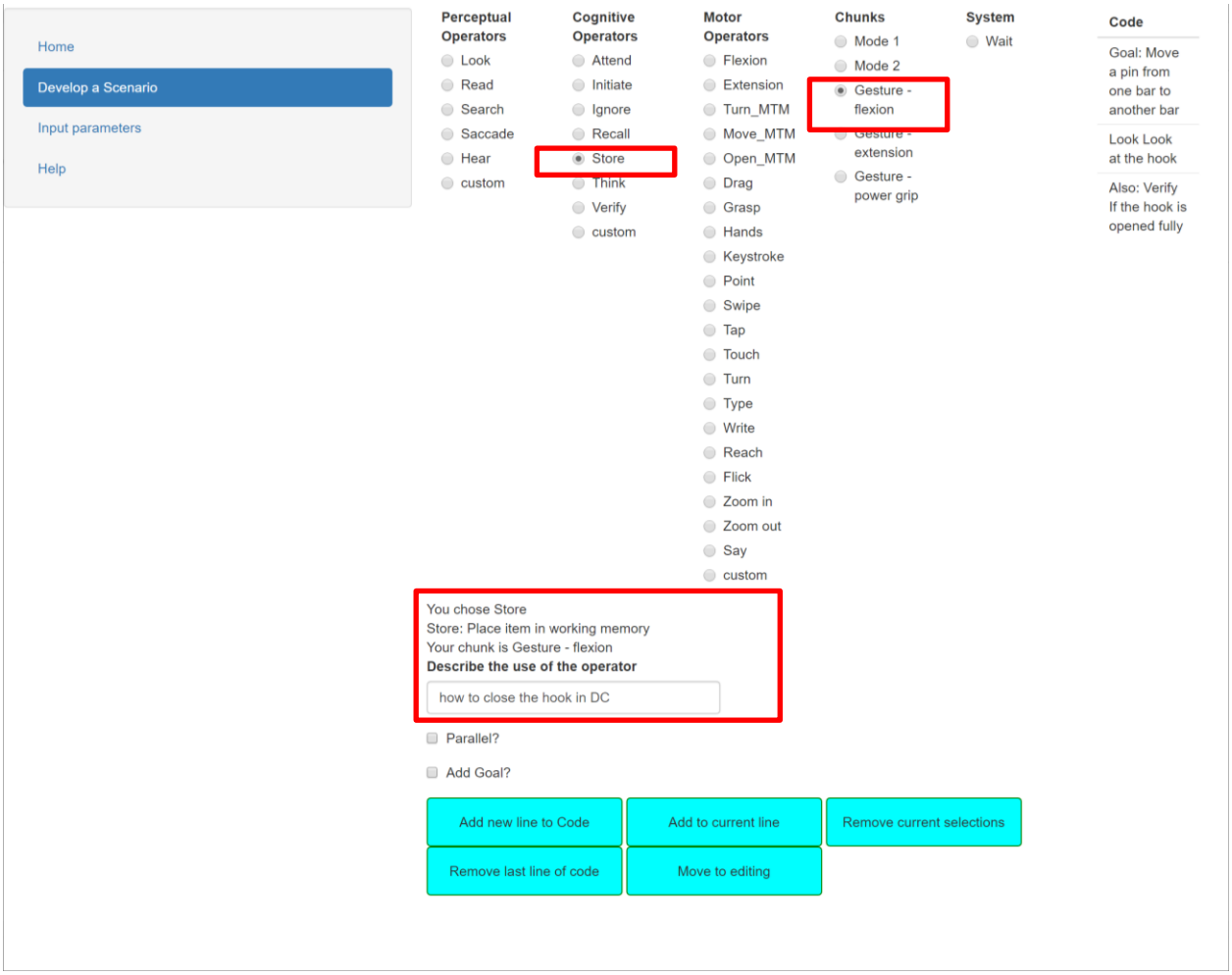


Figure 15. "Develop a Scenario" tab – Adding a chunk to the code

The added chunk will be displayed with a bracket (< and >) in the code (Figure 16).

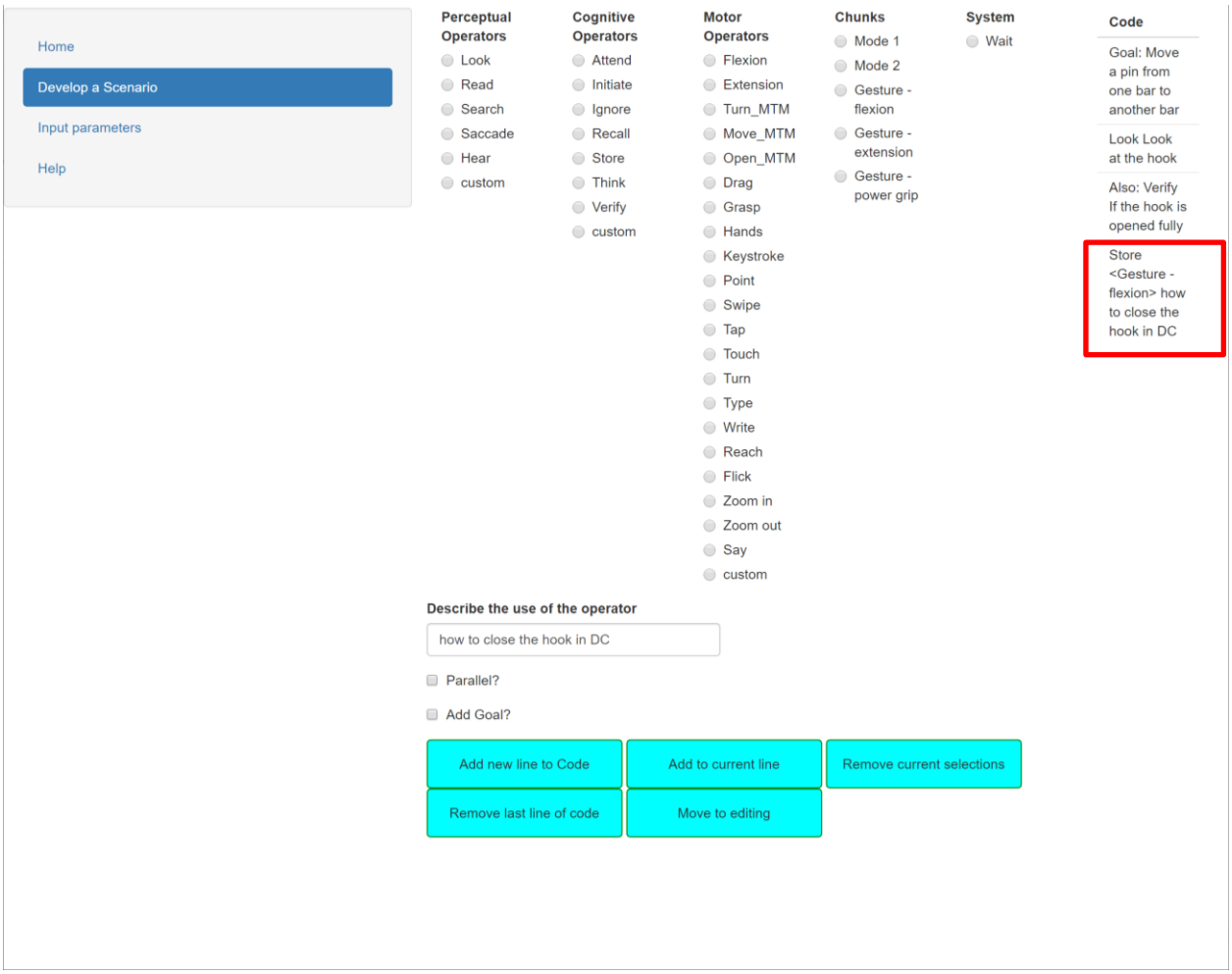


Figure 16. "Develop a Scenario" tab – Added a chunk to a line

Analysts can also add a custom chunk to the scenario. Without clicking a specific chunk under the Chunks column, a new chunk can be added directly when describing an operator (Figure 17). The added custom chunk will also be shown with brackets in the code column (Figure 18).

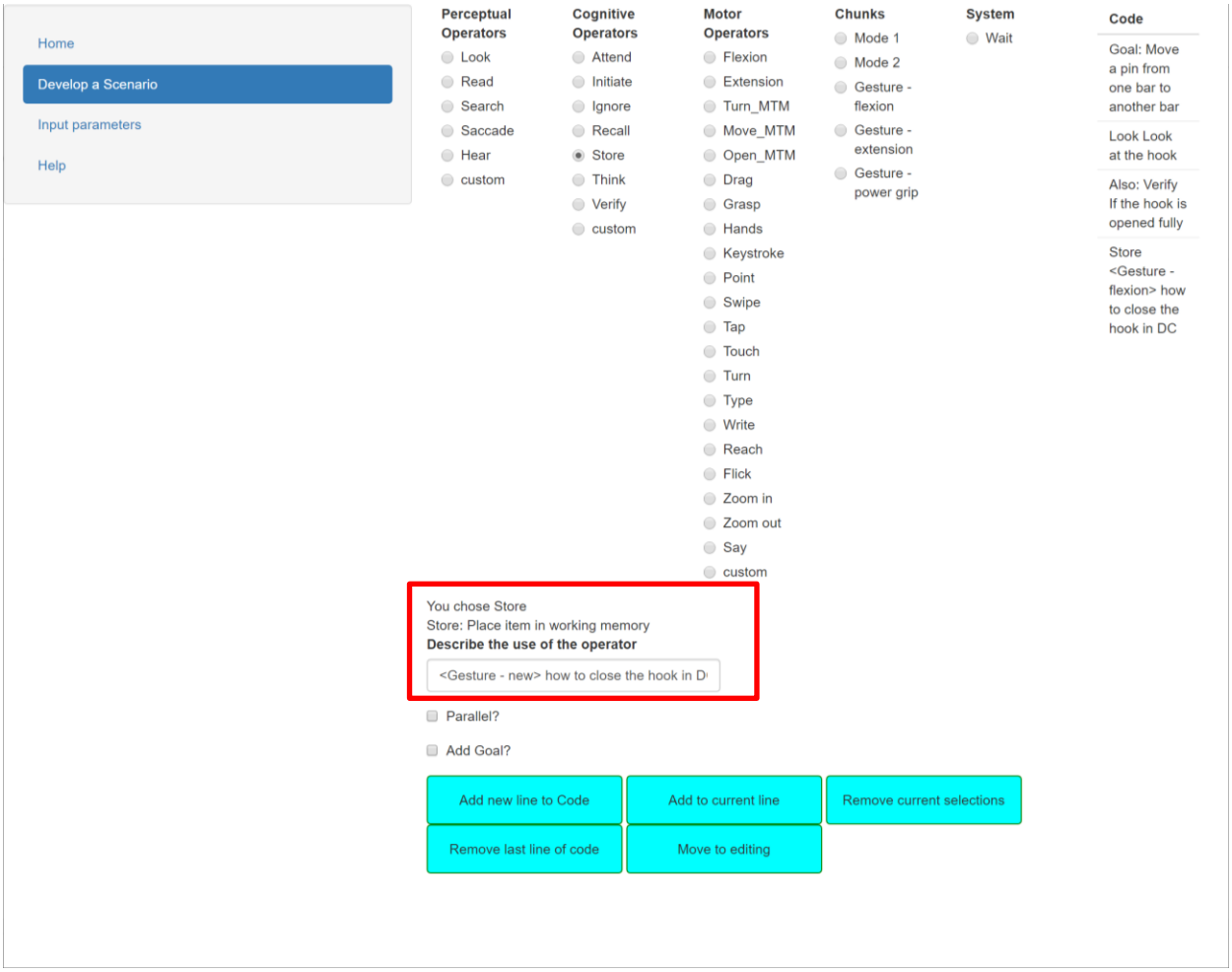


Figure 17. "Develop a Scenario" tab – Adding a custom chunk

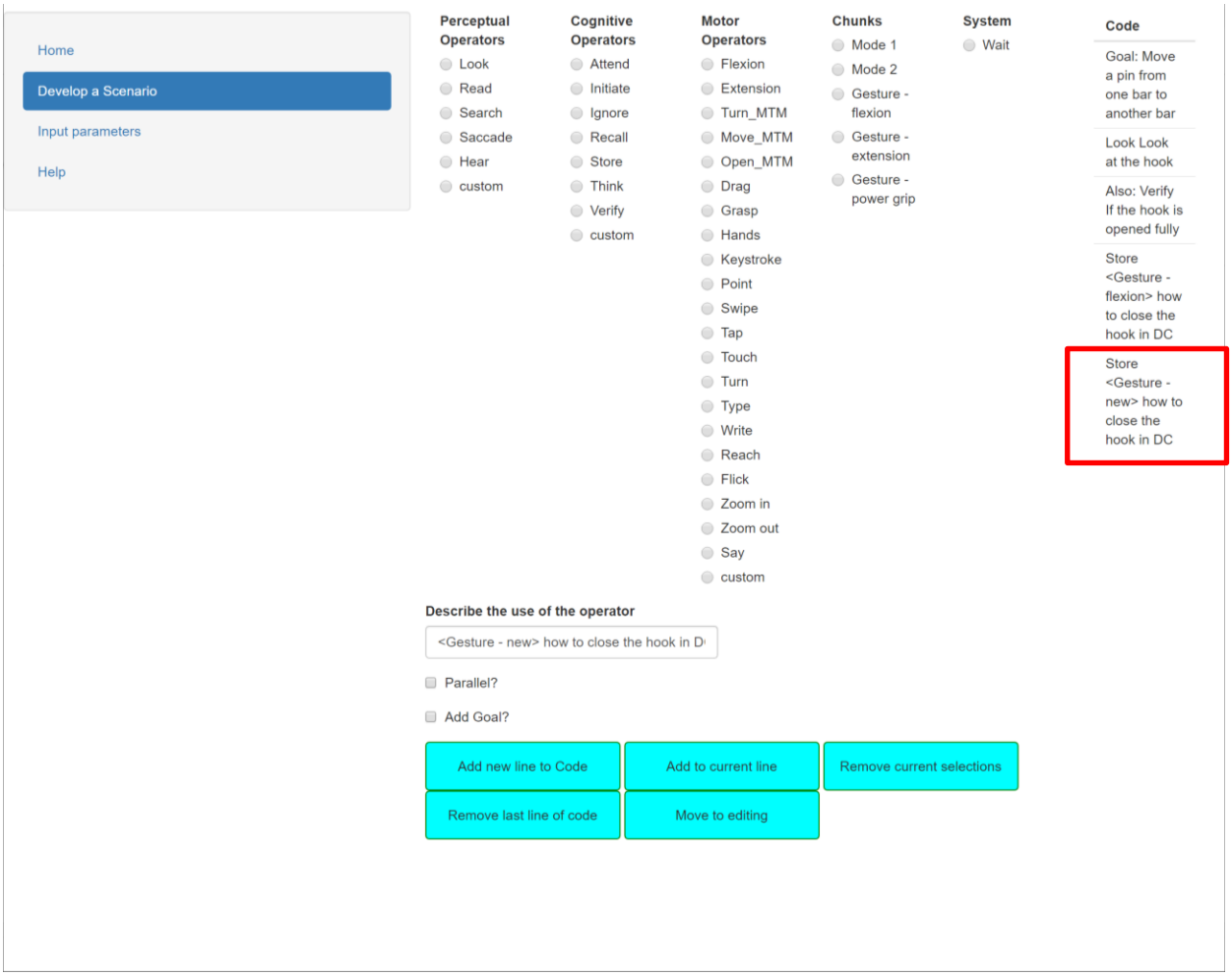


Figure 18. "Develop a Scenario" tab – Added custom chunk

If there is a need to add new or customized operators, the analyst can choose "custom" at the bottom of each column of perceptual, cognitive, or motor operators (Figure 19). Then, the name of operator and duration can be specified. Once this information is added, the operator will be added to the scenario if analyst clicks the "Confirm Custom Operator" button.

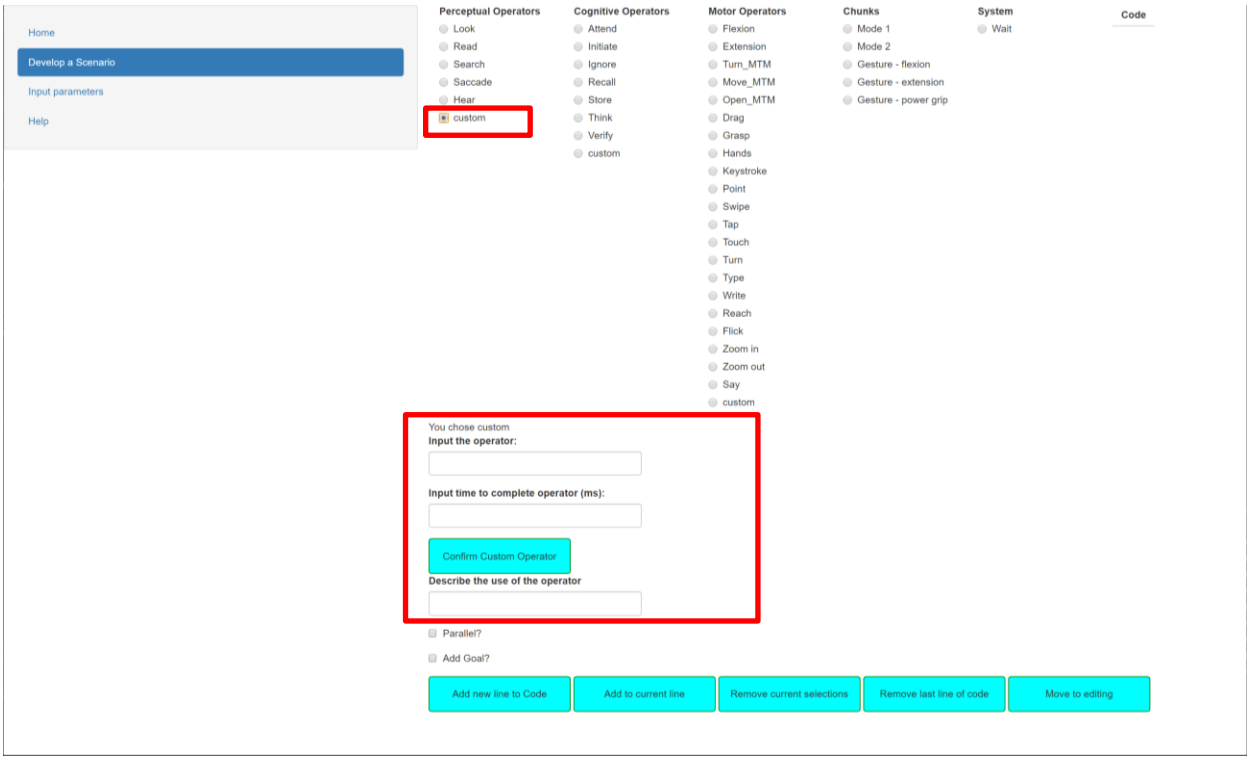


Figure 19. "Develop a Scenario" tab – Custom operator

Once the analyst completes the draft model, they can click "Move to editing" to finalize the scenario development. Then, the "Edit a Scenario" tab will be shown on the screen (Figure 20) where they can add/delete a specific line of code.

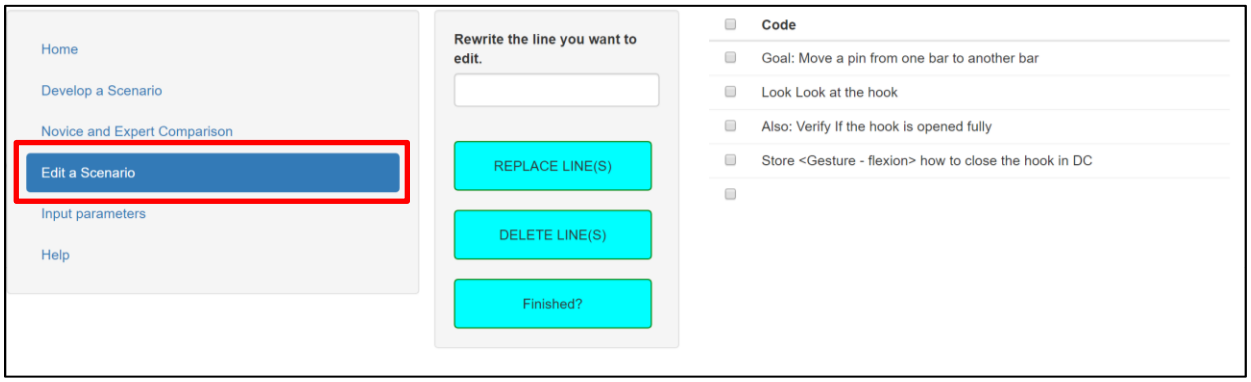


Figure 20. "Edit a Scenario" tab

The scenario can also be downloaded in a CSV format (Figure 21). This is useful because working or editing directly on a CSV format file might be necessary when the analysts are developing more complex scenarios. The CSV format file can also be loaded from the HPM-UP main screen.

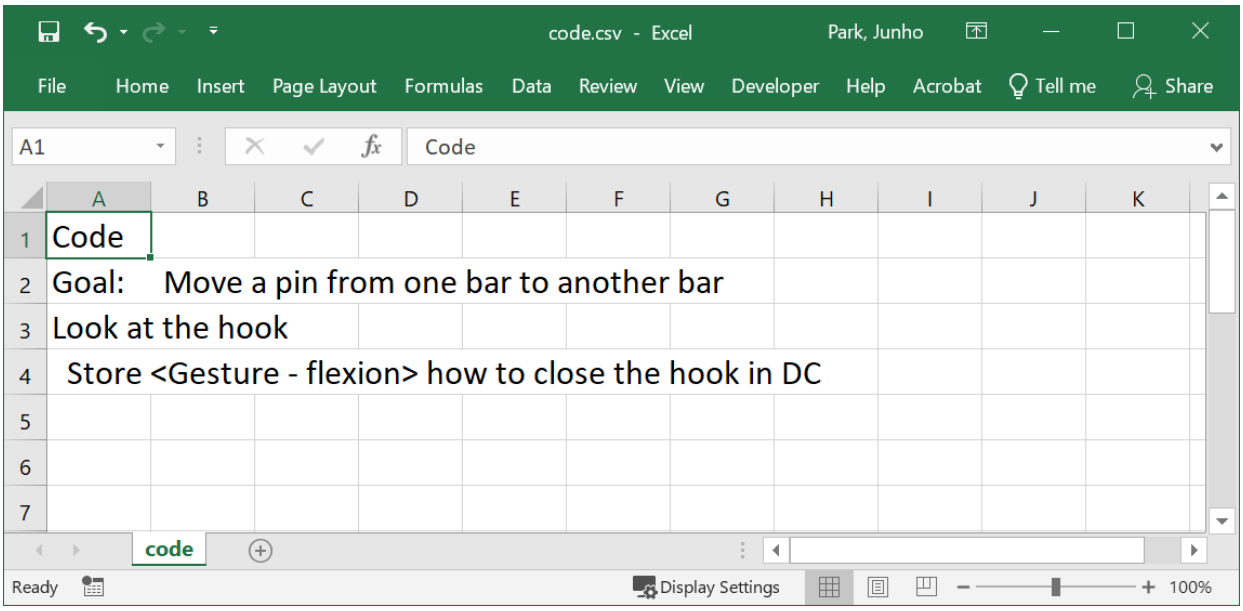


Figure 21. Downloaded scenario

3.3.2.2. Loading an existing scenario

HPM-UP can also be run with a developed scenario in a CSV format or downloaded from Cogulator. In creating these scenarios, the analyst should follow the grammar in the Cogulator software (Figure 22).

```
Goal: Move a pin from one bar(A) to another bar(B)
.Attend to task
.Initiate the task
.Also: Look at the hook
.Goal: Reach to a pin located at bar A with horizontal hook status
..Goal: Adjust the hook
...Reach to the pin
…Open_MTM
..Verify if the hook direction is appropriate
.Goal: Pick up the pin
..Grasp the pin using the hook
..if the pin is fully clamped by the hook
.Goal: Move the arm
..Move_MTM from A to B
..Also: Turn_MTM hand to make the hook perpendicular(or parallel) to the table
.Goal: Drop off the pin to B
..Look at the pin
..Verify if the pin is located on the bar
..Open_MTM the hook to drop off the pin
..Verify if the pin moved successfully
..Move_MTM the arm to_the_original_position
```

Figure 22. A sample HPM-UP scenario developed in Cogulator for moving a clothes pin from a horizontal bar to a vertical bar using the PR configuration

Once the analyst completes the scenario in a CSV format, that scenario can be loaded from the main screen of HPM-UP by clicking the "Browse" button, choosing the scenario file, and clicking the "File uploaded?" button (Figure 23).

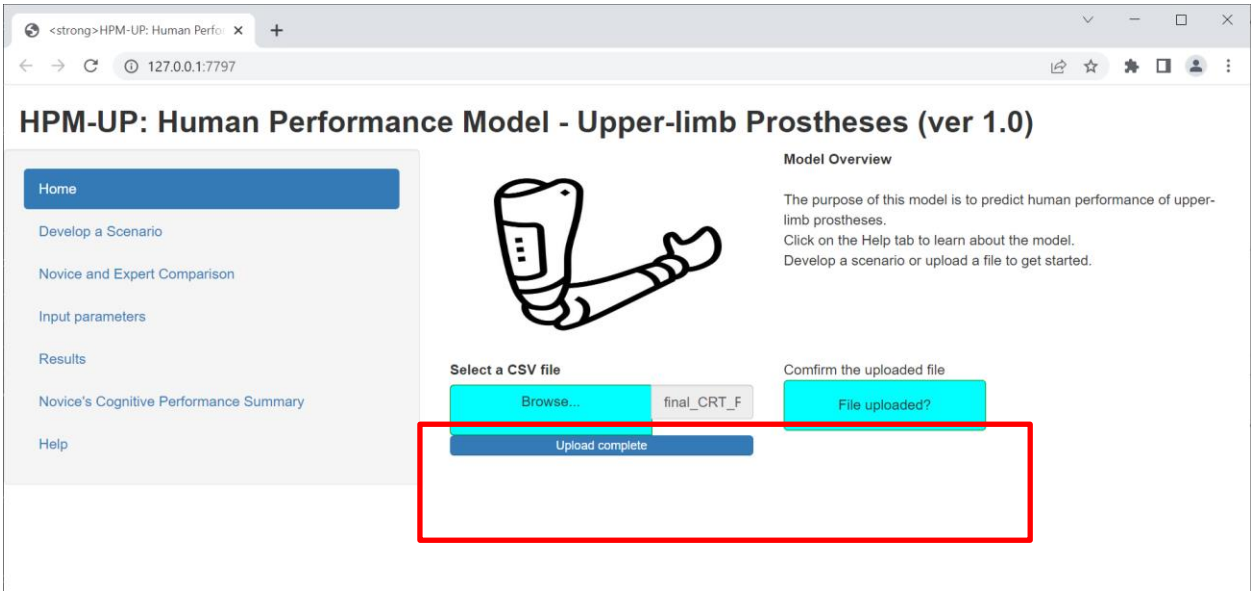


Figure 23. Loading an existing scenario

3.3.3. Input Parameters

To calculate the usability dimensions, user input is required (Figure 24). First, the output control mode (physical or virtual prosthetic device), tasks (e.g., CRT or SHAP), and control scheme (DC, PR, or CC) should be selected. The minimum and maximum training time duration should be specified based on analysts' previous knowledge or pilot test results. Device calibration quality (0-1), first impression (a number between -1 and 1), and effort (0-1) should be determined based on the end users' interaction with the device.

Two questions are related to end users' perception of physical and mental workload when using the device (e.g., after conducting some pilot tests with the prosthetic device). They can also answer these questions based on the provided information, photos, or videos of the prostheses without the actual usage.

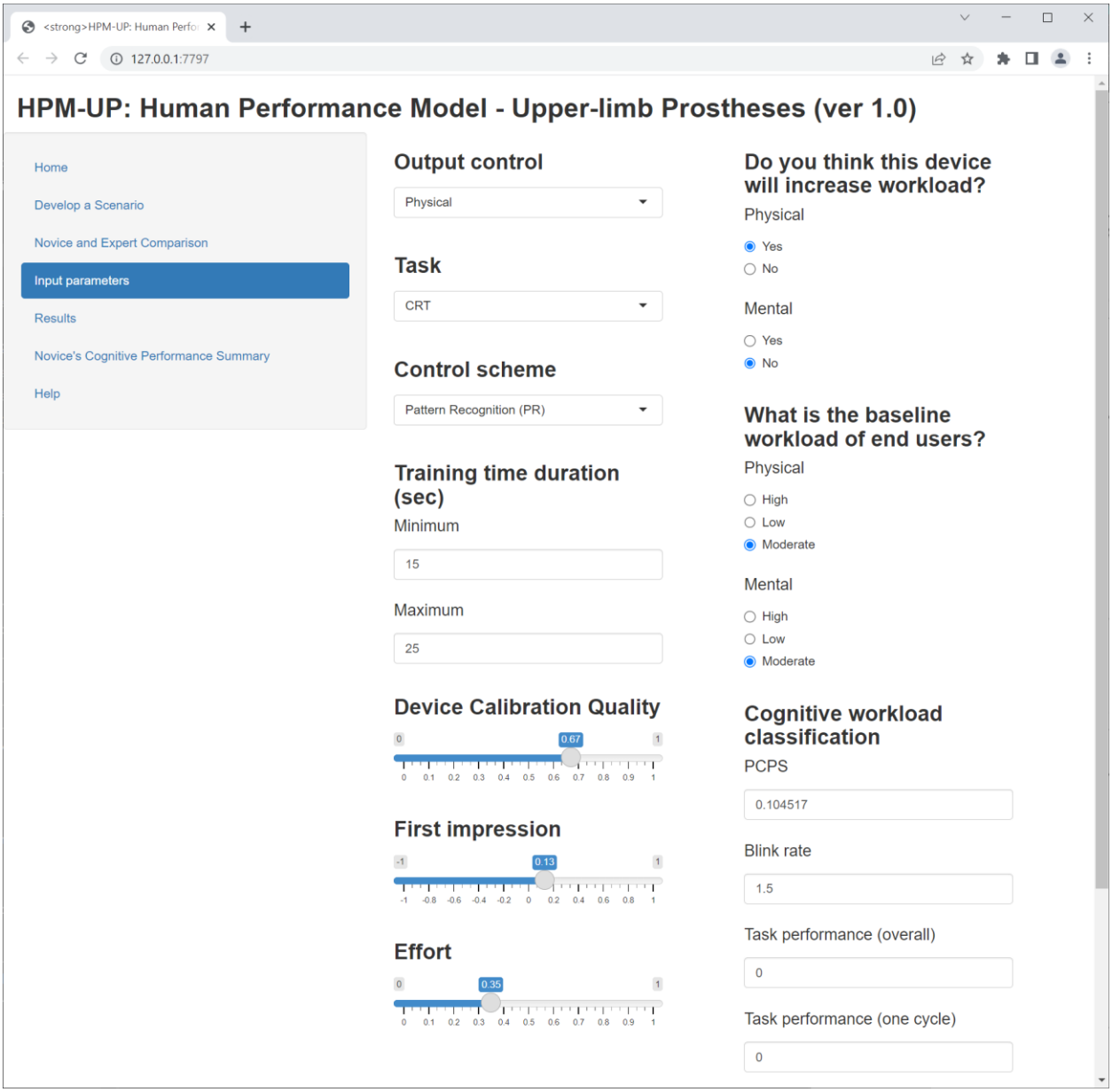


Figure 24. Input parameters

To classify cognitive workload, the variables mentioned in section 3.2.6.1 (Data labeling) should be added as input measures including: PCPS, blink rate, task performance, number of cognitive/perceptual/motor operators, memory load, and number of training trials (Figure 24). Experimental data (i.e., PCPS, blink rate, task performance, and time to accomplish one cycle) can be gathered from pilot tests. CPM outcomes (i.e., number of cognitive/perceptual/motor operators,

memory load) can be generated once the analysts develop a scenario from the scenario development tab or load a developed scenario (CSV format file).

3.3.4. Model Output

3.3.4.1. Six usability dimensions

Once all the input parameters are added, analysts can see the outcomes in terms of the six usability dimensions (Figure 25).

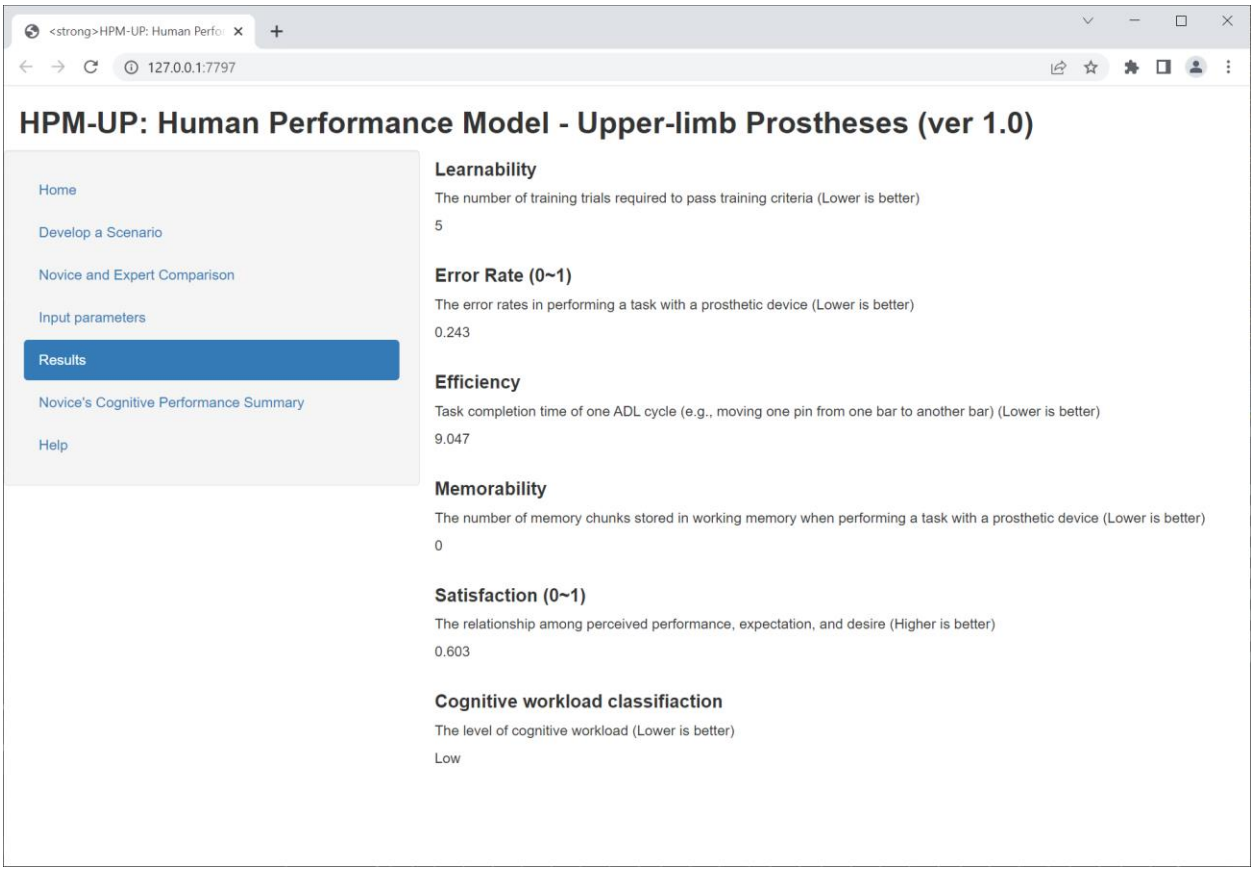


Figure 25. Results – Six usability dimensions

A literature review was conducted to provide a rule of thumb to interpret the outcomes of HPM-UP (Table 11). However, it is important to note that the outcomes depend on several factors

such as individuals' physical condition, the amount of time they spend using the prosthesis per day, and the complexity of the tasks they are performing with it. For example, if the users pass the training criteria in 3-5 training trials, the device has acceptable learnability (Park et al., 2020). The error rate refers to the percentage of times the device fails to perform a task correctly. A previous study with the PR configuration found that the error rate reduced from 12.85% to 11.55% after using the device in a home trial (Mohebbian et al., 2021). The threshold of memory load was defined as 3 to 5 chunks of information (Cowan, 2010). For the efficiency dimension, prior studies recommended to have at least 80% of experts' performance (without errors) (Park et al., 2022; White et al., 2017). The minimum satisfaction scores were defined based on clinicians' evaluation of different prosthetic device types (Rekant et al., 2022) as shown in Table 11. Lastly, the desired CW was set to "low" as CW is one of the challenges with using prosthetic devices.

Table 11. Thresholds to interpret the findings

Dimension	Threshold	Reference
Learnability	≤ 3-5 training trials	Park et al. (2020)
Error Rate	< 15%	Mohebbian et al. (2021)
Memorability	3-5 chunks of information	Cowan (2010)
Efficiency	~ 80% of the experts' performance	Park et al. (2022); White et al. (2017)
Satisfaction	Body powered device: > 45% Myoelectric device: > 50% Cosmetic device: > 50%	Rekant et al. (2022)
Cognitive workload	Low	Geurts et al. (1991); Heller et al. (2000) ; Hofstad et al. (2009);

3.4. Benchmark Model Development

A benchmark model was developed using the CPM-GOMS method and ACT-R (working memory module) in Cogulator software (Estes, 2017) to be compared with HPM-UP and human-subject experiment outcomes.

4. MODEL VALIDATION WITH EXPERIMENT 1: HUMAN SUBJECT EXPERIMENT WITH A PHYSICAL PROSTHESIS[**]

4.1. Objective

The objective of this experiment was to collect human-subject data with a physical prosthetic device for performing ADLs and use these data as a basis for the validation of the HPM-UP.

4.2. Participants

Thirty (Males=18, Females=12) able-bodied participants were recruited for this experiment (Age: *M*=22.4 yrs.; *SD*=2.4 yrs.) from North Carolina State University. All participants had 20/20 or corrected vision with no prior experience using a prosthetic arm or a myoelectric exoskeleton for upper limbs. The study protocol was approved by Texas A&M institutional review board (IRB) (IRB2021-0665).

Based on our prior literature review (Park and Zahabi, 2022a), it was found that about 60% of prior studies were conducted with able-bodied participants using bypass devices. Approximately 12% of the studies included both able-bodied and amputee participants. About 28% of the studies were conducted with amputees. For those studies with able-bodied participants, bypass devices were developed using various input signals such as EMG, IMU, FMG, and motion tracking. In addition, bypass devices were used to study the effects of feedback modality and training schedule on cognitive workload. Therefore, bypass devices with able-bodied participants

[**] Part of this chapter is reprinted with permission from J. Park et al., "Cognitive Workload Classification of Upper-limb Prosthetic Devices," *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*, Orlando, FL, USA, 2022, pp. 1-6, doi: 10.1109/ICHMS56717.2022.9980676. Copyright 2023 by IEEE.

were included in this study as they are devices that allow an able-bodied user to activate a terminal device with similar controls that an amputee would use to operate a custom-made prosthesis (Bloomer et al., 2020). Furthermore, based on our literature review and previous studies on prosthetic devices, recruiting amputee participants for human subject experiments is challenging, and therefore, several studies used able-bodied participants to assess prosthetic devices' usability and cognitive workload (White et al., 2017; Zhang et al., 2016b).

A priori power calculations were conducted using G*Power 3.1 to determine the sample size (Buchner et al., 2017; Faul et al., 2009; Serdar et al., 2021) for repeated measures ANOVA statistical test, with α=.05, power ($1-\beta$) of .8, effect size of .25, and correlation among repetitive measures of .5. The effect size was determined based on prior studies assessing the usability of upper limb prosthetic devices (White et al., 2017). In addition, the sample size was larger than the average number of participants used in prior studies assessing cognitive workload of prosthetic devices with able-bodied subjects (i.e., M=13.46, SD=6.49) (Park and Zahabi, 2022a).

4.3. Apparatus

4.3.1. Prosthetic Device

The experiment used the North Carolina State University Neuromuscular Rehabilitation Engineering Laboratory Lab prosthetic devices (i.e., Utah Motion Control Standard Electric Terminal Device) as shown in Figure 26. The open/close and pronation/supination motions could be controlled via wrist flexion and extension in the DC mode. To switch between the modes (e.g., open/close to pronation/supination), the user must co-contract their muscles, i.e., make a fist. Thus, two EMG signal inputs are required for this mode, including one on a wrist extensor muscle and the other on a wrist flexor muscle.

In the PR mode, the open/close and pronation/supination motions are controlled via their natural hand motions (e.g., prosthesis pronation is achieved by pronation of the intact hand for able-bodied subjects or imagined pronation of the missing hand for amputees). Four EMG signal inputs are typically used for this mode. Electrodes' locations were selected via palpation of the forearm during hand open/close and wrist pronation/supination.

The CC mode was generally similar to the PR configuration. The major difference was that the PR is a classifier that can only predict one gesture at a time (i.e., hand open, hand close, wrist pronate, wrist supinate, no movement), however, the CC can continuously predict velocities for both degrees of freedom, which means that a participant can control both the hand and wrist at the same time. The CC configuration uses a neural network algorithm that constantly and simultaneously predicts joint angles for open/close and pronation/supination. The number of hidden layers and number of neurons used in each hidden layer of the neural networks was increased during the setup by using feedback from participants until they reported good and consistent performance. Velocity was estimated at each time stamp (every 100ms) by taking the difference between the current and last predicted angles and dividing by the update period (100ms). The voltage to the motors was set proportional to the estimated velocity.

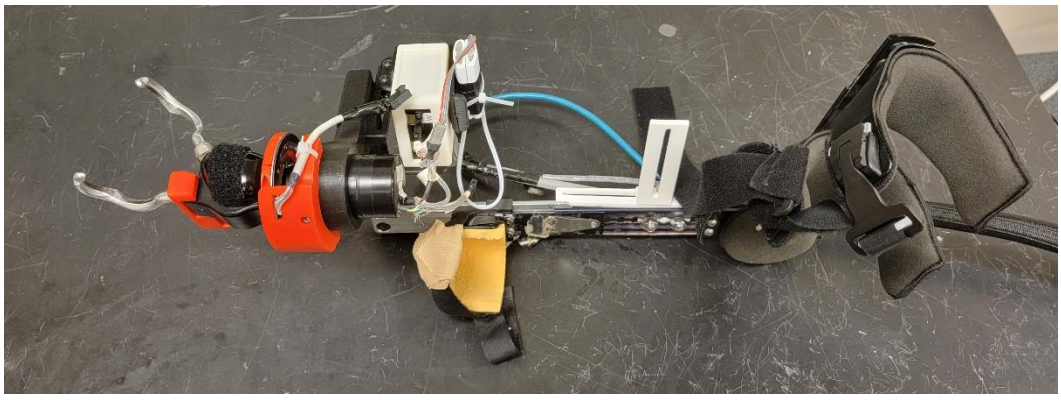


Figure 26. The prosthetic device used for the human subject experiment (Reprinted from Park et al., (2022). © 2023 IEEE)

4.3.2. EMG Sensors

To capture the EMG signals, Motion Lab Systems MA300 Desktop Unit was used. Four-channel surface EMG signals were acquired from four extrinsic hand- and wrist-related muscles, including extensor carpi radialis longus (ECRL), extensor digitorum (ED), flexor carpi radialis (FCR), and flexor digitorum (FD), as shown in Figure 27.

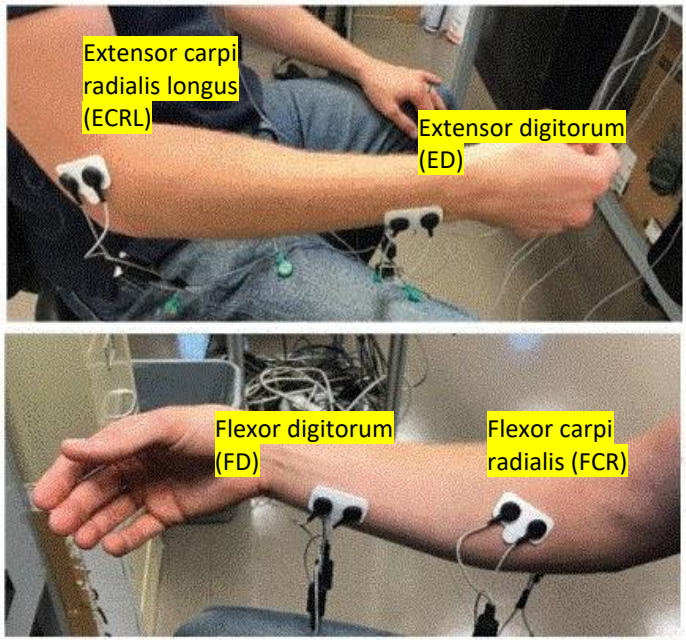


Figure 27. EMG sensor placement

Surface EMG was used to capture input signals for each control scheme. EMG signals were measured using gelled bipolar electrodes. Selected EMG recording sites were cleaned with alcohol wipes before electrode placement. A ground electrode was placed over the subject's right collar bone. The analog EMG signals were sampled at 1,000 Hz.

For the DC scheme, two EMG electrodes were placed over the belly of the extensor and flexor digitorum muscles based on palpation and the anatomical locations, respectively. Electrode placement was checked to capture clear EMG signals with sufficient signal-to-noise ratios and that

individual muscle activations could be identified by various recording channels. Signals were filtered with a 20–450 Hz bandpass filter. The signal magnitude was estimated by calculating the mean absolute value (MAV) of 50ms samples of EMG data. If the magnitude of one muscle was more prominent than a predefined threshold value, a corresponding prosthetic motor was activated; the speed of the motor was proportional to the magnitude of the EMG signal. If the magnitudes of both EMG signals were above threshold values, the prosthesis control mode (either wrist rotator or hand) was switched. Hence, the prosthesis user controlled two directions of movement with one DOF (e.g., wrist pronate and supinate) using finger extension and flexion, whereas forearm muscle co-contraction (power grip/making a fist) was used to switch between DOFs.

For the EMG PR control scheme, the targeted muscles included the flexor carpi ulnaris, flexor carpi radialis, extensor carpi ulnaris, extensor carpi radialis, extensor pollicis longus, and palmaris longus. The EMG PR algorithm was insensitive to EMG crosstalk, therefore, targeting the exact muscles' EMG recording was unnecessary. Instead, selected EMG electrode sites were accepted if EMG patterns during hand open, hand close, supination, and pronation were visually distinguishable from one another. The input EMG signals were filtered first and then segmented by overlapped sliding windows. In each window, the EMG signals extracted four time-domain features (MAV, number of zero crossings, waveform length, and number of slope sign changes) from each input channel.

All features were connected as vectors and then fed into a linear discriminant analysis (LDA) classifier. The classifier determined a user's intended movement. There were four active classes of movement (hand open, hand close, wrist pronation, and wrist supination) and one static class (no movement). The LDA classification decision was passed to a prosthesis motor selector, which activated the motor according to the intended movement, and set the speed of the motor

proportional to the magnitude of EMG signals. In addition, a sensor fault-tolerant mechanism was included to ensure system robustness against disturbances at the sensor level. Movement decisions were made every 50ms on features extracted from 150ms of the EMG data using the PR control strategy. The users controlled the DOF of the prosthesis using intuitive muscle contractions. Note that the LDA methodology and TD features were selected based on previous research (Butt et al., 2018). They indicated comparable EMG pattern classification accuracy, as compared to other classifiers and EMG features. In addition, the LDA method is simple to compute and requires less computational power for real-time implementation. Hand gestures and hook movements for each control scheme is summarized in Table 12.

Table 12. Hand gestures and its hook movement

Prosthetic Movement	Pattern Recognition & Continuous Control Hand Movement	Direct Control Hand Movement
 Open Prosthetic	 Open Hand	 Extend Hand at Wrist
 Close Prosthetic	 Close Hand	 Flex Hand at Wrist
Mode Change	Not Applicable	 Power Grip to Change Between Mode 1 and 2
 Supinate (Rotate Clockwise) Prosthetic	 Supinate Hand	 Extend Hand at Wrist
 Pronate (Rotate Counterclockwise) Prosthetic	 Pronate Hand	 Flex Hand at Wrist

4.3.3. Eye Tracker

A pupil-core eye-tracking system was used to capture pupillometry measures as a basis for inferring the CW of participants while using prosthetic devices and performing ADLs (Figure 28). The Pupil-core system consisted of two cameras and an infrared light-emitting pod. When reflected on the eyes, the light emitted from the pod is captured by the cameras and the pupil's outline. Eye movements were captured at a frequency of 120 Hz for each pupil with a gaze accuracy of 0.6°.

Figure 28. Eye-tracking glasses

4.4. Task

Previous studies on usability evaluation of prosthetic devices used a variety of testbeds, such as Box & Block (B&B), Clothespin Relocation Task (CRT), Jebsen Hand Function Test (JHFT), and Southampton Hand Assessment Procedure (SHAP). In this experiment, the CRT and SHAP tasks were selected based on their coverage of various upper-limb movements (Park et al., 2020), such as (a) Shoulder abduction-adduction; (b) Shoulder flexion-extension; (c) Shoulder internal-external rotation; (d) Flexion-extension of the elbow; (e) Pronation-supination of the forearm; (f) Flexion-Extension of the wrist; and (g) Radial-Ulnar deviation.

4.4.1. Clothespin Relocation Task

CRT is a commonly applied ADL for assessing upper limb prostheses (Stubblefield et al., 2005; Zahabi et al., 2019b). It requires participants to move as many pins as possible from one bar to another within 2 minutes. The experiment included three trials. Between each trial, there was a 2-minute rest. The CRT workstation (Figure 29) was mounted on a table and was adjusted to a comfortable height for the participant.

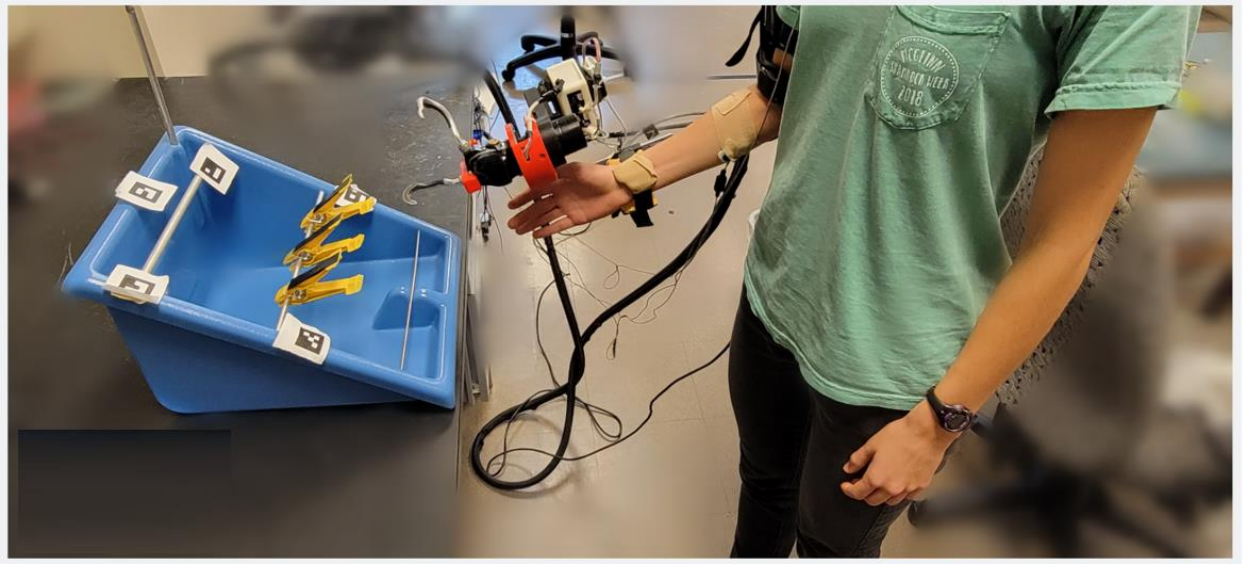


Figure 29. The clothespin relocation task (Reprinted from Park et al., (2022). © 2023 IEEE)

4.4.2. Southampton Hand Assessment Procedure – Door Handle

The SHAP task required participants to rotate the door handle using a power grip until it was fully open, then release the handle as quickly as possible. The participants were asked to do this task five times as quickly as possible. Similar to the CRT, the experiment included three trials. Between each trial, there was a 2-minute rest (Figure 30).

For the SHAP door handle task, the participant's elbows should be at a 90° angle. The SHAP form-board was placed in front of the participant with the blue side facing upward, approximately 8cm from the front edge of the table. The door handle task was demonstrated to the

participant using slow, precise movements, ensuring that the participant is aware of the proper grip for completing the task. The demonstration was carried out using the corresponding hand under assessment to avoid any confusion for the participant.

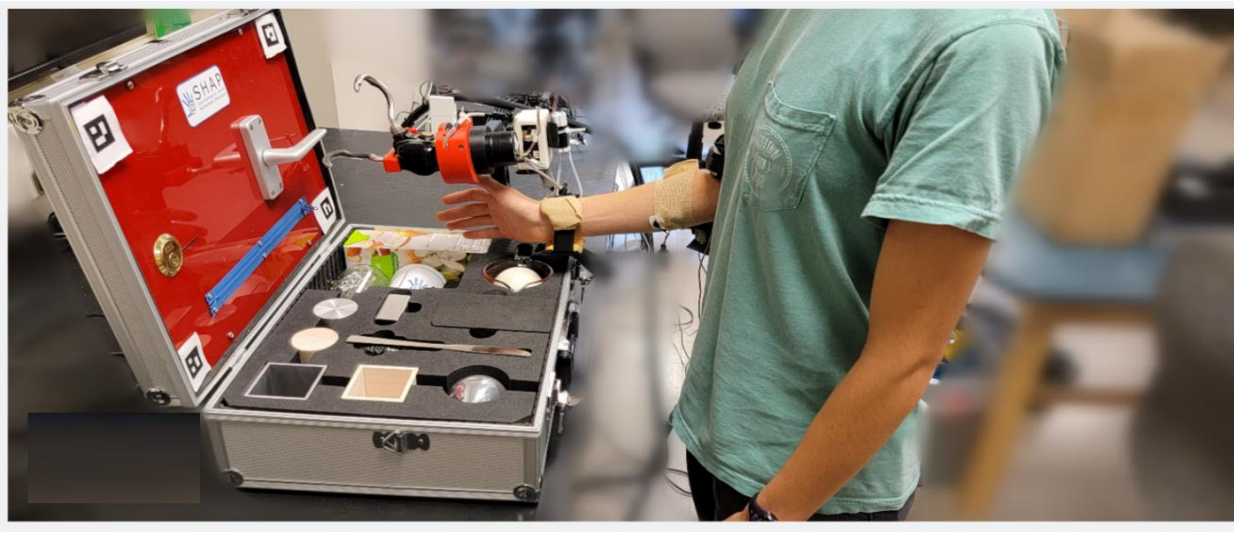


Figure 30. The SHAP door handle task

4.5. Experiment Design

The experiment followed a between-subject design in which each participant was randomly assigned to one of three types of prosthesis (i.e., DC, PR, or CC). This approach was selected to reduce learning effects that might occur for participants as a result of working with different prostheses across multiple test trials. It is recommended to use a between-subjects design to avoid "demand effects" in behavioral studies (Zahabi, 2017). Participants can develop a sense of an experimenter's intention during the progress of the experiment as a result of being exposed to all manipulations and may adapt their behavior accordingly (Rosenthal, 1976). In addition, between-subject designs are more conservative than within-subject designs in terms of potential subject-condition bias (Charness et al., 2012). Another motivation for using a between-subject design is that this design is more appropriate for long experiments in that it only provides one manipulation

to a participant. In the present study, if participants were exposed to all three control schemes, the duration of the experiment would exceed 5 or 6 hours, which could increase the potential for fatigue. Upon being assigned to a specific type of prosthesis, all participants experienced two tasks (i.e., CRT and SHAP door handle tasks), including three trials for each task.

4.6. Independent Variables

The only independent variable in this study was the device configuration with three levels including (1) DC, (2) PR, and (3) CC.

4.7. Dependent Variables

The dependent variables can be categorized into four types: task performance, eye-tracking measures, usability measures, and perceived workload ratings. Task performance measures included the number of pins moved within 2 minutes for the CRT and time to rotate the door handle five times sequentially for the SHAP task.

Eye-tracking measures included the percent change in pupil size (PCPS) and blink rate. PCPS has been used in previous studies to assess the effect of device configurations on cognitive workload (Zhang et al., 2016b). It was found that the PCPS has a higher value in mentally complex tasks than in more manageable tasks (Palinko et al., 2010). Blink rate has also been frequently used as an indicator of cognitive workload (Cardona and Quevedo, 2014; Fogarty and Stern, 1989; Martins and Carvalho, 2015). Blink rate is defined as the number of eye closures in a given period (White et al., 2017). Blink rate can be used to measure cognitive workload (Sirevaag et al., 1993; Van Orden et al., 2001), however, some studies found it to be more sensitive to assess visual

workload (Brookings et al., 1996). Eye blinks and blink duration decrease as visual workload increases (De Waard and Brookhuis, 1996).

Usability was measured using two questionnaires, including (1) QUEST 2.0 (Appendix A) (Demers et al., 2002), which assesses a person's positive or negative evaluation of those distinct dimensions of the assistive device that are influenced by one's expectations, perceptions, attitudes, and personal values, and (2) USE (Appendix B) (Lund, 2001) which measures the subjective usability of a product or service, thus, it can be applied not only for prosthetic devices but also other domains. Participants were asked to rate the usability of the device after the last trial.

NASA-TLX (Appendix C) was used to measure subjective workload, as this measure has been used extensively in prior studies in the prosthesis device context (Connan et al., 2016; Deeny et al., 2014a; Markovic et al., 2018). Participants were asked to rate their perceived workload using the NASA-TLX questionnaire after each trial.

4.8. Procedure

Both the University of North Carolina and Texas A&M Institutional Review Boards approved the experiment protocol, and all participants signed informed consent before any experimental procedure. The experiment took place in a laboratory without windows to limit the effect of illuminance on pupillometry data. The illuminance level was relatively consistent over time with photometer readings of 170–200 lx in the area where participants experimented. The experimental setup included the prosthesis device, CRT workstation, SHAP workstation, and the eye-tracking system (Pupil-core, Germany).

At first, participants signed the informed consent form, an informed consent form addendum for research during the COVID-19 pandemic, and a demographic questionnaire. After

the participants signed all documents, they were asked to complete the Edinburgh Handedness Test (Oldfield, 1971) and the Purdue Pegboard Test (PPT) (Tiffin and Asher, 1948; White et al., 2017). The PPT was conducted three times to determine if they fell within the range of "normal" manipulative dexterity. Participants were recruited for the experiment if they received a right-hand dominance score of 0.7 or greater based on the Edinburgh Handedness Test, and their PPT score was no more than one standard deviation below the normal mean dexterity for their age and gender group (Tiffin and Asher, 1948) (Figure 31).

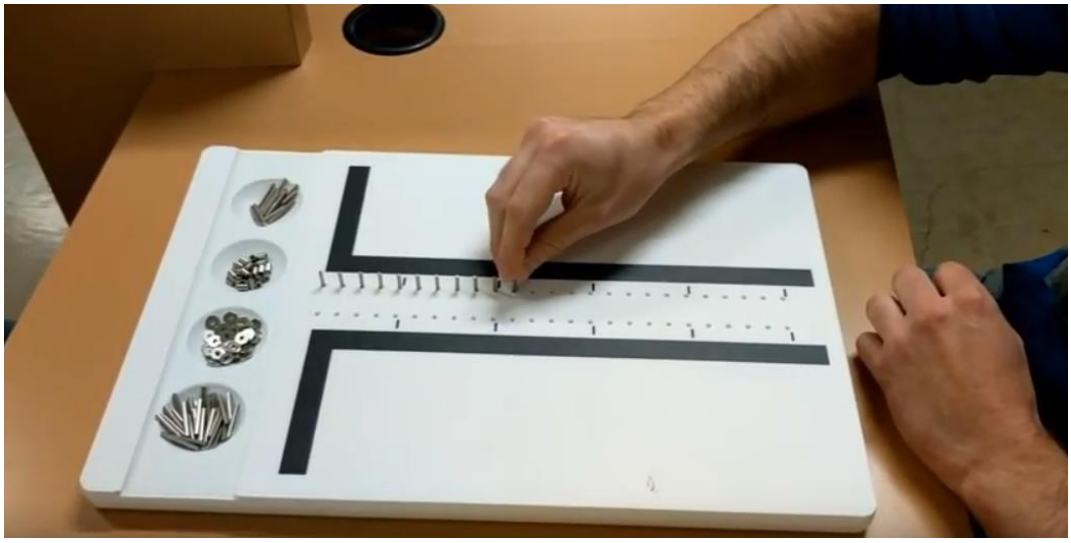


Figure 31. A participant is on dexterity test with Purdue Pegboard Test kit

Participants donned the prosthetic adapter during the experiment, and EMG electrodes were placed on their skin based on the assigned control mode. A verbal description of the prosthesis DOF and control strategy was provided. For participants assigned to the DC group, the prosthesis was activated during the EMG threshold configuration procedure. Participants were allowed to practice controlling the device until they reported comfort with the DC control. Participants then advanced to the formal training period. Participants assigned to the PR group were instructed to perform specific arm motions and to observe a feedback display (hand open,

hand closed, wrist pronated, wrist supinated, and relaxed hand and wrist). In total, 5 s of rest were allowed between each posture. Two sets of data were acquired, with the subject maintaining different arm positions in the sagittal plane. In total, 5 s of rest were enforced between sets of data collection. Participants assigned to the CC group were asked to perform 10s trials three times for each movement type - isolated hand open/close, isolated wrist pronation/supination, and simultaneous movements - at a 0.25Hz tempo set on a metronome, resulting in 9 total trials. Angles of the metacarpophalangeal joints and the wrist's rotation angle were recorded using a Leap Motion Controller placed approximately 4” below the subject’s hand at 120Hz simultaneously with EMG data. The MAV of the EMG was calculated with a 200ms sliding window adjusted in 10ms increments, and the joint angle data were down-sampled to 100Hz to match the EMG data. The processed EMG and motion data were used to train two neural networks for the 2 DOF. Gains for the controller's output and thresholds to reduce small unintentional movements from the user were adjusted using feedback from them. After the classifier was trained, users were allowed to practice controlling the device until they reported comfort with the control.

Participants received training for their assigned control mode. The task-specific training assessed participant mastery of device handling and the respective control mode while completing the CRT. The training session required participants to use the prosthesis to move three clothespins from a horizontal bar at the base of the workstation to a vertical bar extending upward on the clothespin apparatus. They began with the movement of the rightmost clothespin and, as quickly as possible, completed all pins. An experimenter recorded the time to move the three consecutive clothespins. If a participant dropped a clothespin, they were required to restart the trial. A training criterion was established based on pilot test data generated from learning curve analysis, including when participants achieved asymptotic performance with the device and at what level (task time).

If the average task completion time of three sequential trials was within 15–25s for the PR, 20–35s for the DC, and 16-23s for the CC mode, the participant passed the training and proceeded to the actual experimental trials. Upon completion of the training trials, the eye-tracking system was calibrated for the participants, and they could begin the experiment trials after having 5 minutes of rest. During the training and actual trials, participants were standing in front of the task workstation.

Participants were provided instructions on how to complete the two tasks for experiment trials. For CRT trials, the instruction included moving as many clothespins as possible from the horizontal rod to the vertical rod and back within 2 minutes. The number of successfully relocated clothespins was recorded at the end of each trial. For SHAP – Door Handle, participants were instructed to rotate the handle five times as fast as possible. The participant's eyes were tracked throughout each trial. All participants completed a total of three trials for each task and were provided with a 2-min rest period after each trial. After each actual trial, participants filled out the NASA-TLX questionnaire. After all trials, they also filled out the QUEST 2.0 and USE forms.

4.9. Hypotheses

The following hypotheses (H) were formulated for this study. Since both human subject data and benchmark model do not generate all usability dimensions, hypotheses were generated to enable comparisons between the HPM-UP and either human data or the benchmark model.

- Learnability: The results of HPM-UP learnability dimension would be similar to the human-subject data (H1)

- Error rate: Use of the CC configuration would lead to the lowest error rate followed by the PR and DC configurations (i.e., CC < PR < DC) (H2) (White et al., 2017; Zahabi et al., 2019b; Zhang et al., 2016b)
- Memorability: The results of HPM-UP memorability dimension would be similar to the benchmark model (H3)
- Efficiency: The results of HPM-UP efficiency dimension would be similar to the human subject data (H4-1) but would be significantly different from the benchmark model results (H4-2).
- Satisfaction: The results of HPM-UP satisfaction dimension would be similar to the human-subject data (H5)

4.10. Data Analysis

Due to the limited number of data points for each device configuration, nonparametric analysis was conducted to assess the differences in usability dimensions among the human subject data, HPM-UP, and benchmark model. For the comparison between two sets of data, Wilcoxon rank sum test was conducted (Wilcoxon, 1992). The Wilcoxon test statistic "*W*" was used to determine the significance of the difference. Kruskal-Wallis rank sum test was conducted if there were more than two groups of data (Kruskal and Wallis, 1952). "*H"* statistic was used to determine the significance of the difference of the median of each group. "*H*" statistics was compared with a critical cutoff point determined by the chi-square distribution (chi-square is used because it is a good approximation of *H*, especially if each group's sample size is bigger than 5). For the post-hoc analysis, Dunn's Kruskal-Wallis multiple comparison was used (Dunn, 1964). All the statistical analysis was conducted using R 4.0.5. Effect size for Wilcoxon signed-rank test was

calculated with $r = \frac{Z}{\sqrt{n}}$, where Z-score is a test statistic and standardized score of *U*-value calculated from Mann-Whitney *U*-test (Tomczak and Tomczak, 2014) and *n* is the total number of observations. The effect size of Kruskal-Wallis test was calculated using Eta-squared (Rosenthal, 1986).

4.11. Results

4.11.1. Hypothesis Test Results

Table 13 illustrates the descriptive statistics results regarding the comparison among the human subject data, the HPM-UP modeling approach, and the benchmark model. The benchmark model does not provide learnability, error rate, satisfaction, and CW, and therefore, these cells are marked with "N/A" in Table 13. However, HPM-UP was able to generate all six dimensions.

Table 13. Descriptive statistics from Experiment 1 (mean (sd))

Factors (definition)		Human subject data			HPM-UP			Benchmark model		
		DC	PR	CC	DC	PR	CC	DC	PR	CC
Learnability		7.8 (3.46)	5.50 (1.50)	9.80 (4.40)	8.20 (6.54)	7.00 (3.87)	10.00 (5.00)	N/A		
Error rate		N/A			36.43 (3.36)	25.8 (3.55)	28.14 (8.76)	N/A		
Memorability		N/A			4.79 (0.04)	0	0	4.80 (0.00)	0	0
Efficiency	CRT	17.91 (6.76)	12.90 (5.77)	13.40 (2.30)	19.03 (2.48)	11.19 (3.91)	14.19 (5.62)	13.3 (0.00)	6.9 (0.00)	6.9 (0.00)
	SHAP	11.03 (2.31)	11.26 (4.95)	18.91 (7.10)	11.96 (0.70)	12.22 (1.42)	13.18 (1.71)	8.2 (0.00)	8.6 (0.00)	8.6 (0.00)
Satisfaction (0-1)		0.75 (0.12)	0.64 (0.23)	0.65 (0.18)	0.61 (0.06)	0.61 (0.11)	0.56 (0.16)	N/A		
Accuracy of Cognitive Workload Classification (%)		N/A			88.89	66.67	77.78	N/A		

A summary of hypothesis test results is shown in Table 14. All the hypotheses were supported except for H2. For the *learnability* dimension (H1), there was no significant difference between the human subject data and HPM-UP generated data based on the Wilcoxon Signed-Ranks Sum test ($W = 436.5$, $p > .05$). On average, all configurations required 7 to 10 training trials

to pass the device training criteria. The benchmark model does not have this functionality and therefore, was marked as "Not Applicable (N/A)".

Table 14. Summary hypothesis test results (Experiment 1)

Hypothesis ID	Hypothesis	Test Result	Test statistics, p-value, effect size
H1 (Learnability)	The results of HPM-UP learnability dimension would be similar to the human-subject data.	Supported	W = 436.5, p = .85, r = .03
H2 (Error rate)	Use of the CC configuration would lead to the lowest error rate followed by the PR and DC configurations.	Refuted (DC=PR=CC)	$H(2)$ = 1.57, p = .47, η^2 = .06
H3 (Memorability)	The results of HPM-UP memorability dimension would be similar to the benchmark model	Supported	W = 65, p = .23, r = .28
H4 (Efficiency)	(H4-1) The results of HPM-UP efficiency dimension would be similar to the human subject data	Supported	Z = 0.26, p = .79, r = .03
	(H4-2) There would be a significant difference between the HPM-UP efficiency dimension results and the benchmark model results	Supported	Z = -4.54, p < .001, r = .64
H5 (Satisfaction)	The results of HPM-UP satisfaction dimension would be similar to the human-subject data	Supported	W=413, p = .59, r = .07

H2 was refuted as there was no significant difference among configurations in terms of *error rate* ($H(2) = 1.57$, $p > .05$). According to the computational logic of error rate in HPM-UP, the error rate depends on learnability. There was no significant difference in *learnability* among different configurations from the human subject data, which led to not having any significant difference in error rate as well. This trend was also found from the HPM-UP outcomes ($H(2) = 1.57$, $p > .05$). The benchmark model does not provide error rate and therefore, was not included in this comparison.

There was no significant difference in *memorability* between the HPM-UP and the benchmark model (H3) ($W = 65$, $p > .05$). In the DC configuration, participants needed to memorize and recall two mode changes (supination/pronation or open/close) and three gestures (open, close, rotation). Using the PR and CC configurations did not involve memory chunks as

these configurations were intuitive. This information could not be captured from the human subject data or video analysis, and therefore, was not included in the comparison.

For efficiency, both hypotheses (i.e., H4-1 and H4-2) were supported. Based on Dunn's Kruskal-Wallis multiple comparison, there was no significant difference between the human subject data and HPM-UP efficiency outcomes (H4-1) ($Z = 0.26$, $p > .05$). However, there was a significant difference between HPM-UP and the benchmark model (H4-2) ($Z = -4.54$, $p < .001$). There was also no significant difference in *satisfaction* between the human subject data (USE questionnaire – Satisfaction dimension) and HPM-UP ($W=41$, $p > .05$).

4.11.2. Cognitive Workload Classification

The outcomes of each classifier with different targets and tasks are presented in Table 15. The best model was NB with two classes and it resulted in 76% of classification accuracy for the CRT (model No. 3 in Table 15), considering all metrics including AUC, precision, recall, and F1 score (Grandini et al., 2020; Sokolova and Lapalme, 2009). In addition, RF models with two classes showed 70% classification accuracy in CRT (model No. 1) and SHAP (model No 4) tasks. The selected features in the best performance models included pupillometry data, training and task performance measures, some CPM generated outcomes such as the number of memory chunks, and device configuration.

To improve the reliability and generalizability of ML results, each model was run with 15 random seeds per suggestion from Colas et al. (2019) and the average prediction performance was calculated. In 12 out of 15 NB runs, the model showed significantly higher classification accuracy than random guessing (0.56) (McDonald et al., 2019).

Regarding the target variable, in general, classifying the NASA-TLX scores into smaller number of classes led to better algorithm performance than having larger number of classes under the clustering algorithms (i.e., algorithms in *NbClust* package).

Table 15. Summary of classification results by taking different classes as targets

Task	No.	Classifier	Target	Accuracy	AUC	Precision	Recall	F1-Score
CRT	1	RF	Two classes	0.70	0.61	0.66	0.66	0.64
	2	SVC	Two classes	0.55	0.48	0.49	0.55	0.46
	3	NB	Two classes	0.76	0.67	0.67	0.70	0.67
SHAP	4	RF	Two classes	0.70	0.74	0.70	0.71	0.68
	5		Three classes	0.53	0.44	0.29	0.39	0.31
	6	SVC	Two classes	0.60	0.45	0.54	0.58	0.51
	7		Three classes	0.55	0.49	0.32	0.41	0.33
	8	NB	Two classes	0.62	0.61	0.56	0.60	0.54
	9		Three classes	0.48	0.55	0.43	0.44	0.40

The grid search time for every combination of classifiers, targets, and feature selectors suggested that the SVC and NB algorithms outperformed the RF in terms of computational cost (Table 16). Both SVC and NB performed within a few seconds. Among the three features selectors, SFS exhibited significantly longer computational time as compared to other two selectors.

Overall, the NB algorithm with two classes was selected as the best model (model No. 3 in Table 15). However, if the training time is not restricted, RF with two classes can also be a good model (models No. 1 and 4 in Table 15).

Table 16. Grid search time (seconds)

Classifier	Class	Feature selector		
		RFE	K-Best	SFS
RF	Two	4,092.6	1,282.2	21767.4
	Three	5,107.2	2,023.2	13,851.6
SVC	Two	25.2	22.8	747.6
	Three	33.6	35.4	428.4
NB	Two	23.4	26.4	582
	Three	20.4	19.8	303.6

4.12. Discussion

Overall, the hypothesis test results revealed that the HPM-UP generated outcomes were similar to the human subject data. This implies that the logic behind HPM-UP worked properly to estimate each usability dimension. However, the findings of the benchmark model were significantly different from the HPM-UP model and human subject data. The focus of the benchmark model was for modeling expert behavior without any errors. Unlike the benchmark model, HPM-UP showed closer results to the human data, especially for the TCT, as HPM-UP incorporated error rates based on the learnability dimension.

In the HPM-UP, learnability was used as an input for calculating other usability dimensions. This concept was based on the halo effect (Clifford and Walster, 1973; Thorndike, 1920). Including the SDCQ and FI in the equation was appropriate because the training criteria could not capture the individual differences. This means that although all the participants could pass the training sessions, they were not on the same level of the expertise in terms of controlling the prosthesis. Therefore, the SDCQ and FI factors were used to capture these individual differences. Furthermore, incorporating the physical and mental workload into the learning curve slope was effective to customize the model.

In general, the CC control scheme was similar to the PR configuration, in terms of error rate and efficiency. Although the capability of CC to drive multiple DOFs simultaneously was expected to allow participants to adopt more natural motion strategies to efficiently complete tasks, participants had a hard time to control the device. This was because sometimes the hook was rotating even though the participants had a neutral gesture and therefore, it was difficult for them to recover from errors. The other reason might be that the ADL tasks were too simple and therefore, could not show the differences between these configurations. Although the CC mode allowed simultaneous joint operations and natural arm motion in control, since the number of controllable joints in our study was limited to two and the task duration was short, the perceptual, cognitive, and motor demand in operating CC and PR control could be similar. There might be differences between the PR and CC configurations if the tasks become more complicated.

Not surprisingly, the overall performance of the CRT between PR and CC was similar. This is because the gestures to control the hook was the same in both configurations. In the SHAP task, however, the PR configuration exhibited better performance than the CC mode. In CRT, to pick up or release a pin, participants could either pronate or supinate. However, the SHAP task demanded participants to rotate the hook only in one direction in order to grab the door handle. Therefore, participants needed to spend more time adjusting the hook using the CC configuration than the PR.

4.12.1. Classification Performance

The findings suggested that CW of using prosthetic devices can be classified with reasonable accuracy and low computational cost. This study is the first investigation that included CPM outcomes as input features in ML algorithms. Some CPM outcomes (i.e., number of cognitive operators) and task performance features were included in the best models. This can

suggest the possibility of predicting CW of prosthetic devices without conducting human-subject experiments because task performance can also be modeled from the CPM outcomes. Some CPM outcomes such as the number of perceptual operators were not selected in the best models. This might be because the perceptual operators only appeared in the DC control scheme. In PR and CC configurations, there was no perceptual operators in the outcome of cognitive models because all perceptual operators were in parallel with cognitive or motor operators. However, if the task is more complex or with other prosthetic device configurations, more CPM outcomes might be included as important features in the algorithm.

There are several advantages of using CPM over human-subject experiments. For example, the analyst can conduct CPM in the early design process. It is a faster and safer approach than the experimental approach as it can minimizes human participant's involvement. It can also quantify and predict human behavior in natural tasks with simple tools such as Cogulator (Estes, 2017) or CogTool (John and Suzuki, 2009) based on human information processing theory. Lastly, CPM can also generate task performance related features without the need of conducting human-subject experiment and by using the results of task analysis and operator times from the literature (Estes, 2017).

This study suggested that multiple metrics should be considered to evaluate the ML algorithms and find the best model(s). For example, although the accuracy of some models was above 70% (e.g., model No. 1 in Table 15), their AUC was relatively low (e.g., 0.65). Precision and recall were also helpful to test the robustness of ML algorithms and to avoid "accuracy paradox" (due to unbalanced classes) (Afonja, 2017; Valverde-Albacete et al., 2013). For example, model 24 exhibited reasonable accuracy (0.67) among other algorithms for the SHAP task. However, its recall percentage was low (around 0.5), which implies that those models are not

useful for classifying CW when the target variable is not well-balanced. Considering only precision or recall scores individually is also not sufficient for evaluating ML algorithms. For example, we can have a recall score of 100% even though the accuracy of the model is low. In this case, precision will be close to 0. Thus, F1-score should be used to reflect the imbalance between precision and recall because it is a harmonic average between these two measures.

The results also revealed that task performance measures were more promising in predicting CW as compared to other features that were collected from the experiment. This finding is in line with the results of prior studies that found primary task measures as a key indicator of CW for prosthetic devices. Wood and Parr (2022) recently developed a questionnaire for measuring CW of prostheses as an extension of NASA-TLX, which is called prosthesis task load index (PROS-TLX). While validating their questionnaire, the authors used task performance as an indicator of CW as there was a high correlation between the task performance and the evaluated scores on PROS-TLX. Deeny et al. (2014a) also found high positive correlation under the complicated task condition between the task performance and the self-report workload score. Task performance measures have advantages in that they evaluate participants' performance on the task of interest directly. However, these measures often lack scientific rigor, making interpretation of the results difficult as unknown or uncontrolled factors may affect results rather than the intended manipulations in the study (Park and Zahabi, 2022a; Wilson and Schlegel, 2004; Wood and Parr, 2022). Therefore, some studies suggested using physiological measures of workload instead (Cain, 2007). We found that pupillometry measures especially the blink rate was selected as important features in the models. The results support the findings of previous studies that used eye-tracking data for measuring CW of prosthetic devices (White et al., 2017; Zahabi et al., 2019b; Zhang et al., 2016b). Eye-tracking measures have been widely applied to other domains to measure CW of

operators such as simulations for emergency responders (Appel et al., 2019), construction (Li et al., 2020), and fetal ultrasound examination (Sharma et al., 2021). One possible explanation why PCPS was not included as an important feature is that the task (e.g., moving a pin from one bar to another bar) was so rudimentary that the pupillometry data were not sensitive enough to differentiate the degree of CW among different device control schemes. This is in line with a previous study that found PCPS has a higher value in mentally complex tasks than in more manageable tasks (Palinko et al., 2010). Siegle et al. (2008) also suggested that pupil dilation can better reflect sustained information processing.

It was also found that the models with two classes performed better than models with three classes. This is intuitive from a general classification stance, since two classes are simpler than several classes to be classified as it has only one threshold. This is in line with previous studies that found smaller number of labels led to high classification accuracy (Nourbakhsh et al., 2013a; Wang et al., 2013).

Although the sample size was small, the NB algorithm exhibited reasonable average performance across multiple runs, which is in line with prior studies that found NB was more accurate than the SVM algorithm in classifying CW (Nourbakhsh et al., 2013b; Raufi, 2019). There are several advantages of NB that resulted in classification accuracy above 70%. First, NB can compensate for class imbalance (Murphy, 2006). Second, NB can perform well with small datasets (Huang and Li, 2011) and it is a fast and computationally effective approach (Jadhav and Channe, 2016; McCallum and Nigam, 1998). Third, the complement NB classifier used in this study were accurate in classifying CW, which suggests the possibility of using this classifier for other small datasets in the future.

The RF model did not perform as well as NB and some of the models had overfitting issues, which was mainly due to the detailed hyperparameter tuning on an extremely small dataset. Prior studies found that with small and imbalanced datasets, RF could generate either poor results due to a lack of diversity in the dataset or might cause overfitting (Tang et al., 2018). SVC also performs poorly when the dataset is imbalanced. This is mainly due to the weakness of the soft margin optimization (Batuwita and Palade, 2013) that allows SVC to make a certain number of mistakes and keep margin as wide as possible so that other points can still be classified correctly. This could result in the hyperplanes being skewed to the minority class when imbalanced data is used for training. The second reason is related to the issue of an imbalanced support vector ratio. That is, the ratio between the positive and negative support vectors becomes imbalanced and as a result, datapoints at the decision boundaries of the hyperplanes have a higher chance of being classified as negative. The major reason why RF generated longer computational time is that it included more hyperparameters, especially the number of trees in the forest and their levels, than the other two algorithms. Basically, training time complexity of RF is faster than SVC (Kumar, 2019). However, RF took much longer time than SVC due to the burden in hyperparameter tuning. In addition, the main limitation of RF is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions (Donges, 2021). SFS demanded extensive computational time because it is a wrapper method which needs to train the classifier for each feature subset, and therefore the method can be impractical.

The findings suggested two ML algorithms (RF or NB) for classification of CW for prostheses. Our intention was not to propose one specific algorithm or feature selector which should be used for all types of tasks mainly because depending on the characteristics of the dataset, several factors can affect the algorithm performance, including size and quality of the dataset,

complexity of the models, and potential biases in the dataset (Dietterich, 2000; Goodfellow et al., 2016; Murphy, 2012). We suggest researchers to use the findings of this study as a starting point in estimating CW of prosthetic devices and explore other models depending on the characteristics of their dataset.

4.12.2. Limitations and Future Work

The first limitation of this study was the small dataset that was used for training the models. Future studies with larger dataset are necessary to validate the findings of this investigation. Second, the models were generated based on the performance of able-bodied participants. The decision to work with an able-bodied population was made due to the limited number of trans-radial amputees in the surrounding area. In addition, since most patients currently use devices with DC modes (commonly used in myoelectric control), recruiting such patients could have produced a bias in their performance. Therefore, there is a need for further investigation with amputees, as an actual user population, to validate the models.

4.13. Contributions of Experiment 1

Evaluating usability of prosthetic devices early in the design process is crucial considering the cost and difficulty of testing with human subjects. Previous studies relied on questionnaires after conducting human subject experiments. Furthermore, recruiting amputee participants for human subject experiments are challenging and therefore, several studies used able-bodied participants to assess usability and cognitive workload of prosthetic devices (Park and Zahabi, 2022a).

The most unique contribution of this research was that it provided a method for early and objective usability evaluation of prosthetic devices. Previous studies focused on human subject

experiment and subjective evaluations for usability and mental workload assessment of amputees (Park and Zahabi, 2020; Park and Zahabi, 2022a). HPM-UP can be used to improve the usability of prosthetic devices using pilot tests and early-stage prototypes especially since conducting human subject experiments with amputee participants can be time-consuming and challenging.

The second contribution of this research is that it developed a human performance model for upper limbs (HPM-UP) based on top-down (theories) and bottom-up (data driven) approaches and the model was validated with a human subject experiment. Although previous HPMs such as ACT-R or QN-MHP could generate task completion time, memory load, or cognitive workload (i.e., NASA-TLX score), they were not able to provide estimation on learnability, error rate, and satisfaction. Furthermore, the dimensions in HPM-UP are interconnected. Learnability is the key to compute other dimensions, as the theoretical background to calculate all dimensions in HPM-UP is based on the learning performance. This connectivity has not been captured in previous methods. The third contribution of this study is that HPM-UP is the first HPM that has been developed in R-shiny package format. This can help researchers, designers, or practitioners use HPM-UP easily with a GUI without the need for hard coding.

The HPM-UP developed in this research has several advantages as compared to the benchmark human performance model. Including the *Learnability* dimension provided the possibility of measuring learning efforts of prosthesis devices. The model was successful as it could predict learnability accurately even for cases where participants needed more training (i.e., more than 10 trials) due to fatigue or low device calibration quality.

Error rate estimation feature of HPM-UP is closely related to learnability. The unique contribution of HPM-UP with error rates is that it tried to model the effect of errors not only on task performance (i.e., efficiency) but also *satisfaction* because the efficiency was used as an input

for the satisfaction dimension. The predicted satisfaction scores from the HPM-UP were similar to the results from human data and were significantly different from the benchmark model outcomes. Calculating the number of errors manually (by watching videos) was not feasible for the HPM-UP as the analysis on two minutes of the task requires more than several hours of work in millisecond. Also, it was not possible to figure out whether the specific hand gestures were errors as the videos did not provide any information on participants' intention.

In addition, since error rate affects efficiency, the model outcomes became closer to the human subject data than the benchmark model. In a previous cognitive modeling study (Zahabi et al., 2019b), it was assumed that participants were experts (which means that they do not have errors) in a certain task after they passed the training sessions, which is one of the main assumptions of many HPMs (Park and Zahabi, 2022b). However, this research found that participants could still make errors even after passing the training sessions. These errors were added to estimate the TCT and to calculate the *efficiency* dimension. Therefore, the model outcomes were closer to the human subject data than the benchmark model.

The benchmark model could calculate memory chunks. However, it only enables one cycle of the task, which is a limitation in simulating repetitive routine tasks. For example, the benchmark model was not capable of setting the end of the task with a time parameter. To do that, "for-loop" or "while-loop" are needed to set the time limitation (e.g., 2 minutes). However, there is no grammar for this in the benchmark model. Hence, to make comparisons, the same cycle of task was copied and pasted many times to calculate the number of memory chunks for 2 minutes. Also, fitting the task completion time exactly to 2 minutes in the benchmark model required removal of some lines of the code manually.

Another unique feature of HPM-UP is using a computational approach for quantifying *satisfaction* based on theories. Unlike previous studies that used questionnaires with a Likert-scale (QUEST 2.0 or USE), HPM-UP calculates satisfaction based on the expectation-confirmation theory.

Classification of cognitive workload with ML was another unique feature of HPM-UP. Although the dataset was extremely small, it was possible to train and generate models which could have a reasonable classification accuracy (i.e., above 70%). This module can shed light on the prosthetic device development with emphasis on cognitive workload as previous studies were mainly focused assessing physical workload of these devices.

5. MODEL VALIDATION WITH EXPERIMENT 2: HUMAN SUBJECT EXPERIMENT WITH A VIRTUAL PROSTHESIS

5.1. Objective

The objective of this experiment was to collect human-subject data in a virtual reality (VR) setting when performing ADLs and use these data as a basis for validating the results of HPM-UP.

5.2. Participants

Twenty (Males=13, Females=7) able-bodied participants were recruited for this experiment (Age: *M*=26.9 yrs.; *SD*=4.6 yrs.). The study was conducted at Texas A&M University. All participants had 20/20 or corrected vision with no prior experience using a prosthetic arm or a myoelectric exoskeleton for upper limbs. The study protocol was approved by Texas A&M IRB (IRB2021-0990D). This validation experiment was conducted with the DC and PR configurations since there was no significant difference between the CC and PR modes from Experiment 1 results.

5.3. Apparatus

The experiment setup included three modules: 1) the EMG/kinematic data collection and processing module, 2) the server module, and 3) the VR module. The expected input and output data formats for each module are described in the following sections to allow researchers to modify or replace any module while maintaining compatibility with the others.

Figure 32 provides an overview of the system architecture and is presented in the form of a flowchart to visualize the progression of generated EMG signals to virtual action commands.

Arrows pointing to and from the Ultraleap Leap Motion Controller and the HTC VIVE controller are dotted to indicate optional use within the system.

Module 1 is responsible for collecting the necessary physiological signals for different configurations as well as applying the respective algorithms to implement each method. The most recently detected action classification is then sent to the server module, which delivers it to the VR application module. Module 3 is responsible for translating the action command into animations and state changes for the virtual prosthesis. A data logging script records the various interactions within the VR application, action commands received from the server, and eye tracking data. The log file is represented by the session data element in Module 3 of Figure 34. Detailed hardware, software, and server implementation are described in Music (2022).

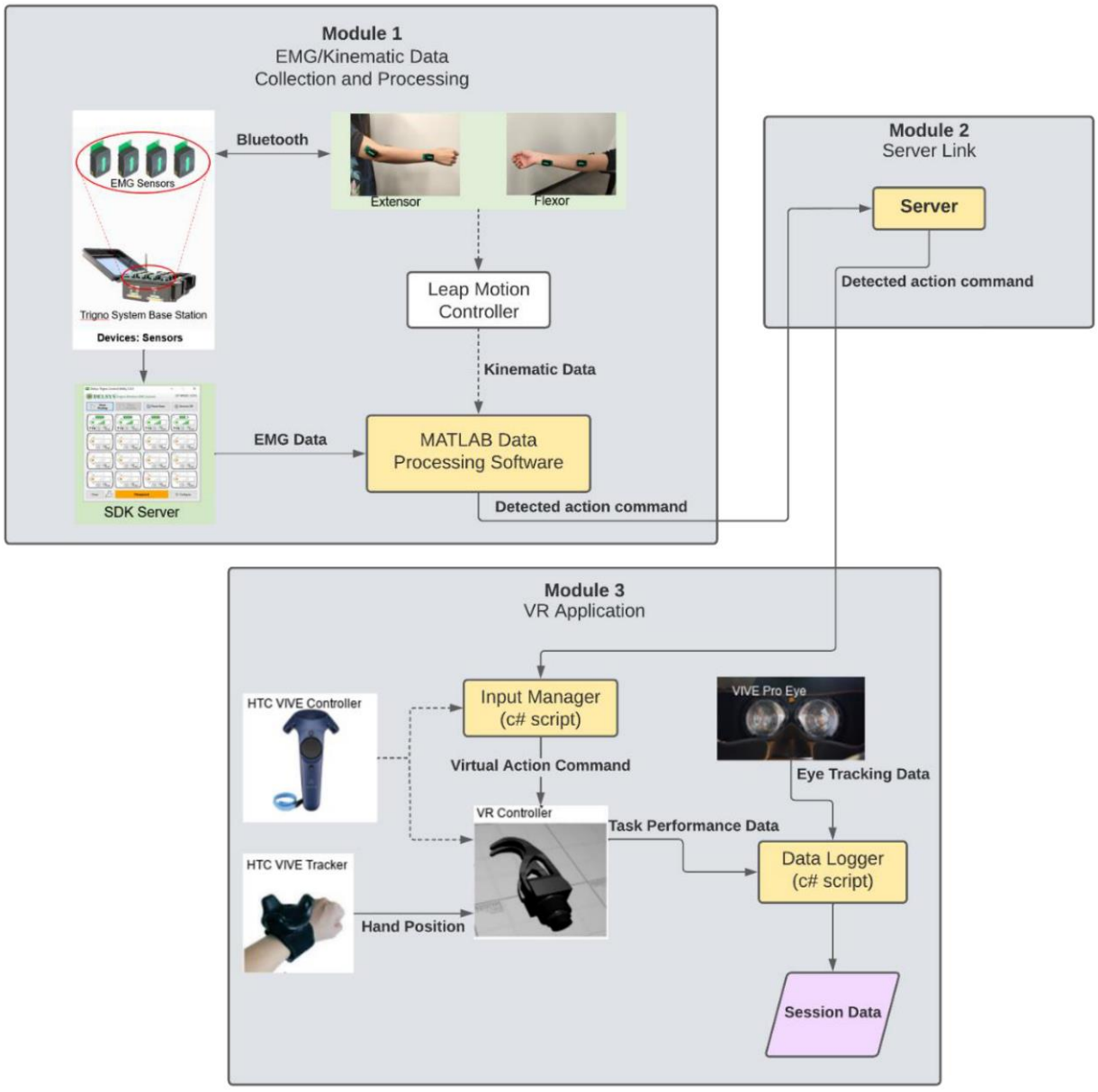


Figure 32. Flowchart of complete system architecture for EMG-based VR human-machine interface

5.3.1. Virtual Prosthesis Development

For the VR simulation, an HTC VIVE Pro Eye head-mounted interface (HMD) was chosen as the optimal VR system as it affords built-in eye tracking technology capable of recording gaze and pupillometry data at 120 Hz. The virtual reality application was developed on the Unity Game

Engine v2019.4.28f, a real-time development platform for 2D, 3D, VR, and augmented reality (AR) interactive applications (Music, 2022).

The virtual prosthesis was modeled after the Fillauer Motion Control Electric Terminal Device (ETD) 2. The ETD 2 was chosen as the model for the virtual prosthesis as it affords two degrees of freedom, and this is the minimum number of degrees of freedom required to make meaningful comparisons between the compatible prosthesis control modes. A side-by-side view of the virtual prosthesis model and the ETD 2 is shown in Figure 33.

Figure 33. The Motion Control ETD 2 prosthesis. The real-world ETD 2 prosthetic device (A) and the virtual ETD 2 model (B) are in the inactive motion state. Source: https://fillauer.com/products/proplus-mc-etd2/

The virtual model of the ETD 2 prosthetic device can perform the same motions as its real-world counterpart. The VR application provides mirrored models to support left- or right-handed subjects. A VIVE Tracker 3.0 device is secured to the dorsal side of the hand with athletic tape to track the position of the prosthetic model in virtual space to the position of an able-bodied user's hand, as shown in Figure 34. For transradial or wrist disarticulation amputees, the VIVE Tracker can be secured to the end of the vestigial limb in a similar manner.

Figure 34. VIVE Tracker 3.0 secured to the dorsal side of the hand via athletic tape for virtual prosthesis position tracking

The VIVE Tracker 3.0 provides only positional data to the virtual prosthesis, and the prosthesis rotation is locked so that the prosthesis extends from the user's body perpendicularly. This constraint is in place to allow supination and pronation to be controlled manually through the various control algorithms.

When the DC or PR control configurations are enabled, one of the five possible action commands are received from the server module at a time. These commands are received by a C# script that acts as an input manager, which updates five Boolean values which correspond to the five motion classes. A simple switch statement determines which Boolean value should be set to true and sets all others to false based on the most recently received command. All other scripts responsible for the virtual prosthesis animations and interactions determine their internal state based on these five Booleans in the input manager. The current implementation allows the system operator to choose whether to use proportional control or ON-OFF control. Proportional control is

enabled by default. Traditional ON-OFF control does not take EMG signal magnitude into account, meaning the speed of the prosthesis motion is constant and not controlled by the user.

5.3.2. EMG Sensors

For EMG signal collection, a Delsys Trigno Wireless Biofeedback system was used in conjunction with four Trigno Avanti Sensors. The sampling rate of the Trigno Avanti sensors was set to 1,111 Hz. The sensor placement was the same with Experiment 1. As DC only requires analysis of EMG signals from an agonist-antagonist muscle pair (Resnik et al., 2018), one sensor is placed on the flexor carpi radialis, and another is placed on the extensor carpi radialis longus for this method (Figure 35).

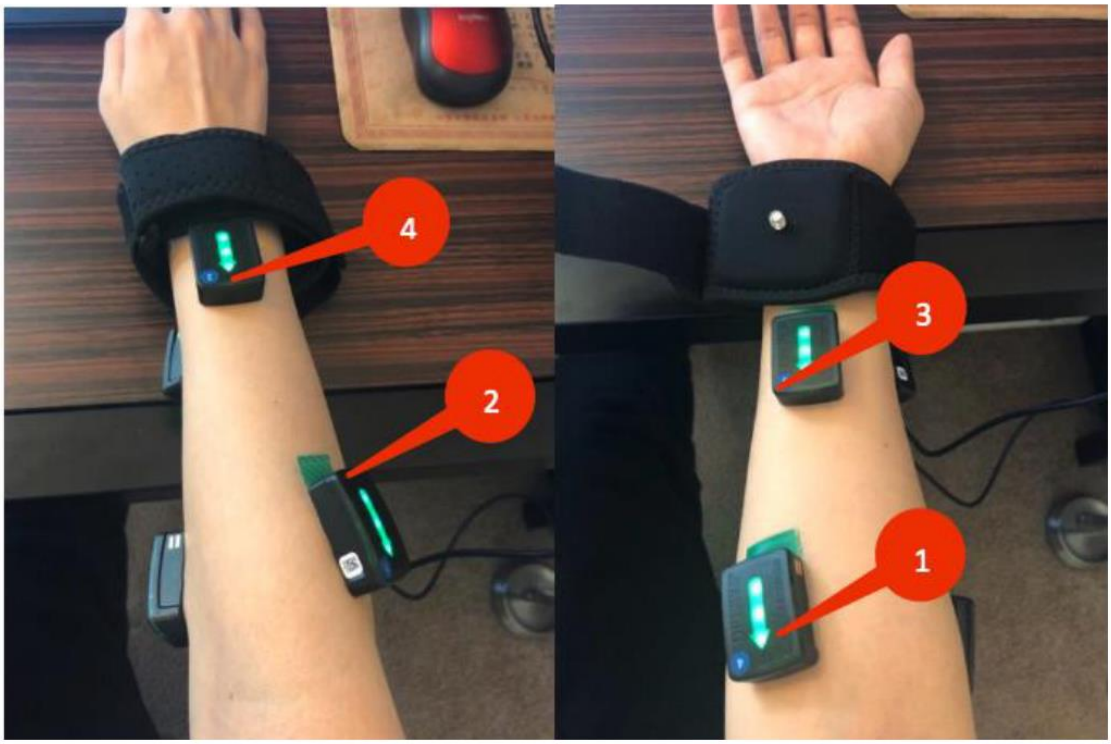


Figure 35. EMG sensor placement on flexor carpi radialis (1), extensor carpi radialis longus (2), flexor digitorum (3), and extensor digitorum (4)

5.3.3. VR Headset and Eye Tracker

The VIVE Pro Eye HMD provides built-in eye tracking features, making it the optimal VR hardware for this system (Figure 36). Eye tracking data is retrieved from the hardware using the SRanipal SDK at the maximum frequency of 120 Hz. The system automatically detects and logs blink rate and pupil diameter in millimeters to generate the necessary data to estimate the level of CW required when performing virtual ADLs.

Figure 36. HTC VIVE Pro Eye HMD

5.4. Task

5.4.1. CRT

The VR application features virtual versions of the CRT and SHAP. The virtual prosthesis must be in the open position and close enough to a clothespin to see the highlighted outline cue to pick up a clothespin in the VR environment (Figure 37). This yellow outline is a visual indicator that the virtual prosthesis is close enough to grip a clothespin. Visual cues for interaction are necessary features as there is no tactile feedback afforded by the VR environment. The participant

must then generate the command to close the hand to grip the clothespin. Clothespins in hand can be released by generating another open command. If a clothespin is released in a position in which it clamps onto any one of the bars of the base station, it will lock to that position until it is gripped again. If a clothespin is released anywhere other than onto one of the four bars, it will automatically respawn in the last valid position in which it was placed. If a clothespin is dropped immediately after removing it from the start position, it will return to the starting position.

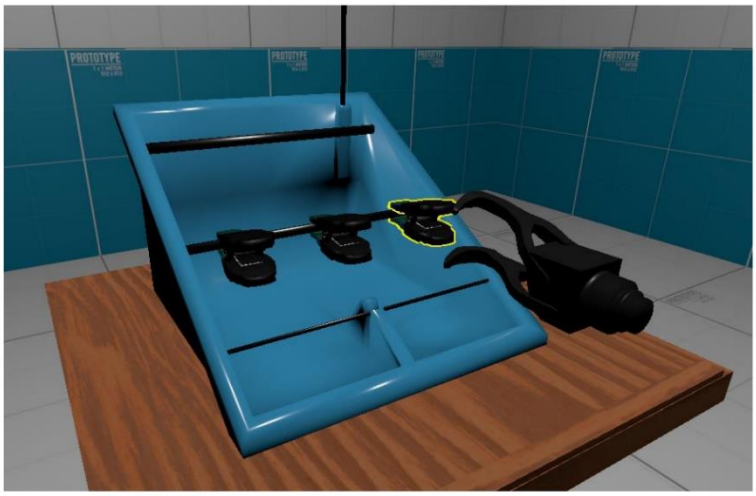


Figure 37. Highlighted outline of virtual clothespin. Serves as a visual cue to alert user of proximity to interactable object

Figure 38. A participant performing the virtual CRT task

5.4.2. SHAP

Similar to Experiment 1, the virtual SHAP door handle task shown in Figure 39 and 40 required participants to initiate a close action close enough to the door handle to grip it. Then, they must rotate the handle clockwise 90° via wrist supination, rotate it counterclockwise 90° back to the original position via wrist pronation, and finally perform an open action to release the handle. Meeting the above criteria (90 degrees) defined one successful rotation of the door handle. If the handle was released before these criteria were met or if the user moves the virtual prosthesis away from the door handle after gripping it, the system will recognize this as a drop or failed attempt and generate an appropriate log. Like the virtual CRT task, a highlighted outline of the door handle appears as a visual cue to indicate a close enough proximity to manipulate the handle (Figure 39).

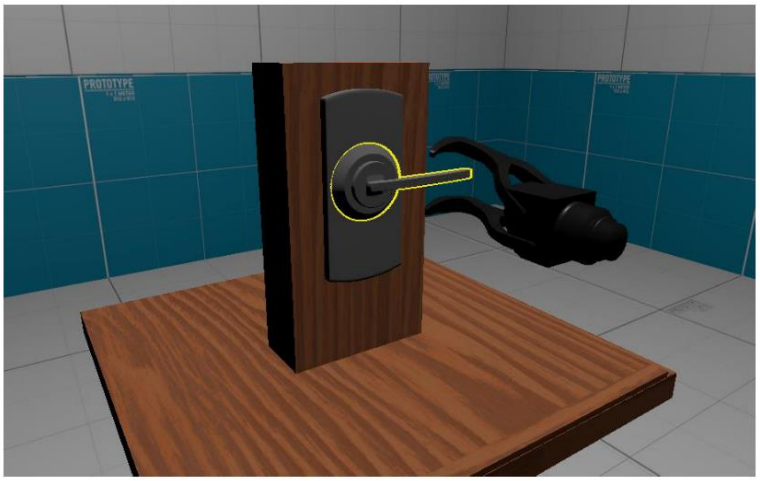


Figure 39. Highlighted outline of virtual door handle. Serves as a visual cue to alert user of proximity to interactable object

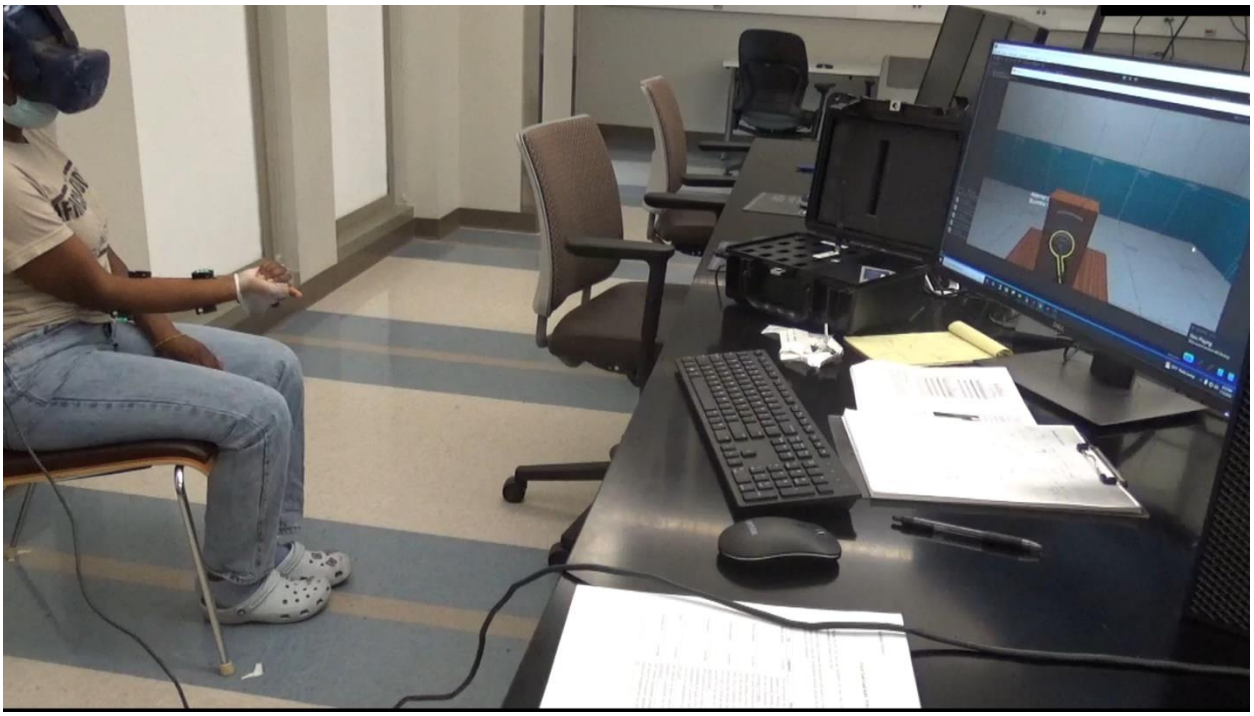


Figure 40. A participant is performing the virtual SHAP-Door handle task

5.5. Experiment Design and Variables

The experiment design and variables were similar to Experiment 1. Please see Sections 4.5, 4.6, and 4.7.

5.6. Procedure

The experiment procedure was similar with Experiment 1 (Section 4.8). The experimental setup included a VR headset and PC to run the experiment and collect raw data. For participants assigned to the DC group, participants needed to exert maximum strength for five seconds with dynamometer to measure MVC of each muscle. Using the MVC instead of MAV resulted in having more consistent EMG signals (Sabri et al., 2014). Therefore, we decided to collect MVC to define appropriate thresholds to activate or deactivate gestures in Experiment 2. During MVC measurement, the participant placed their feet on the dynamometer chassis as shown in Figure 41. The experimenter then adjusts the length of the chain so that the participant can still hold the chain while the muscles are relaxed. Three MVCs were collected and there was 1-minute break between each trial.

Figure 41. A participant is raising a handle to exert maximum strength

All participants were instructed to perform specific arm motions and to observe a feedback display (hand open, hand closed, wrist pronated, wrist supinated, and relaxed hand and wrist). The training and test sessions were the same as Experiment 1 (section 4.8). However, there was a need to develop a new training threshold for the VR setting. Based on the pilot test data, if the average task completion time of three sequential trials was within 17-27s for the PR and within 44-54s for the DC in the virtual CRT task, the participant passed the training and proceeded to the experimental trials. During the training and actual trials, participants were seated on a chair without an armrest to avoid any interference with upper limb motions.

5.7. Hypotheses

The following hypotheses (H) were formulated for this study. Since both human subject data and benchmark model do not generate all usability dimensions, hypotheses were generated to enable comparisons between the HPM-UP and either human data or the benchmark model.

- Learnability: The results of HPM-UP learnability dimension would be similar to the human-subject data (H1)
- Error rate: Use of the PR configuration would lead to the lower error rate than DC (H2) (White et al., 2017; Zahabi et al., 2019b; Zhang et al., 2016b)
- Memorability: The results of HPM-UP memorability dimension would be similar to the benchmark model (H3)
- Efficiency: The results of HPM-UP efficiency dimension would be similar to the human subject data (H4-1) but would be significantly different from the benchmark model results (H4-2).

- Satisfaction: The results of HPM-UP satisfaction dimension would be similar to the human-subject data (H5)

5.8. Data Collection and Analysis

Data collection and analysis were similar to Experiment 1. Pupil size and the number of blinks was captured at a frequency of 200 Hz from HTC VIVE headset. The VR system also automatically logged the task performance of each task. The same non-parametric and post-hoc analysis methods were applied in Experiment 2.

5.9. Results

5.9.1. Hypothesis Test Results

The data were collected from twenty able-bodied participants (10 participants for DC and 10 for PR), and the results were compared with outcomes generated from the benchmark model and the HPM-UP. Table 17 illustrates the descriptive statistics results regarding the comparison among the data from human subject experiment (Experiment 2), the HPM-UP, and the benchmark model. All the hypotheses were supported. The PR configuration led to a significantly lower error rate than the DC mode ($H(1)=5.47$, $p < .05$). Table 18 summarized the result of the hypothesis tests.

Table 17. Descriptive statistics from Experiment 2 (mean (sd))

Factors (definition)		Human subject data		HPM-UP		Benchmark model	
		DC	**PR**	**DC**	**PR**	**DC**	**PR**
Learnability		5.5 (2.33)	3.6 (0.49)	3.4 (0.92)	4.3 (1.10)	N/A	
Error rate		N/A		0.25 (0.06)	0.30 (0.07)	N/A	
Memorability		N/A		3.48 (0.13)	0	3.60 (0.00)	0
Efficiency	CRT	26.27 (10.19)	9.05 (3.06)	17.83 (1.67)	8.42 (0.94)	13.3 (0.00)	5.6 (0.00)
	SHAP	10.65 (3.19)	7.97 (2.36)	9.85 (0.52)	7.30 (0.38)	7.5 (0.00)	5.1 (0.00)
Satisfaction		0.68 (0.16)	0.74 (0.14)	0.74 (0.12)	0.75 (0.11)	N/A	
Accuracy of Cognitive Workload Classification (%)		N/A		65.00	80.00	N/A	

Table 18. Summary hypothesis test results (Experiment 2)

Hypothesis ID	Hypothesis	Test Result	Test statistics, p-value, effect size
H1 (Learnability)	The results of HPM-UP learnability dimension would be similar to the human data.	Supported	$W = 236$, $p = .31$, $r = .16$
H2 (Error rate)	Use of the PR configuration would lead to the lower error rate than DC.	Supported	$H(1) = 5.47$, $p = .02$, $r = .52$
H3 (Memorability)	The results of HPM-UP memorability dimension would be similar to the benchmark model.	Supported	$W = 65$, $p = .23$, $r = .28$
H4 (Efficiency)	(H4-1) The results of HPM-UP efficiency dimension would be similar to the human subject data.	Supported	$Z = -0.68$, $p = .50$, $r = .12$
	(H4-2) There would be a significant difference between the HPM-UP efficiency dimension results and the benchmark model results.	Supported	$Z = -2.64$, $p = .02$, $r = .33$
H5 (Satisfaction)	The results of HPM-UP satisfaction dimension would be similar to the human data.	Supported	$W=258$, $p = .12$, $r = .25$

5.9.2. Cognitive Workload Classification

The trained model NB model from the Experiment 1 was used to classify CW of the data collected from Experiment 2. The model exhibited 73.83% average classification performance (*SD*

= 6.96%) across 15 random seeds (Colas et al., 2019). Similar to Experiment 1 results, the NB model exhibited a decent performance across multiple runs.

5.9.3. Discussion

Experiment 2 results revealed that the HPM-UP can also be used in to assess the usability of virtual prosthetic controls. Although the hypothesis test results were similar to Experiment 1 findings, the descriptive results in Experiment 2 were better than that of Experiment 1, especially for the learnability dimension. In Experiment 2, participants reached the training threshold much faster than Experiment 1. This might be because the VR could improve learning for participants. Previous studies revealed the effectiveness of using VR in different domains including: prosthetic device control (Lambrecht et al., 2011), prosthetic rehabilitation training (Dhawan et al., 2019), fire response training (Sankaranarayanan et al., 2018), surgery (Seymour et al., 2002), and rescue team training (Katz et al., 2020). Another reason is the effect of fatigue. Unlike Experiment 1, participants did not use the physical prosthetic device in Experiment 2 and therefore, were much faster in mastering the control schemes.

The usability of virtual prosthetic device in Experiment 2 was superior to the physical prosthetic device used in Experiment 1 based on the outcomes of the HPM-UP. For example, the efficiency was high when performing both the CRT and SHAP tasks. Even with the DC configuration, participants' performance in the SHAP task substantially improved as compared to the performance in Experiment 1. The relationship among learnability, error rate, and efficiency dimensions existed in both Experiment 1 and 2 results. For example, in Experiment 2, high learnability led to having high efficiency in performing the tasks and eventually led to high satisfaction levels.

The CW classification model developed from Experiment 1 was validated with data from the Experiment 2. The trained ML model in HPM-UP could accurately predict the CW in use of the virtual prosthetic device. This is an indication of model's generalizability as the model has learned meaningful patterns and relationships in the training data. It also means the hyperparameter setting and the ML modeling structure (e.g., having feature selectors or cross validation) was effective in finding the best model. However, future studies should validate the model with other ADL tasks and experiment settings.

5.10. Contributions of Experiment 2

HPM-UP is the first human performance model that can estimate usability of prosthetic devices and has been validated with the data from both physical and virtual environments. Results of Experiment 2 also supported the merits of using VE for training as suggested by prior studies (Hargrove et al., 2018). Unlike previous prosthetic device studies that used 2D-displays (Deeny et al., 2014a; Deeny et al., 2014b; Rezazadeh et al., 2012; Rezazadeh et al., 2011), this study used an immersive VE (i.e., virtual reality headset). VE could be used for testing the capability of human, through practice, to acquire new sensorimotor mappings to adapt to novel kinematics or dynamics as well as to learn how to manipulate a device (Park and Zahabi, 2022a).

The ML algorithm can provide valuable inputs regarding the CW of prosthetic devices to designers, clinicians, and researchers. First, the algorithm exhibited its generalizability to different circumstances (i.e., physical or virtual environments). Thus, designers or clinicians can run the model only with EMG sensors to predict the level of CW with a reasonable accuracy. This could reduce experimental cost which has been normally accompanied by human-subject experiments (Park et al., 2022). From the users' point of view, prosthetic device users can also easily test the

level of CW with the immersive VE. No prior ML algorithm has been validated in both physical and VE. Thus, the developed algorithm can be a starting point for future research. For example, the model can be tested with other tasks or additional datasets to improve its versatility. Other researchers could improve the accuracy of the classification with modifications on the model or with additional data. Third, the ML model can improve prosthetic device control and experience if the model is expanded to provide CW in real-time. For instance, prosthetic device users can be notified with the current level of CW and can adjust the device control strategy. Clinicians can also use the model to estimate the level of CW during a task. Tracking the fluctuation of the level of CW is possible and the trajectory of the change of CW can also be plotted on the screen in real-time to inform clinicians which part of the task is demanding for amputee patients.

5.11. Practical Implications of HPM-UP

There are several practical implications of HPM-UP for clinicians, device developers, or researchers in the cognitive modeling domain. The model can be used by clinicians and device developers using the GUI and with mouse-clicks. This feature can be especially useful for those without any knowledge of programming. In addition, in the “Results” tab, a guideline table is provided, which can provide practical recommendations regarding the range of each usability dimension score. With this table, clinicians can determine whether to recommend a certain prosthesis to a patient. Finally, under the “Help” tab of HPM-UP, several tutorial videos are provided on how to use the model. The model is available on Github (https://github.com/hsilab/hpmup/tree/master) and other researchers can modify or update it. Once they have added all the data and created the scenario, they could assess the usability and CW of any prosthetic device.

Whenever clinicians have new amputee patients and before recommending any prosthetic device to patients, they can test or predict which device could be the best in terms of the usability and CW for the amputee. That is, HPM-UP could reduce the work of clinicians to find, test, analyze, and recommend a prosthetic device. Once clinicians collect the input parameters (e.g., first impression) for each prosthetic device from the patients, they could run the model and see the predicted usability and CW values. Then, based on the results and the guideline table, they can recommend the best device to amputees.

For designers of prosthetic devices, HPM-UP could be a quick and practical guidance for a prototype-level usability and CW assessment. Once they have defined tasks and concept for the prosthetic devices, they could predict the human performance of the device at the early stages of the design process. They can also change the input parameters based on the characteristics of the target group. Based on the results, they can make changes to the device configurations. Designers could do this iterative process at the early stage of the design process to adjust the usability and CW of prosthetic devices to improve human use.

6. CONCLUSION

Previous studies for measuring performance of prosthetic device users relied on human subject experiments and subjective evaluations. Especially for cognitive workload or usability assessments, subjective evaluation methods were heavily employed. While these methods could provide useful outcomes, early estimation of usability and CW is critical to reduce future device rejection due to high CW or usability issues.

This research advanced the fundamental knowledge of estimating usability and CW of upper-limb prostheses in EMG-based human-machine interfaces. Established methodologies, theories, and experimental methods were used to formulate the equations to quantify usability and cognitive workload of upper-limb prosthetic devices. This research not only quantified each usability dimension (learnability, errors, memorability, efficiency, and satisfaction) but also connected them in a computational way.

The HPM-UP model was developed with top-down and bottom-up approaches. Especially for estimating the level of CW, machine learning algorithms were trained, tested, and incorporated in HPM-UP. The model was validated by experimental studies. The findings of Experiment 2 also supported previous studies that argued potential merits of having virtual environment instead of physical environment to train prosthetic device users. HPM-UP can be run using a GUI and does not require hard-coding. It is the first HPM that was developed in R Shiny package format and released to GitHub to be used by other researchers, designers, or clinicians. HPM-UP provides the capability to predict human performance of prostheses at the early stage of the design process. Clinicians can test and analyze the human performance of several commercial prostheses to find and recommend a best device(s) for the patients.

6.1. Limitations and Future Research

There are several aspects of this study that may limit the generalizability of findings. First, HPM-UP has some free parameters, especially in learnability and satisfaction dimensions. The reason to include them in the models was to personalize the outcome of the model to improve model performance. The initial/default values included in the current version of HPM-UP was calculated based on the pilot tests in this research. Although this approach has been used in other HPMs such as QN-MHP (e.g., preliminary estimates of the perceptual memory access time) (Feyen, 2003), MHP (Card et al., 1986a), or ACT-R (Bothell, 2020), the outcomes of the model depend on these values.

To quantify qualitative dimensions such as learnability and satisfaction, some assumptions have been made. For example, *SDCQ* was calculated based on the average of the responses to questions Q3 (easiness in adjusting the device (fixing, fastening)), Q6 (easiness of using the device), and Q8 (effectiveness of using the device (the degree to which the device meets a user's needs) of the USE questionnaire. *FI* was calculated from the difference between the SDCQ and participant's training performance. Furthermore, HPM-UP estimates the immediate learnability and satisfaction after using prostheses, which is different from the retention effect or long-term/sustained satisfaction, which is the original concept behind the ECT. Future studies should validate these assumptions with additional experiments and considering long-term satisfaction with prosthetic devices.

The decision to work with an able-bodied population was made due to the limited number of trans-radial amputees in the surrounding area. In addition, since most patients currently use devices with DC modes (commonly used in myoelectric control), recruiting such patients could have produced a bias in their performance. There is a need to validate the results of this research

with amputee patients. In addition, the dataset used to train ML algorithms and to find the best model to classify cognitive workload in the HPM-UP was small. Future studies with larger datasets are necessary to validate the accuracy of the model.

Although with the GUI analysts can develop scenarios only with mouse clicks, they need to have basic knowledge of human performance modeling. If not, they need to have external support to validate the codes for accuracy. In addition, clinicians might need help to set up the software. That is, to run HPM-UP, they should install R, R-studio, and several packages. This could be challenging for clinicians who do not have backgrounds in statistics. Although the installation guidelines are available on GitHub, the process could still be challenging for some individuals.

The HPM-UP introduced in this dissertation was mainly designed for assessing upper-limb prosthetic devices while performing CRT and SHAP tasks, which are widely used testbeds of ADLs. Considering the generalizability of the logic behind HPM-UP (CPM-GOMS and ACT-R), it is possible to expand the scope and application of HPM-UP. Future research should extend HPM-UP to other tasks in the SHAP testbed such as food cutting, glass jug pouring, or lifting a tray. After validating or updating HPM-UP for other tasks, it is also desirable to apply the model to naturalistic ADL tasks such as eating or dressing at home.

HPM-UP simulated the "transradial (below elbow) upper-limb" amputee behavior, although there exists a wide range of amputations. Possible target populations are related to trans-humeral (above elbow), forequarter amputation, shoulder disarticulation, or wrist disarticulation. HPM-UP can be further extended to model cognitive workload of lower limb prostheses. For example, transtibial (below the knee), transfemoral (above knee), foot amputations, knee disarticulation, and hip disarticulation can be modeled in the extended HPM-UP.

Although HPM-UP provides estimates of device usability and CW, it cannot guarantee the fitness or feeling of embodiment of a prosthesis to amputees. Therefore, future studies could consider incorporating the fitness into the model, as it is one of the critical factors in acceptance of prostheses (Hagberg et al., 2004; Stratford, 2001). To enable this, researchers can collect data with established questionnaires such as QUEST 2.0, CSD-OPUS, questionnaires for persons with a transfemoral amputation (Q-TFA) (Hagberg et al., 2004), or upper extremity functional index (UEFI) as a ground-truth after amputees perform some ADLs. Then, some of the factors in these questionnaires can be added as parameters to the equation to calculate the satisfaction dimension of HPM-UP. The model's outcome can then be compared with the ground-truth to for validation.

Lastly, the current version of HPM-UP only estimates usability and CW. While validating the model under various conditions mentioned above, future studies should enhance HPM-UP with the capability to estimate physical workload. To enable this, there is a need to collect anthropometry data and EMG signals and analyze the pattern of signals to identify physically demanding activities. Then, the task scenario should be compared with the analyzed EMG pattern. In this process, we could match each motor operator with EMG signals, which could lead to an estimate of muscle activities or physical workload during the task. Studying ACT-Phi (physiological measurement) (Dancy, 2018; Dancy and Kaulakis, 2013; Dancy et al., 2015) or ACT-R/F (Fatigue) (Gunzelmann et al., 2009; Gunzelmann and Gluck, 2008) which are some of the family members of ACT-R could be good starting points for this integration.

REFERENCES

Afonja, T. (2017). Accuracy paradox. *Towards Data Science.*

Aggarwal, C. C. (2018). *Machine learning for text* (Vol. 848): Springer.

Albert, W., & Tullis, T. (2013). *Measuring the user experience: collecting, analyzing, and presenting usability metrics*: Newnes.

Altmann, E. M., & Schunn, C. D. (2019). *Integrating decay and interference: A new look at an old interaction.* Paper presented at the Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society.

Amputee Coalition. (2021). Limb Loss & Limb Difference in the U.S.

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human–Computer Interaction, 12*(4), 439-462.

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive psychology, 30*(3), 221-256.

Appel, T., Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., Kasneci, E., Gerjets, P., & Assoc Comp, M. (2019, Oct 14-18). *Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures.* Paper presented at the 21st ACM International Conference on Multimodal Interaction (ICMI), Suzhou, PEOPLES R CHINA.

Arenas, J. A. (2015). Evaluation of a Novel Myoelectric Training Device.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*(3), 258.

Bark, K., Hyman, E., Tan, F., Cha, E., Jax, S. A., Buxbaum, L. J., & Kuchenbecker, K. J. (2014). Effects of vibrotactile feedback on human learning of arm motions. *IEEE Transactions on neural systems and rehabilitation engineering, 23*(1), 51-63.

Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector machines. *Imbalanced learning: Foundations, algorithms, and applications*, 83-99.

Beauvois, J.-L. (1982). Théories implicites de la personnalité, évaluation et reproduction idéologique. *L'Année psychologique, 82*(2), 513-536.

Berntsson, L. (2019). Evaluation of cognitive workload using EEG: Investigation of how sensory feedback improves function of osseo-neuromuscular upper limb prostheses. In.

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS quarterly*, 351-370.

Biddiss, E., Beaton, D., & Chau, T. (2007). Consumer design priorities for upper limb prosthetics. *Disability and Rehabilitation: Assistive Technology, 2*(6), 346-357.

Bloomer, C., Wang, S., & Kontson, K. (2020). Kinematic analysis of motor learning in upper limb body-powered bypass prosthesis training. *Plos One, 15*(1), e0226563.

Bothell, D. (2017). ACT-R 7 reference manual. *Available at act-r. psy. cmu. edu/wordpress/wpcontent/themes/ACT-R/actr7/reference-manual. pdf*.

Bothell, D. (2020). *ACT-R 7.21+ reference manual*. Retrieved from

Bowker, J. (2004). The art of prosthesis prescription. *Smith DG, Michael JW, Bowker JH. American Academy of Orthopaedic Surgeons. Atlas of Amputations and Limb Deficienciens Surgical, Prosthetic and Rehabilitation Principles. 3st ed. Rosemont IL: Bone and Joint Decade*, 739-744.

Braarud, P. O., Bodal, T., Hulsund, J. E., Louka, M. N., Nihlwing, C., Nystad, E., Svengren, H., & Wingstedt, E. (2021). An Investigation of Speech Features, Plant System Alarms, and Operator-System Interaction for the Classification of Operator Cognitive Workload During Dynamic Work. *Human Factors, 63*(5), 736-756. doi:10.1177/0018720820961730

Bravini, E., Franchignoni, F., Ferriero, G., Giordano, A., Bakhsh, H., Sartorio, F., & Vercelli, S. (2014). Validation of the Italian version of the Client Satisfaction with Device module of the Orthotics and Prosthetics Users' Survey. *Disability and health journal, 7*(4), 442-447.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry, 189*(194), 4-7.

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology, 42*(3), 361-377.

Buchner, A., Erdfelder, E., Faul, F., & Lang, A. (2017). G* Power 3.1 manual. *Düsseldorf, Germany: Heinrich-Heine-Universitat Dusseldorf.*

Butt, A. H., Rovini, E., Dolciotti, C., De Petris, G., Bongioanni, P., Carboncini, M., & Cavallo, F. (2018). Objective and automatic classification of Parkinson disease with Leap Motion controller. *BioMedical Engineering Online, 17*(1), 1-21.

Cain, B. (2007). *A review of the mental workload literature*. Retrieved from

Camm, J. D. (1985). A note on learning curve parameters. *Decision sciences, 16*(3), 325-327.

Card, MORAN, & Newell. (1986a). The model human processor- An engineering model of human performance. *Handbook of perception and human performance., 2*(45–1).

Card, Newell, A., & Moran, T. P. (1983). The Psychology of Human-Computer Interaction. In: L. Erlbaum Associates Inc.

Card, S., MORAN, T., & Newell, A. (1986b). The model human processor- An engineering model of human performance. *Handbook of perception and human performance., 2*(45–1).

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.

Cardona, G., & Quevedo, N. (2014). Blinking and driving: the influence of saccades and cognitive workload. *Current eye research, 39*(3), 239-244.

Carlson, T., Tonin, L., Perdikis, S., Leeb, R., & Millán, J. d. R. (2013). *A hybrid BCI for enhanced control of a telepresence robot.* Paper presented at the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: a systematic review. *Applied Ergonomics, 74*, 221-232.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of economic behavior & organization, 81*(1), 1-8.

Childress, D. S. (1980). Closed-loop control in prosthetic systems: historical perspective. *Annals of biomedical engineering, 8*(4), 293-303.

Clifford, M. M., & Walster, E. (1973). The effect of physical attractiveness on teacher expectations. *Sociology of education*, 248-258.

Colas, C., Sigaud, O., & Oudeyer, P.-Y. (2019). A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms. *arXiv preprint arXiv:1904.06979*.

Connan, M., Ruiz Ramírez, E., Vodermayer, B., & Castellini, C. (2016). Assessment of a wearable force-and electromyography device and comparison of the related signals for myocontrol. *Frontiers in neurorobotics, 10*, 17.

Cordella, F., Ciancio, A. L., Sacchetti, R., Davalli, A., Cutti, A. G., Guglielmelli, E., & Zollo, L. (2016). Literature Review on Needs of Upper Limb Prosthesis Users. *Frontiers in Neuroscience, 10*, 209-209. doi:10.3389/fnins.2016.00209

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current directions in psychological science, 19*(1), 51-57.

Crea, S., Edin, B. B., Knaepen, K., Meeusen, R., & Vitiello, N. (2017). Time-Discrete Vibrotactile Feedback Contributes to Improved Gait Symmetry in Patients With Lower Limb Amputations: Case Series. *Physical Therapy, 97*(2), 198-207. doi:10.2522/ptj.20150441

Dancy, C. (2018). Towards a physio-cognitive model of slow-breathing.

Dancy, C. L., & Kaulakis, R. (2013). *Towards adding bottom-up homeostatic affect to ACT-R.* Paper presented at the proceedings of the 12th International Conference on Cognitive Modeling.

Dancy, C. L., Ritter, F. E., & Gunzelmann, G. (2015). *Two ways to model the effects of sleep fatigue on cognition.* Paper presented at the 13th International Conference on Cognitive Modeling, ICCM 2015, April 9, 2015 - April 11, 2015, Groningen, Netherlands.

Davidson, M. L. (2017). *Development of a Novel Prosthetic Wrist Device Incorporating the Dart Thrower's Motion*: University of Colorado at Denver.

De Waard, D., & Brookhuis, K. (1996). The measurement of drivers' mental workload.

Deeny, S., Barstead, M., Chicoine, C., Hargrove, L., Parrish, T., & Jayaraman, A. (2014a). EEG as an outcome measure for cognitive workload during prosthetic use.

Deeny, S., Chicoine, C., Hargrove, L., Parrish, T., & Jayaraman, A. (2014b). A Simple ERP Method for Quantitative Analysis of Cognitive Workload in Myoelectric Prosthesis Control and Human-Machine Interaction. *Plos One, 9*(11). doi:10.1371/journal.pone.0112091

Dehban, A., Menhaj, M., & Sajedin, A. (2015). Neuro-ACT cognitive architecture applications in modeling driver's steering behavior in turns. *AUT Journal of Modeling and Simulation, 47*(2), 21-29.

Dehban, A., Sajedin, A., & Menhaj, M. B. (2016). *A cognitive based driver's steering behavior modeling.* Paper presented at the 2016 4th International Conference on Control, Instrumentation, and Automation (ICCIA).

Demers, L., Weiss-Lambrou, R., & Ska, B. (2002). The Quebec User Evaluation of Satisfaction with Assistive Technology (QUEST 2.0): an overview and recent progress. *Technology and Disability, 14*(3), 101-105.

Dhawan, D., Barlow, M., & Lakshika, E. (2019). *Prosthetic rehabilitation training in virtual reality.* Paper presented at the 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH).

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning, 40*(2), 139-157.

Din, A. (2015). Usable Security using GOMS: A Study to Evaluate and Compare the Usability of User Accounts on E-Government Websites.

Ding, Y., Cao, Y. Q., Duffy, V. G., Wang, Y., & Zhang, X. F. (2020). Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning. *Ergonomics*, *63*(7), 896-908. doi:10.1080/00140139.2020.1759699

Dix, A., Finlay, J., Abowd, G. D., & Beale, R. (2000). Human-computer interaction. *Harlow ua*.

Donges, N. (2021). Random Forest Algorithm: A Complete Guide. Retrieved from https://builtin.com/data-science/random-forest-algorithm

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*(3), 241-252.

Duysens, J., Potocanac, Z., Hegeman, J., Verschueren, S., & McFadyen, B. J. (2012). Split-second decisions on a split belt: does simulated limping affect obstacle avoidance? *Experimental brain research*, *223*(1), 33-42. doi:10.1007/s00221-012-3238-x

Engdahl, S. M., Christie, B. P., Kelly, B., Davis, A., Chestek, C. A., & Gates, D. H. (2015). Surveying the interest of individuals with upper limb loss in novel prosthetic control techniques. *Journal of neuroengineering and rehabilitation*, *12*(1), 53.

Estes, S. (2015). The workload curve: Subjective mental workload. *Human factors*, *57*(7), 1174-1187.

Estes, S. (2017). Cogulator. *The MITRE Corporation*.

Estes, S. (2021). Cogulator - A Cognitive Modeling Calculator. Retrieved from https://github.com/Cogulator/Cogulator

Estes, S., & Masalonis, A. J. (2003). *I See What You're Thinking: Using Cognitive Models to Represent Working Memory Usage for Traffic Flow Management Decision Support Prototypes.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods, 41*(4), 1149-1160.

Ferreira, A. J., & Figueiredo, M. A. (2012). Efficient feature selection filters for high-dimensional data. *Pattern recognition letters, 33*(13), 1794-1804.

Feyen, R. G. (2003). Modeling human performance using the queuing network-model human processor (QN-MHP).

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology* (Vol. 23, pp. 1-74): Elsevier.

Fogarty, C., & Stern, J. A. (1989). Eye movements and blinks: their relationship to higher cognitive processes. *International Journal of Psychophysiology, 8*(1), 35-42.

Gaskins, C., Kontson, K., Shaw, E. P., Shuggi, I. M., Ayoub, M. J., Rietschel, J. C., Miller, M. W., & Gentili, R. (2018). Mental Workload Assessment During Simulated Upper Extremity Prosthetic Performance. *Archives of physical medicine and rehabilitation, 99*(10), e33.

Geurts, & Mulder, T. W. (1994). Attention Demands in Balance Recovery following Lower Limb Amputation. *Journal of Motor Behavior, 26*(2), 162-170. doi:10.1080/00222895.1994.9941670

Geurts, Mulder, T. W., Nienhuis, B., & Rijken, R. (1991). Dual-task assessment of reorganization of postural control in persons with lower limb amputation. *Arch Phys Med Rehabil, 72*(13), 1059-1064.

Gil, G. H. (2010). An Accessible Cognitive Modeling Tool for Evaluation of Human-Automation Interaction in the Systems Design Process.

Gonzalez, J., Soma, H., Sekine, M., & Yu, W. (2012a). Psycho-physiological assessment of a prosthetic hand sensory feedback system based on an auditory display: a preliminary study. *Journal of neuroengineering and rehabilitation, 9*(1), 33.

Gonzalez, J., Suzuki, H., Natsumi, N., Sekine, M., & Yu, W. (2012b). *Auditory display as a prosthetic hand sensory feedback for reaching and grasping tasks.* Paper presented at the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*: MIT press.

Götze, T., Gürtler, M., & Witowski, E. (2020a). How to Deal with Small Data Sets in Machine Learning: An Analysis on the CAT Bond Market. *Available at SSRN 3528082*.

Götze, T., Gürtler, M., & Witowski, E. (2020b). Improving CAT bond pricing models via machine learning. *Journal of Asset Management, 21*(5), 428-446.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Gronier, G. (2016). Measuring the First Impression: Testing the Validity of the 5 Second Test. *Journal of Usability Studies, 12*(1).

Gunzelmann, G., Byrne, M. D., Gluck, K. A., & Moore Jr, L. R. (2009). Using computational cognitive modeling to predict dual-task performance with sleep deprivation. *Human Factors, 51*(2), 251-260. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/0018720809334592

Gunzelmann, G., & Gluck, K. A. (2008). *Approaches to modeling the effects of fatigue on cognitive performance.* Paper presented at the 17th Conference on Behavior Representation in Modeling and Simulation, BRIMS, April 14, 2008 - April 17, 2008, Providence, RI, United states.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1), 389-422.

Hagberg, K., Brånemark, R., & Hägg, O. (2004). Questionnaire for Persons with a Transfemoral Amputation (Q-TFA): initial validity and reliability of a new outcome measure. *Journal of Rehabilitation Research & Development*, *41*(5).

Hargrove, L., Losier, Y., Lock, B., Englehart, K., & Hudgins, B. (2007). *A real-time pattern recognition based myoelectric control usability study implemented in a virtual environment.* Paper presented at the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Hargrove, L., Miller, L., Turner, K., & Kuiken, T. (2018). Control within a virtual environment is correlated to functional outcomes when using a physical prosthesis. *Journal of neuroengineering and rehabilitation*, *15*. doi:10.1186/s12984-018-0402-y

Hargrove, L. J., Miller, L. A., Turner, K., & Kuiken, T. A. (2017). Myoelectric pattern recognition outperforms direct control for transhumeral amputees with targeted muscle reinnervation: A randomized clinical trial. *Scientific reports*, *7*(1), 13840.

Hart, S. G. (2006). *NASA-task load index (NASA-TLX); 20 years later.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183): Elsevier.

Hassenzahl, M., & Monk, A. (2010). The inference of perceived usability from beauty. *Human–Computer Interaction, 25*(3), 235-260.

Heller, B. W., Datta, D., & Howitt, J. (2000). A pilot study comparing the cognitive demand of walking for transfemoral amputees using the Intelligent Prosthesis with that using conventionally damped knees. *Clinical rehabilitation, 14*(5), 518-522. doi:10.1191/0269215500cr345oa

Herberts, P., & Körner, L. (1979). Ideas on sensory feedback in hand prostheses. *Prosthetics and Orthotics International, 3*(3), 157-162. Retrieved from https://journals.sagepub.com/doi/pdf/10.3109/03093647909103104

Hofstad, C. J., Weerdesteyn, V., van der Linde, H., Nienhuis, B., Geurts, A. C., & Duysens, J. (2009). Evidence for bilaterally delayed and decreased obstacle avoidance responses while walking with a lower limb prosthesis. *Clinical Neurophysiology, 120*(5), 1009-1015. doi:https://doi.org/10.1016/j.clinph.2009.03.003

Huang, Y., & Li, L. (2011). *Naive Bayes classification algorithm based on small sample set.* Paper presented at the 2011 IEEE International conference on cloud computing and intelligence systems.

ISO. (2019). Ergonomics of human-system interaction-Part 11: Usability: Definitions and concepts (ISO 9241-11: 2018) Irish Standard ISO. *Recuperado de https://infostore. saiglobal. com/preview/is/en/2018/is eniso9241-11-2018. pdf.*

Jadhav, S. D., & Channe, H. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR), 5*(1), 1842-1845.

Jain, K. (2021). How to Improve Naive Bayes? Retrieved from https://medium.com/analytics-vidhya/how-to-improve-naive-bayes-9fa698e14cba

Jang, C. H., Yang, H. S., Yang, H. E., Lee, S. Y., Kwon, J. W., Yun, B. D., Choi, J. Y., Kim, S. N., & Jeong, H. W. (2011). A survey on activities of daily living and occupations of upper extremity amputees. *Annals of rehabilitation medicine, 35*(6), 907-921. doi:10.5535/arm.2011.35.6.907

Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). *Facing imbalanced data--recommendations for the use of performance metrics.* Paper presented at the 2013 Humaine association conference on affective computing and intelligent interaction.

Jeong, H., & Liu, Y. (2018). *Cognitive Modeling of Remote-manual and Voice Controls for In-vehicle Human-automation Systems.* Paper presented at the 13th Annual ACM/IEEE International Conference on Human Robot Interaction, HRI 2018, March 5, 2018 - March 8, 2018, Chicago, IL, United states.

John, B. E. (1990). *Extensions of GOMS Analyses to Expert Performance Requiring Perception of Dynamic Visual and Auditory.* Paper presented at the Empowering People: CHI'90 Conference Proceedings [on] Human Factors in Computing Systems: Seatle, Washington, April 1-5, 1990.

John, B. E. (2005). *Cognitive human performance modeling by demonstration.* Paper presented at the 49th Annual Meeting of the Human Factors and Ergonomics Society, HFES 2005, September 26, 2005 - September 30, 2005, Orlando, FL, United states.

John, B. E., & Suzuki, S. (2009). *Toward cognitive modeling for predicting usability.* Paper presented at the 13th International Conference on Human-Computer Interaction, HCI International 2009, July 19, 2009 - July 24, 2009, San Diego, CA, United states.

Joyce, A. (2019). How to Measure Learnability of a User Interface. Retrieved from https://www.nngroup.com/articles/measure-learnability/

Kaczorowska, M., Plechawska-Wojcik, M., & Tokovarov, M. (2021). Interpretable Machine Learning Models for Three-Way Classification of Cognitive Workload Levels for Eye-Tracking Features. *Brain Sciences*, *11*(2), 22. doi:10.3390/brainsci11020210

Kannenberg, A., & Zacharias, B. (2014). *Difficulty of performing activities of daily living with the Michelangelo Multigrip and traditional myoelectric hands.* Paper presented at the American Academy of Orthotists & Prosthetists 40th Academy Annual Meeting &Scientific Symposium, FPTH14.

Katz, D., Shah, R., Kim, E., Park, C., Shah, A., Levine, A., & Burnett, G. (2020). Utilization of a voice-based virtual reality advanced cardiac life support team leader refresher: prospective observational study. *Journal of Medical Internet Research*, *22*(3), e17425.

Khalid, U. (2014). Development And Human Performance Evaluation Of Control Modes Of An Exo-Skeletal Assistive Robotic Arm (esara).

Kieras, D. (1994). *GOMS Modeling of User Interfaces Using NGOMSL. Tutorial Notes.* Paper presented at the CHI Conference on Human Factors in Computing Systems, Boston, MA, April.

Kieras, D. (2005). A survey of cognitive architectures. In.

Kieras, D. (2006). *A Guide to GOMS Model Usability Evaluation using GOMSL and GLEAN4*.

Kieras, D. E. (1988). Towards a practical GOMS model methodology for user interface design. In *Handbook of human-computer interaction* (pp. 135-157): Elsevier.

Kieras, D. E., & Meyer, D. E. (1995). *Predicting human performance in dual-task tracking and decision making with computational models using the EPIC architecture.* Paper presented at the Proceedings of the First International Symposium on Command and Control Research and Technology, National Defense University, June. Washington, DC: National Defense University.

Kim, H., & Fesenmaier, D. R. (2008). Persuasive design of destination web sites: An analysis of first impression. *Journal of Travel research, 47*(1), 3-13.

Kim, J.-W. (2005). Introduction to human computer interaction. *Ahn graphics*.

Knaepen, K., Marusic, U., Crea, S., Guerrero, C. D. R., Vitiello, N., Pattyn, N., Mairesse, O., Lefeber, D., & Meeusen, R. (2015). Psychophysiological response to cognitive workload during symmetrical, asymmetrical and dual-task walking. *Human movement science, 40*, 248-263.

Kotseruba, I., & Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Core cognitive abilities and practical applications. *arXiv preprint arXiv:1610.08602*.

Krewer, C., Müller, F., Husemann, B., Heller, S., Quintern, J., & Koenig, E. (2007). The influence of different Lokomat walking conditions on the energy expenditure of hemiparetic patients and healthy subjects. *Gait & posture, 26*(3), 372-377. doi:https://doi.org/10.1016/j.gaitpost.2006.10.003

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association, 47*(260), 583-621.

Kuiken, T., Miller, L., & Turner, K. (2015). *A comparison of direct control and pattern recognition control in Transhumeral TMR subjects.* Paper presented at the ISPO World Congress.

Kuiken, T. A., Miller, L. A., Turner, K., & Hargrove, L. J. (2016). A comparison of pattern recognition control and direct control of a multiple degree-of-freedom transradial prosthesis. *IEEE journal of translational engineering in health and medicine, 4*, 1-8.

Kumar, P. (2019). Computational Complexity of ML Models. Retrieved from https://medium.com/analytics-vidhya/time-complexity-of-ml-models-4ec39fad2770

Laird, J., Hucka, M., Huffman, S., & Rosenbloom, P. (1991). An analysis of Soar as an integrated architecture. *ACM SIGART Bulletin, 2*(4), 98-103.

Lambrecht, J. M., Pulliam, C. L., & Kirsch, R. F. (2011). Virtual reality environment for simulating tasks with a myoelectric prosthesis: an assessment and training tool. *Journal of prosthetics and orthotics: JPO, 23*(2), 89.

Leeb, R., Tonin, L., Rohm, M., Desideri, L., Carlson, T., & Millán, J. d. R. (2015). Towards Independence: A BCI Telepresence Robot for People With Severe Motor Disabilities. *Proceedings of the IEEE, 103*(6), 969-982. doi:10.1109/JPROC.2015.2419736

Leiden, K., & Best, B. (2005). A cross-model comparison of human performance modeling tools applied to aviation safety. *Micro Analysis & Design, Inc. Boulder, CO, 80301*, 2005.

Leiden, K., Laughery, K. R., Keller, J., French, J., Warwick, W., & Wood, S. D. (2001). A review of human performance models for the prediction of human error. *Ann Arbor, 1001*, 48105.

Li, J., Li, H., Umer, W., Wang, H. W., Xing, X. J., Zhao, S. K., & Hou, J. (2020). Identification and classification of construction equipment operators' mental fatigue using wearable

eye-tracking technology. *Automation in Construction, 109*, 15. doi:10.1016/j.autcon.2019.103000

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. R News 2 (3): 18–22. *URL: http://CRAN. R-project. org/doc/Rnews*.

Liu, C., White, R. W., & Dumais, S. (2010). *Understanding web browsing behaviors through Weibull analysis of dwell time.* Paper presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.

Liu, Y., Feyen, R., & Tsimhoni, O. (2006). Queueing Network-Model Human Processor (QN-MHP) A computational architecture for multitask performance in human-machine systems. *ACM Transactions on Computer-Human Interaction (TOCHI), 13*(1), 37-70.

Liu, Y., Wang, Y., Li, W., Xu, W., & Gui, J. (2009). *Improve driver performance by experience of driver cognitive behavior model's practice.* Paper presented at the 2009 IEEE Intelligent Vehicles Symposium.

Lock, B., Englehart, K., & Hudgins, B. (2005). *Real-time myoelectric control in a virtual environment to relate usability vs. accuracy*.

Lowry, P. B., Spaulding, T., Wells, T., Moody, G., Moffit, K., & Madariaga, S. (2006). *A theoretical model and empirical results linking website interactivity and usability satisfaction.* Paper presented at the Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06).

Lund, A. M. (2001). Measuring usability with the use questionnaire12. *Usability interface, 8*(2), 3-6.

Lusardi, M. M., Jorge, M., & Nielsen, C. C. (2013). *Orthotics and Prosthetics in Rehabilitation-E-Book*: Elsevier Health Sciences.

Majka, M. (2018). naivebayes: High performance implementation of the Naive Bayes algorithm. R package version 0.9. 2. In.

Markovic, M., Schweisfurth, M. A., Engels, L. F., Bentz, T., Wüstefeld, D., Farina, D., & Dosen, S. (2018). The clinical relevance of advanced artificial feedback in the control of a multi-functional myoelectric prosthesis. *Journal of neuroengineering and rehabilitation, 15*(1), 28.

Martins, R., & Carvalho, J. (2015). Eye blinking as an indicator of fatigue and mental load-a systematic review. *Occupational safety and hygiene III, 10*.

Maynard, H. B., Stegemerten, G. J., & Schwab, J. L. (1948). Methods-time measurement.

McCallum, A., & Nigam, K. (1998). *A comparison of event models for naive bayes text classification.* Paper presented at the AAAI-98 workshop on learning for text categorization.

McDonald, A. D., Ferris, T. K., & Wiener, T. A. (2019). Classification of Driver Distraction: A Comprehensive Analysis of Feature Generation, Machine Learning, and Input Measures. *Human Factors, 0*(0), 0018720819856454. doi:10.1177/0018720819856454

Meteier, Q., Capallera, M., Ruffieux, S., Angelini, L., Abou Khaled, O., Mugellini, E., Widmer, M., & Sonderegger, A. (2021). Classification of Drivers' Workload Using Physiological Signals in Conditional Automation. *Frontiers in Psychology, 12*, 18. doi:10.3389/fpsyg.2021.596038

Meyer, D. (2017). Support Vector Machines: The Interface to libsvm in package e1071, R package version 1.6-8. *URL https://cran. r-project. org/web/packages/e1071/vignettes/svmdoc. pdf*.

Michalco, J., Simonsen, J. G., & Hornbæk, K. (2015). An exploration of the relation between expectations and user experience. *International Journal of Human-Computer Interaction, 31*(9), 603-617.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review, 63*(2), 81.

Mislick, G. K., & Nussbaum, D. A. (2015). *Cost estimation: methods and tools*: John Wiley & Sons.

Mohebbian, M. R., Nosouhi, M., Fazilati, F., Esfahani, Z. N., Amiri, G., Malekifar, N., Yusefi, F., Rastegari, M., & Marateb, H. R. (2021). A Comprehensive Review of Myoelectric Prosthesis Control. *arXiv preprint arXiv:2112.13192*.

Montagnani, F., Controzzi, M., & Cipriani, C. (2015). Is it Finger or Wrist Dexterity That is Missing in Current Hand Prostheses? *IEEE Transactions on neural systems and rehabilitation engineering, 23*(4), 600-609. doi:10.1109/tnsre.2015.2398112

Morgan, S., Kelly, V., & Hafner, B. (2014). *The impact of transfemoral amputation on the cognitive load associated with level walking.* Paper presented at the 40th Annual Academy Meeting AAOP.

Moustafa, K., Luz, S., & Longo, L. (2017). *Assessment of mental workload: A comparison of machine learning methods and subjective assessment techniques.* Paper presented at the 1st International Symposium on Human Mental Workload: Models and Applications, H-WORKLOAD 2017, June 28, 2017 - June 30, 2017, Dublin, Ireland.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia*, *18*(60), 1-8.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*: MIT press.

Music, C. A. (2022). ELECTROMYOGRAPHY-BASED ASSISTIVE VIRTUAL REALITY HUMAN-MACHINE INTERFACE. *UNIVERSITY OF FLORIDA*.

Nadi, A., & Moradi, H. (2019). Increasing the views and reducing the depth in random forest. *Expert Systems with Applications*, *138*, 112801.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition*, *1*(1981), 1-55.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175-220.

Nielsen, J. (2005). Ten usability heuristics. In: http://www. nngroup. com/articles/ten-usability-heuristics/(acc-essed ….

Nielsen, J. (2012). Usability 101: Introduction to usability.

Nourbakhsh, N., Wang, Y., & Chen, F. (2013a, Sep 02-06). *GSR and Blink Features for Cognitive Load Classification.* Paper presented at the 14th IFIP TC 13 INTERACT International Conference on Designing for Diversity, Cape Town, SOUTH AFRICA.

Nourbakhsh, N., Wang, Y., & Chen, F. (2013b). *GSR and blink features for cognitive load classification.* Paper presented at the IFIP conference on human-computer interaction.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113.

Oliver, R. L. (1977). Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of applied psychology, 62*(4), 480.

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of marketing research, 17*(4), 460-469.

Olsen, N. R., George, J. A., Brinton, M. R., Paskett, M. D., Kluger, D. T., Tully, T. N., Duncan, C. C., & Clark, G. A. (2019). An Adaptable Prosthetic Wrist Reduces Subjective Workload. *bioRxiv*, 808634.

Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). *Estimating cognitive load using remote eye tracking in a driving simulator.* Paper presented at the Proceedings of the 2010 symposium on eye-tracking research & applications.

Park, J., Berman, J., Dodson, A., Liu, Y., Matthew, A., Huang, H., Kaber, D., Ruiz, J., & Zahabi, M. (2022). *Cognitive Workload Classification of Upper-limb Prosthetic Devices.* Paper presented at the 2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS).

Park, J., & Zahabi, M. (2020, 11-14 Oct. 2020). *Comparison of Cognitive Workload Assessment Techniques in EMG-based Prosthetic Device Studies.* Paper presented at the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC).

Park, J., & Zahabi, M. (2022a). Cognitive Workload Assessment of Prosthetic Devices: A Review of Literature and Meta-Analysis. *Ieee Transactions on Human-Machine Systems*.

Park, J., & Zahabi, M. (2022b). A review of human performance models for prediction of driver behavior and interactions with in-vehicle technology. *Human factors*, 00187208221132740.

Park, J., Zahabi, M., Kaber, D., Ruiz, J., & Huang, H. (2020). *Evaluation of Activities of Daily Living Tesbeds for Assessing Prosthetic Device Usability*. Paper presented at the 2020 IEEE International Conference on Human-Machine Systems (ICHMS).

Parr, J., Vine, S. J., Wilson, M. R., Harrison, N., & Wood, G. (2019). Visual attention, EEG alpha power and T7-Fz connectivity are implicated in prosthetic hand control and can be optimized through gaze training. *Journal of neuroengineering and rehabilitation, 16*(1), 52.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

Prada, L. R., & Boehm-Davis, D. A. (2004). *GOMS on the flight deck: A case study of the Boeing 777 MCP*. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Probst, P. (2019). *Hyperparameters, tuning and meta-learning for random forest and other machine learning algorithms.* lmu,

Pruziner, A. L., Shaw, E. P., Rietschel, J. C., Hendershot, B. D., Miller, M. W., Wolf, E. J., Hatfield, B. D., Dearth, C. L., & Gentili, R. J. (2019). Biomechanical and neurocognitive performance outcomes of walking with transtibial limb loss while challenged by a concurrent task. *Experimental brain research, 237*(2), 477-491.

Raihan-Al-Masud, M., & Mondal, M. R. H. (2020). Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. *Plos one, 15*(2), e0228422.

Raita, E., & Oulasvirta, A. (2011). Too good to be bad: Favorable product expectations boost subjective usability ratings. *Interacting with computers, 23*(4), 363-371.

Raufi, B. (2019). *Hybrid models of performance using mental workload and usability features via supervised machine learning.* Paper presented at the International Symposium on Human Mental Workload: Models and Applications.

Raveh, E., Friedman, J., & Portnoy, S. (2018). Evaluation of the effects of adding vibrotactile feedback to myoelectric prosthesis users on performance and visual attention in a dual-task paradigm. *Clinical rehabilitation, 32*(10), 1308-1316.

Rekant, J., Fisher, L. E., Boninger, M. L., Gaunt, R. A., & Collinger, J. L. (2022). Amputee, clinician, and regulator perspectives on current and prospective upper extremity prosthetic technologies. *Assistive Technology*, 1-13.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). *Tackling the poor assumptions of naive bayes text classifiers.* Paper presented at the Proceedings of the 20th international conference on machine learning (ICML-03).

Resnik, L., Huang, H. H., Winslow, A., Crouch, D. L., Zhang, F., & Wolk, N. (2018). Evaluation of EMG pattern recognition for upper limb prosthesis control: a case study in comparison with direct myoelectric control. *Journal of neuroengineering and rehabilitation, 15*(1), 23.

Rezazadeh, I. M., Firoozabadi, M., Hu, H., & Golpayegani, S. M. R. H. (2012). Co-adaptive and affective human-machine interface for improving training performances of virtual myoelectric forearm prosthesis. *IEEE Transactions on affective computing, 3*(3), 285-297.

Rezazadeh, I. M., Firoozabadi, S., Golpayegani, S. H., & Hu, H. (2011). *Controlling a virtual forehand prosthesis using an adaptive and affective Human-Machine Interface.* Paper presented at the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society.

Ritter, F. E. (2009). Two cognitive modeling frontiers. *Information and Media Technologies, 4*(1), 76-84.

Ritter, F. E., Tehranchi, F., & Oury, J. D. (2019). ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 10*(3), e1488. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1488 https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/wcs.1488?download=true

Rosenthal, R. (1976). Experimenter effects in behavioral research.

Rosenthal, R. (1986). Meta-Analytic Procedures for Social Science Research Sage Publications: Beverly Hills, 1984, 148 pp. *Educational Researcher, 15*(8), 18-20.

Rosyidah, U., Haryanto, H., & Kardianawati, A. (2019). *Usability Evaluation Using GOMS Model for Education Game "Play and Learn English".* Paper presented at the 2019 International Seminar on Application for Technology of Information and Communication (iSemantic).

Ruiz Ramírez, E. (2016). *Control of a hand prosthesis using mixed electromyography and pressure sensing.* Universitat Politècnica de Catalunya,

Rupp, R., Rohm, M., Schneiders, M., Weidner, N., Kaiser, V., Kreilinger, A., & Müller-Putz, G. (2013). Think2grasp-bci-controlled neuroprosthesis for the upper extremity. *Biomedical Engineering/Biomedizinische Technik*.

Sabri, M., Miskon, M., Yaacob, M., Basri, A. S. H., Soo, Y., & Bukhari, W. (2014). MVC BASED NORMALIZATION TO IMPROVE THE CONSISTENCY OF EMG SIGNAL. *Journal of Theoretical & Applied Information Technology, 65*(2).

Salvucci, D. D., & Lee, F. J. (2003). *Simple cognitive modeling in a complex cognitive architecture.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Salvucci, D. D., Zuber, M., Beregovaia, E., & Markley, D. (2005). *Distract-R: Rapid prototyping and evaluation of in-vehicle interfaces.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Samsonovich, A. (2015). Comparative Table of Implemented Cognitive Architectures. Retrieved from https://bicasociety.org/cogarch/

Sankaranarayanan, G., Wooley, L., Hogg, D., Dorozhkin, D., Olasky, J., Chauhan, S., Fleshman, J. W., De, S., Scott, D., & Jones, D. B. (2018). Immersive virtual reality-based training improves response in a simulated operating room fire scenario. *Surgical endoscopy, 32*(8), 3439-3449.

Saraiji, M., Sasaki, T., Kunze, K., Minamizawa, K., & Inami, M. (2018). *MetaArmS: Body remapping using feet-controlled artificial arms.* Paper presented at the The 31st Annual ACM Symposium on User Interface Software and Technology.

Scheme, E., & Englehart, K. (2011). Electromyogram pattern recognition for control of powered upper-limb prostheses: state of the art and challenges for clinical use. *Journal of Rehabilitation Research & Development, 48*(6).

Schneider, D. J. (1973). Implicit personality theory: A review. *Psychological bulletin, 79*(5), 294.

Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica, 31*(1), 27-53.

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

Shao, S., Wang, T., Li, Y., Song, C., Jiang, Y., & Yao, C. (2021). Comparison Analysis of Different Time-Scale Heart Rate Variability Signals for Mental Workload Assessment in Human-Robot Interaction. *Wireless Communications and Mobile Computing, 2021*. doi:10.1155/2021/8371637

Sharma, H., Drukker, L., Papageorghiou, A. T., & Noble, J. A. (2021). Machine learning-based analysis of operator pupillary response to assess cognitive workload in clinical ultrasound imaging. *Computers in Biology and Medicine, 135*, 14. doi:10.1016/j.compbiomed.2021.104589

Shaw, E. P., Rietschel, J. C., Hendershot, B. D., Pruziner, A. L., Miller, M. W., Hatfield, B. D., & Gentili, R. J. (2018). Measurement of attentional reserve and mental effort for cognitive workload assessment under various task demands during dual-task walking. *Biological Psychology, 134*, 39-51.

Shaw, E. P., Rietschel, J. C., Hendershot, B. D., Pruziner, A. L., Wolf, E. J., Dearth, C. L., Miller, M. W., Hatfield, B. D., & Gentili, R. J. (2019a). A Comparison of Mental Workload in Individuals with Transtibial and Transfemoral Lower Limb Loss during Dual-Task Walking under Varying Demand. *Journal of the International Neuropsychological Society, 25*(9), 985-997. doi:10.1017/S1355617719000602

Shaw, E. P., Rietschel, J. C., Shuggi, I. M., Xu, Y., Chen, S., Miller, M. W., Hatfield, B. D., & Gentili, R. J. (2019b). Cerebral cortical networking for mental workload assessment under various demands during dual-task walking. *Experimental brain research, 237*(9), 2279-2295. Retrieved from https://link.springer.com/content/pdf/10.1007%2Fs00221-019-05550-x.pdf

Sheng, B., & Wan, C.-x. (2013). Comparison of the reaction time of wrist flexion and extension between patients with stroke and age-matched healthy subjects and correlation with clinical measures. *Chinese medical journal, 126*(13), 2485-2488.

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology, 45*(5), 679-687.

Sirevaag, E. J., Kramer, A. F., REISWEBER, C. D. W. M., STRAYER, D. L., & GRENELL, J. F. (1993). Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics, 36*(9), 1121-1140. Retrieved from https://www.tandfonline.com/doi/pdf/10.1080/00140139308967983?needAccess=true

Skaramagkas, V., Ktistakis, E., Manousos, D., Tachos, N. S., Kazantzaki, E., Tripoliti, E. E., Fotiadis, D. I., & Tsiknakis, M. (2021). *Cognitive workload level estimation based on eye tracking: A machine learning approach.* Paper presented at the 21st IEEE International Conference on BioInformatics and BioEngineering, BIBE 2021, October 25, 2021 - October 27, 2021, Kragujevac, Serbia.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of personality and social psychology, 36*(11), 1202.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management, 45*(4), 427-437.

Spreng, R. A., MacKenzie, S. B., & Olshavsky, R. W. (1996). A reexamination of the determinants of consumer satisfaction. *Journal of marketing, 60*(3), 15-32.

Stanley, R. M., Hall, B. G., Baden, W. A., & Exum, M. (2019). *Analyzing Air Traffic Controller Task Times and Capacity Benefits From New Automation Capabilities.* Paper presented at the 2019 Integrated Communications, Navigation and Surveillance Conference (ICNS).

Stratford, P. W. (2001). Development and initial validation of the upper Ectremity functional index. *Physiother Can, 52*, 259-267.

Stubblefield, K., Lipschutz, R., Phillips, M., Heckathorne, C., & Kuiken, T. (2005). *Occupational therapy outcomes with targeted hyper-reinnervation nerve transfer surgery: Two case studies.* Paper presented at the Proceedings of the MyoElectric Controls/Powered Prosthetics Symposium.

Taatgen, N., & Anderson, J. R. (2010). The past, present, and future of cognitive architectures. *Top Cogn Sci, 2*(4), 693-704. doi:10.1111/j.1756-8765.2009.01063.x

Tang, C., Garreau, D., & von Luxburg, U. (2018). When do random forests fail? *Advances in neural information processing systems, 31*.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of applied psychology, 4*(1), 25.

Tiffin, J., & Asher, E. J. (1948). The Purdue Pegboard: norms and studies of reliability and validity. *Journal of applied psychology, 32*(3), 234.

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in sport sciences, 21*(1).

Valverde-Albacete, F. J., Carrillo-de-Albornoz, J., & Peláez-Moreno, C. (2013). *A proposal for new evaluation metrics and result visualization technique for sentiment analysis tasks.* Paper presented at the International Conference of the Cross-Language Evaluation Forum for European Languages.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T.-P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human factors*, *43*(1), 111-121. Retrieved from https://journals.sagepub.com/doi/pdf/10.1518/001872001775992570

Van Rijn, H., Johnson, A., & Taatgen, N. (2011). Cognitive user modeling. *Handbook of human factors in web design*, 523-538.

VandenBos, G. R. (2007). *APA dictionary of psychology*: American Psychological Association.

Volkmar, R., Dosen, S., Gonzalez-Vargas, J., Baum, M., & Markovic, M. (2019). Improving bimanual interaction with a prosthesis using semi-autonomous control. *Journal of neuroengineering and rehabilitation*, *16*(1), 140.

Walambe, R., Nayak, P., Bhardwaj, A., & Kotecha, K. (2021). Employing Multimodal Machine Learning for Stress Detection. *Journal of Healthcare Engineering*, *2021*, 12. doi:10.1155/2021/9356452

Wang, W., Li, Z., Wang, Y., & Chen, F. (2013). *Indexing cognitive workload based on pupillary response under luminance and emotional changes.* Paper presented at the 18th International Conference on Intelligent User Interfaces, IUI 2013, March 19, 2013 - March 22, 2013, Santa Monica, CA, United states.

White, M. M., Zhang, W., Winslow, A. T., Zahabi, M., Zhang, F., Huang, H., & Kaber, D. B. (2017). Usability comparison of conventional direct control versus pattern recognition

control of transradial prostheses. *Ieee Transactions on Human-Machine Systems, 47*(6), 1146-1157.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics* (pp. 196-202): Springer.

Williams, R. M., Turner, A. P., Orendurff, M., Segal, A. D., Klute, G. K., Pecoraro, J., & Czerniecki, J. (2006). Does Having a Computerized Prosthetic Knee Influence Cognitive Performance During Amputee Walking? *Archives of physical medicine and rehabilitation, 87*(7), 989-994. doi:https://doi.org/10.1016/j.apmr.2006.03.006

Wilson, G., & Schlegel, R. (2004). Operator functional state assessment. Paris, FR, North Atlantic Treaty Organisation (NATO). *Research and Technology Organisation (RTO) BP, 25*.

Witteveen, H., de Rond, L., Rietman, J. S., & Veltink, P. H. (2012). Hand-opening feedback for myoelectric forearm prostheses: performance in virtual grasping tasks influenced by different levels of distraction. *J Rehabil Res Dev, 49*(10), 1517-1526.

Wood, G., & Parr, J. (2022). A tool for measuring mental workload during prosthesis use: The Prosthesis Task Load Index (PROS-TLX).

Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human factors, 30*(1), 111-120.

Yuan, H., Li, S., & Rusconi, P. (2020). Cognitive Approaches to Human Computer Interaction. In *Cognitive Modeling for Automated Human Performance Evaluation at Scale* (pp. 5-15): Springer.

Zahabi, M. (2017). Analysis and Redesign of Police Vehicle Mobile Computer Terminal for Minimizing Officer Driving Distraction.

Zahabi, M., & Lyman, A. (2019). *Impact of Electronic Medical Records on Patient-Provider Communication.* Paper presented at the Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care.

Zahabi, M., & McCollum, E. (2019). *An Application of Machine Learning for Police Mobile Computer Terminal Usability Evaluation.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Zahabi, M., Shupsky, T., & Lymanb, A. (2019a). Analysis of Law Enforcement Mobile Computer Terminal Interface.

Zahabi, M., White, M. M., Zhang, W., Winslow, A. T., Zhang, F., Huang, H., & Kaber, D. B. (2019b). Application of Cognitive Task Performance Modeling for Assessing Usability of Transradial Prostheses. *Ieee Transactions on Human-Machine Systems, 49*(4), 381-387. doi:10.1109/THMS.2019.2903188

Zhang, D., Xu, H., Shull, P. B., Liu, J., & Zhu, X. (2015). Somatotopical feedback versus non-somatotopical feedback for phantom digit sensation on amputees using electrotactile stimulation. *Journal of neuroengineering and rehabilitation, 12*(1), 44.

Zhang, W., Ma, W., Brandao, M., Kaber, D. B., Bloomfield, P., & Swangnetr, M. (2016a). Biometric validation of a virtual reality-based psychomotor test for motor skill training. *Assistive Technology, 28*(4), 233-241.

Zhang, W., White, M., Zahabi, M., Winslow, A. T., Zhang, F., Huang, H., & Kaber, D. (2016b). *Cognitive workload in conventional direct control vs. pattern recognition control of an upper-limb prosthesis.* Paper presented at the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC).

Zhang, Y., & Wu, C. (2017). *Learn to Integrate Mathematical Models in Human Performance Modeling.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Zhang, Y., Yang, J., Bai, D., & Wang, Y. (2018). *A Research about the Mental Fatigue of using an Intelligent Artificial Limb based on Functional Near Infrared Spectrum Technique.* Paper presented at the 2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR).

APPENDIX A

QUEST 2.0

Quebec User Evaluation of Satisfaction with assistive Technology

Participant number:

Device configuration (select one): DC/ PR/ CC

Date:

The purpose of the **QUEST** questionnaire is to evaluate how satisfied you are with your assistive device. The questionnaire consists of 8 satisfaction items.

- For each of the 8 items, rate your satisfaction with your assistive device by using the following scale of 1 to 5.

1	**2**	**3**	**4**	**5**
not satisfied at all	not very satisfied	more or less satisfied	quite satisfied	very satisfied

- Please circle or mark the **one number** that best describes your degree of satisfaction with each of the 8 items.
- D**o not** leave any question unanswered.
- For any item that you were not "very satisfied", please comment in the section
comments.

Thank you for completing the QUEST questionnaire.

ASSISTIVE DEVICE ***How satisfied are you with,***	
1. the **dimensions** (size, height, length, width) of your assistive device? *Comments*:	1 2 3 4 5
2. the **weight** of your assistive device? *Comments*:	1 2 3 4 5
3. the **ease in adjusting** (fixing, fastening) the parts of your assistive device? *Comments*:	1 2 3 4 5
4. how **safe and secure** your assistive device is? *Comments*:	1 2 3 4 5
5. the **durability** (endurance, resistance to wear) of your assistive device? *Comments*:	1 2 3 4 5
6. how **easy** it is to use your assistive device? *Comments*:	1 2 3 4 5
7. how **comfortable** your assistive device is? *Comments*:	1 2 3 4 5
8. how **effective** your assistive device is (the degree to which your device meets your needs)? *Comments*:	1 2 3 4 5

- Below is the list of the same 8 satisfaction items. PLEASE **SELECT THE THREE ITEMS** that you consider to be **the most important to you**. Please put an X in the **3 boxes** of your choice.

☐ 1. Dimensions	☐ 7. Comfort
☐ 2. Weight	☐ 8. Effectiveness
☐ 3. Adjustments	☐
☐ 4. Safety	☐
☐ 5. Durability	☐
☐ 6. Easy to use	☐

APPENDIX B

USE

Participant number:
Device (Select one): DC/ PR/ CC
Date:
Please rate your agreement with these statements.

- Respond to all the items
- For items that are not applicable, use: NA

USEFULNESS		1	2	3	4	5	6	7		NA
1. It helps me be more effective.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
2. It helps me be more productive.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
3. It is useful.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
4. It gives me more control over the activities in my life.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
5. It makes the things I want to accomplish easier to get done.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
6. It saves me time when I use it.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
7. It meets my needs.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
8. It does everything I would expect it to do.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○

EASE OF USE		1	2	3	4	5	6	7		NA
9. It is easy to use.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
10. It is simple to use.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
11. It is user friendly.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
12. It requires the fewest steps possible to accomplish what I want to do with it.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
13. It is flexible	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
14. Using it is effortless	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○

15. I can use it without written instructions	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
16. I don't notice any inconsistencies as I use it.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
17. Both occasional and regular users would like it.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
18. I can recover from mistakes quickly and easily.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
19. I can use it successfully every time.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○

EASE OF LEARNING		1	2	3	4	5	6	7		NA
20. I learned to use it quickly.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
21. I easily remember how to use it.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
22. It is easy to learn to use it.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
23. I quickly became skillful with it.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○

SATISFACTION		1	2	3	4	5	6	7		NA
24. I am satisfied with it.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
25. I would recommend it to a friend.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
26. It is fun to use.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
27. It works the way I want it to work.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
28. It is wonderful.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
29. I feel I need to have it.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○
30. It is pleasant to use.	Strongly disagree	○	○	○	○	○	○	○	Strongly agree	○

APPENDIX C

NASA-TLX

During the test you have just completed you may have experienced some difficulties and constraints with regard to the task.

You will be asked to evaluate this experience with regard to 6 factors, which are described below. Please read each factor and its description carefully and ask the experimenter to explain anything you do not fully understand.

Title	Endpoints	Description
Mental demand	Low/high	How much mental and perceptual acdtivity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc)? Was the task easy or demanding, simple or complex, exacting or forviging?
Physical demand	Low/high	How much physical acitvity was required (e.g., pushing, pulling, turning, controlling, activating, etc)? Was the task easy or demading, slow or brisk, slack or strenuous, restful or laborious?
Temporal demand	Low/high	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	Low/high	How successful do you think you were in accomplisihng the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	Low/high	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration level	Low/high	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified , content, relaxed and complacent did you feel during the task?

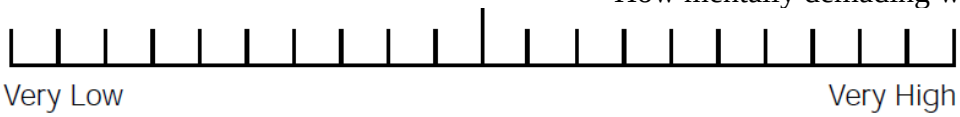
For each of the pairs below, circle the scale title that represents the more important contributor to workload when you are performing the task.

Pair 1 Effort or Performance	Pair 2 Temporal Demand or Frustration
Pair 3 Temporal Demand or Effort	Pair 4 Physical Demand or Frustration
Pair 5 Performance or Frustration	Pair 6 Physical Demand or Temporal Demand
Pair 7 Physical Demand or Performance	Pair 8 Temporal Demand or Mental Demand
Pair 9 Frustration or Effort	Pair 10 Performance or Mental Demand
Pair 11 Performance or Temporal Demand	Pair 12 Mental Demand or Effort
Pair 13 Mental Demand or Physical Demand	Pair 14 Effort or Physical Demand
Pair 15 Frustration or Mental Demand	

For each factor you will be required to rate the level of constraint felt during the test on a scale from “Very low (0)” to “Very high (100)”, with regard to the task. Please circle one of tick marks in each factor.

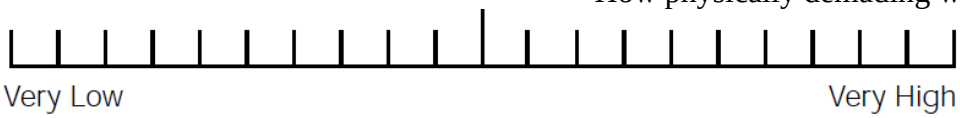
Mental demand:

How mentally demading was the task?



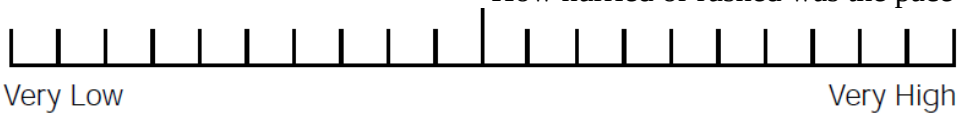
Physical demand:

How physically demading was the task?



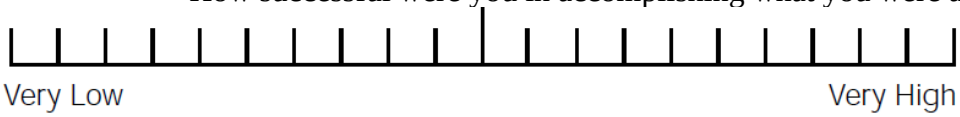
Temporal demand:

How hurried or rushed was the pace of the task?



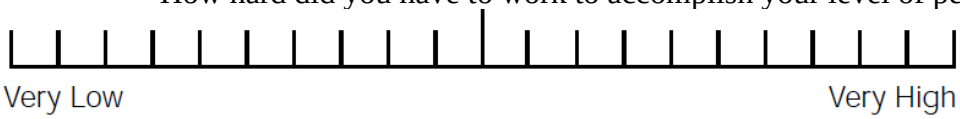
Performance:

How successful were you in accomplishing what you were asked to do?



Effort:

How hard did you have to work to accomplish your level of performance?



Frustration:

How insecure, discouraged, irritated, stressed, and annoyed were you?