Human Computer Interaction for Complex Machine Learning

Kobla Setor Zilevu

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Application

Aisling Kelliher, Co-chair
Thanassis Rikakis, Co-chair
Douglas Bowman
Sang Won Lee
Deana Anglin

February 15, 2022
Blacksburg, Virginia

Keywords: Assistive technologies, TCE, Stroke Rehabilitation, HCI

Human Computer Interaction for Complex Machine Learning

Kobla Setor Zilevu

(ABSTRACT)

This dissertation focuses on taking a human-centric approach to utilize human intelligence best to inform machine learning models. More specifically, the complex relationship between the changes in movement functionality to movement quality. I designed and evaluated the Tacit Computable Empowerment methodology across two domains: in-home rehabilitation and clinical assessment. My methodology has three main objectives: first, to transform tacit expert knowledge into explicit knowledge. Second, to transform explicit knowledge into a computable framework that machine learning can understand and replicate. Third, synergize human intelligence with computational machine learning to empower, not replace, the human. Finally, my methodology uses assistive interfaces to allow clinicians and machine learning models to draw parallels between movement functionality and movement quality. The results from my dissertation inform researchers and clinicians on how best to create a standardized framework to capture and assess human movement data for embodied learning scenarios.

Human Computer Interaction for Complex Machine Learning

Kobla Setor Zilevu

(GENERAL AUDIENCE ABSTRACT)

Artificial intelligence (AI) is increasingly considered an important computational design material in the development of innovative products, systems and services. Recent research emphasizes the potential for computational designers to create new tools, methods and design processes to more adeptly handle AI and machine learning as fundamental, but not exclusive, materials within the design process. This talk adopts a human-centric approach to utilize human intelligence to inform machine learning models within a healthcare context. I describe the novel Tacit Computable Empowerment (TCE) methodology used and evaluated across two healthcare domains: in-home rehabilitation and clinic-based assessment. The TCE methodology comprises three main objectives: 1) to transform tacit expert knowledge into explicit knowledge; 2) to transform explicit knowledge into a computable framework that machine learning can understand and replicate and 3) to synergize human intelligence with computational machine learning to empower (and not replace) the human. This methodology uses assistive interfaces to allow clinicians and machine learning models to draw parallels between movement functionality and movement quality. Outcomes from this work inform researchers and clinicians as to how to best create a standardized framework to capture and assess human movement data for embodied learning scenarios.

Dedication

To Mom and Dad, thank you so much for all your support and love throughout this journey. I owe you both the world and more and I hope this dissertation makes you proud. To my two brothers, thank you for the sacrifices you both made in order to make this a reality. To my sister-in-law, thank you for pushing me throughout the journey and keeping me grounded. To my friends (extended-family), you all stood by me throughout this journey and saw the bigger picture. I reached the finish line because of you all. Whether big or small, you all made sure I never fell down. Thank you. And to my nephew...you are up next! This dissertation could not have been done with you all. Romans 8:28 - And we know that all that happens to us is working for our good if we love God and are fitting into his plans.

Acknowledgments

Thank you to my advisor Aisling Kelliher, co-advisor Thanassis Rikakis, committee members, Doug Bowman, Sang Won, Lee, and Deana Anglin. Also to the INR team!

Contents

List of Figures

List of Tables

List of Abbreviations

ADL Activities of Daily Living

ARAT Action Research Arm Test

F2MQ Function to Movement Quality

HCI Human Computer Interaction

IML Interactive Machine Learning

INR Interactive Neurorehabilitation

IRR Inter Rater Reliability

ML Machine Learning

MQ2F Movement Quality to Function

OT Occupational Therapist

PD Participatory Design

PT Physical Therapist

SARAH Semi Automated Rehabilitation At Home

SDP Structured Decision Process

SRAL Shirley Ryan Ability Lab

TCE Tacit Computable Empowerment

TPS Time per Session

UX User Experience

VAT Video Annotation Tool

Chapter 1

Introduction

Artificial intelligence (AI) is increasingly considered as an important computational design material in the development of innovative products, systems, and services [35]. The framing of AI as a possible key constituent in the design process necessitates a rethinking of the end-goal function and use of computational design solutions in an era of "evolving complementary capabilities and doings" [80]. AI, and in particular, machine learning, has the potential to radically reorient the designers approach in crafting high quality user experiences [85] as a human-centered design stance might now also have to accommodate the needs of machines. Transportation, financial services, retail, entertainment, construction, and logistics are all industries that are rapidly embracing machine learning to gain a competitive edge, better understand their own enterprise, and improve their quality of service.

A growing body of literature within design and human-computer-interaction points to the opportunities and possibilities for these fields to contribute to, and even lead, the development of more human-centered AI systems, grounded within ecologically valid contexts [85], [20]. Recent work emphasizes the potential for computational designers to create new tools, methods, and design processes to more adeptly handle AI and machine learning as fundamental (but not exclusive) materials within the design process [85], [35]. One way to conceptualize how computational designers might handle AI as an interactive material is to focus initially on defining the desired outcome. If, for example, the outcome is to automate a task within a human replacement paradigm, or to locate a photo within a transactional

paradigm, the design constraints are relatively narrow. However, if the desired outcome is more aspirational, such as the parallel growth of humans and machines aimed at improving the human condition, the design constraints and indeed, the design process will be quite different.

Recent work within the design and human computer interaction communities surface a rich variety of perspectives on the potential role of design as an important vector within AI research and practice. Practitioner surveys [20] and HCI literature reviews [85] highlight not only the growing significance of machine learning and artificial intelligence within the discipline, but also point out the paucity of integrative innovation between these technologies and user experience design. Within this burgeoning area of research and practice, preliminary explorations note that the use of plug-and-play AI approaches can result in non-experts implementing poorly understood "black box" models [85], while the lack of accessible AI or machine learning prototyping models can also hinder design lead innovation [20]. Multidisciplinary teams collaborating together on complex AI problems is one possible way to address this shortcoming. In particular, iterative participatory design approaches could provide an opportunity for such teams to work together building cyber-human systems promoting "continuous learning by the human actors in tandem with learning by the AI agent" [66].

Education, healthcare, and social justice are three significant global areas of inquiry that could greatly benefit from ethical innovations in artificial intelligence. The worldwide COVID-19 pandemic brings these issues even more to the fore as recurring lockdowns keep people at home, necessitating the transfer of services to virtual domains. Over the last decade, important progress has been made within the area of telehealth and telemedicine in delivering healthcare in the community and at home at scale. As the global population ages, there is a growing need for rehabilitation services for debilitating illnesses such as stroke, Parkinson's and arthritis [33]. Technology assisted rehabilitation, using commercial

consumer tools (e.g. Fitbit) or custom designed products shows promise, but issues of cost, patient acceptance, and replicating the expertise of the clinician present challenges [40]. Effective smart rehabilitation in the clinic and at home requires that both therapy assessment and training sessions and activities of daily living are captured and analyzed in a manner that effectively supports evidence-based supervision and adaptation of therapy [63].

This dissertation takes a collaborative and human-centric computational design approach towards understanding the application of Human-Computer-Interaction (HCI) for complex Machine Learning (ML) in Embodied Learning (EL) scenarios. The work described in this document has focused on applying HCI methodologies to complex ML contexts to capture and assess how humans move and learn in embodied spaces. HCI and ML can be considered existentially different in their primary objectives, methodologies, and evaluation processes within embodied spaces. HCI typically takes a human-centered approach, which focuses on using human intelligence to best design solutions, whereas ML processes are typically focused on using highly standardized, quantifiable datasets as input for extremely fast processors to provide generalizable outputs.

Human learning in embodied learning spaces is often considered tacit and challenging to uncover. This dissertation takes a three-pronged approach to designing a solution to reveal and augment this form of human knowledge. The first design challenge aims to understand how HCI and design can reveal and transform tacit human knowledge into explicit knowledge. The second challenge is to then transform explicit human expertise into a computable model understandable to an AI. The third and final design challenge is to leverage AI computations to promote further learning and enhance human intelligence in embodied learning spaces The primary objective of my dissertation is to create an integrative HCI method that synergizes human intelligence with computational intelligence.

This dissertation proposes the methodology of Tacit Computable Empowerment (TCE). This

methodology comprises three highly interrelated and iterative processes: First it transforms tacit human knowledge into explicit knowledge. Next, it converts explicit human knowledge into a computable model. Finally it uses computational power to empower the human. To assess the validity of the TCE methodology, we interrogate the approach within two critical healthcare frameworks: stroke rehabilitation movement training and stroke movement assessment.

This process requires the design and evaluation of interrelated HCI metrics and interfaces to reveal the processes:

1. Movement assessment rubrics + rating
2. Movement capture interface
3. Human empowerment summaries

In Figure 1.1 I describe the process I used to answer my primary research question. I have developed a standardized methodological approach to understanding how HCI can be integrated into complex ML applications. This figure showcases that the TCE method is designed to synergize both human and computational intelligence through the use of interactive interfaces.

The future will bring a world where computers and technology are an integral part of human existence within healthcare. COVID-19 has expedited this process, as we are seeing a shift from traditional face-to-face healthcare into a model primarily mediated by technology. However, to make these services explicit and computable, human intelligence needs to be at the forefront of designing for complex machine learning solutions. The global pandemic emphasizes the tremendous need for accurate and meaningful automated tele-rehabilitation tools for home and community-based care. Developing such systems requires integrating

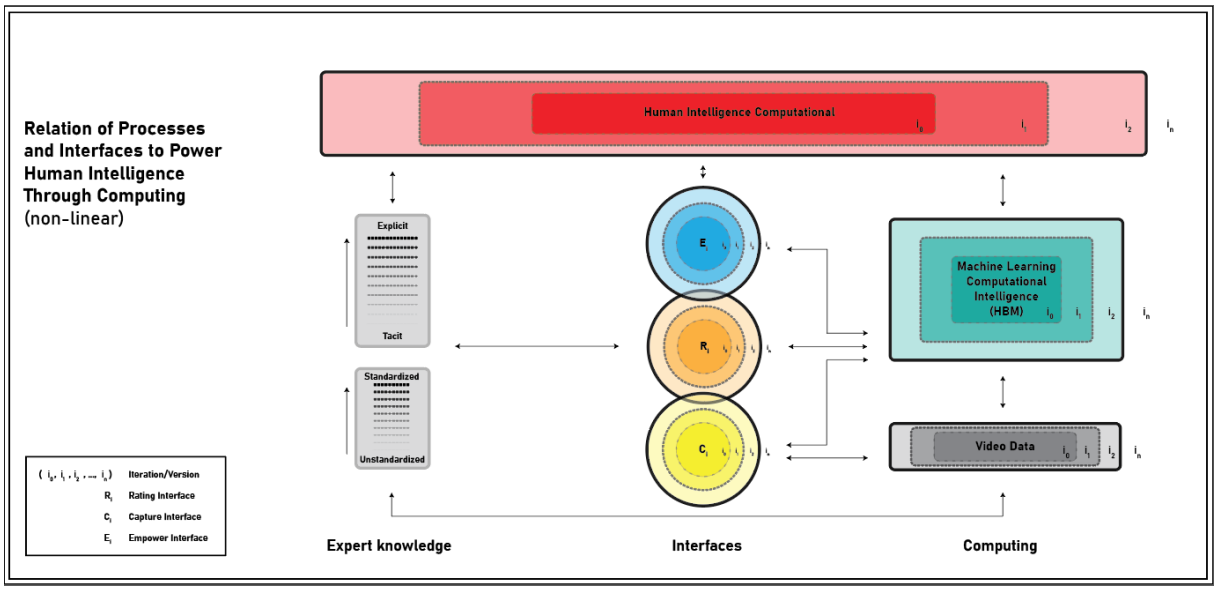

Figure 1.1: The Relationship of Processes and Interfaces to Power Human Intelligence through Computing (non-linear)

expert knowledge with data-driven approaches to produce reliable assistive technologies at scale. This endeavor has numerous challenges, including spare data sets, non-standardized data collection, varying analysis approaches, and the need to support clinicians in revealing their expert implicit knowledge.

In the US alone, Medicare primary care visits via telehealth rose from 1% in February 2021 to almost 50% in April 2021 [8]. However, even pre-pandemic, the field of telehealth was already considerably rising in prominence and prevalence. With the loosening of insurance restrictions in the US concerning routine medical care, telehealth is expected to continue to develop and grow. There are increasing needs for general physical and occupational rehabilitation services for common age-related debilitating illnesses such arthritis, stroke, and osteoporosis. Effective rehabilitation requires intensive training and the ability to adapt the training program based on patient progress and therapeutic judgment. Active participation by the patient is also critical for improving self-efficacy and program adherence. However, intensive and adaptive rehabilitation is challenging to administer in an accessible and af-

fordable way. It necessitates frequent trips by the patient to the clinic (usually reliant on a caregiver) and significant face-to-face time with rehabilitation experts. Ultimately, the biggest problem here is a significant lack of available rehabilitation experts and therapists to cover the needs of a geographically dispersed and aging population.

Clinical assessment in the wild is complex because it is not standardized, making it difficult for clinicians to receive all the necessary patient information required for assessment. Furthermore, the time consuming process of information gathering is strenuous for clinicians. Additionally, this process is complex to execute in a standardized and detailed manner. However, utilizing the TCE method may reduce the capturing and assessment effort for the clinician while still ensuring the accuracy and standarizated of the medical protocol.

Technology-assisted rehabilitation in the clinic offers the possibility of significantly improving the reach and wellness outcomes of rehabilitation therapy, which could also be accessible and affordable at scale. The scaling and structuring of this form of rehabilitation still face significant challenges. However, fully automated rehabilitation for clinical assessment is not currently feasible as the functions of the therapist cannot be comprehensively replicated by a machine. Therapists differ widely in their approaches to evaluation. These approaches are not fully observable as they are partly based on experience and implicit knowledge rather than a standardized quantifiable taxonomy. For example, the assessment of knee replacement by physical therapists or administering an assessment for a stroke survivor by occupational therapist.

Cyber-human approaches leveraging productive collaboration of the therapist with computational intelligence and active participation of patients, however, provide a promising avenue of inquiry. These cyber-human approaches strive not to move the therapist functions to a machine but rather conceptualize and implement a cyber-human system that can allow the scaling of effective and efficient rehabilitation in the home as shown in Fig 1.2. Although the

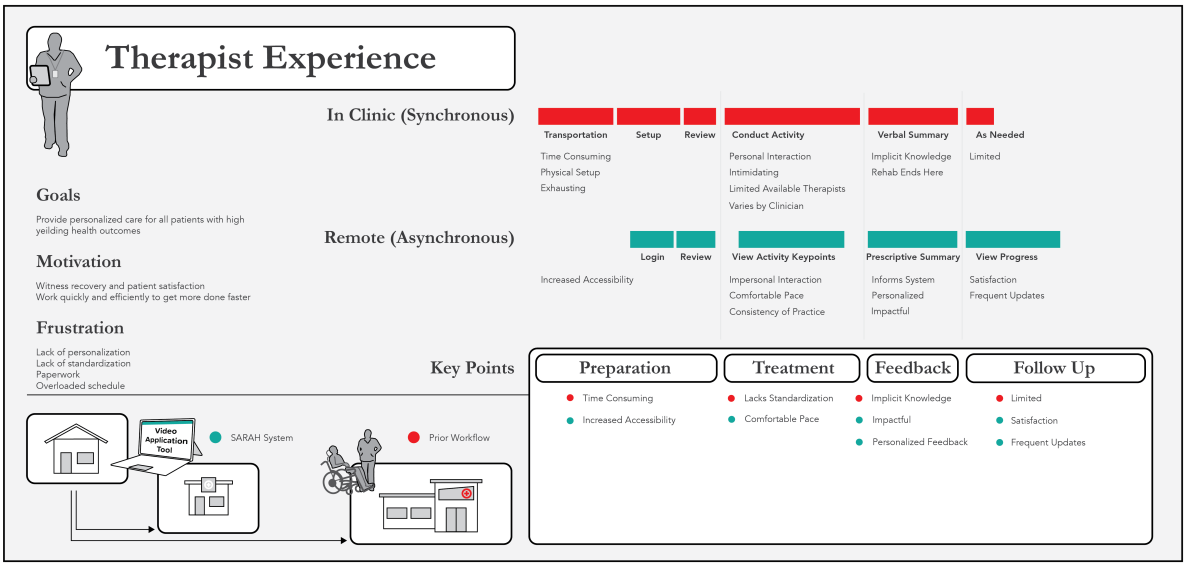

Figure 1.2: The therapist persona as it pertain to their journey with a in-clinic experience and remote asynchronous experience

resulting experience may be quite different from a traditional face-to-face visit in the clinic, the quality and impact will be equally high.

A standardized process needs to be designed and implemented to allow clinical therapists to capture and reflect the different behaviors of the patient throughout the rehabilitation assessment. However, telehealth has limited tools that provide high computational outcomes for human cyber systems and lacks a synthesized process for reflection. To justify the development effort for such hybrid systems, technology used within this cyber-human model needs to show potential for enhancing the rehabilitation process, the therapist/patient experiences, and the therapeutic outcomes.

With this dissertation, I aim to understand how this methodology can improve the experience of the patient, therapist, and machine learning system to create a process in which, through a participatory design approach, each stakeholder gradually improves through each iteration.

Without the use of ML and a human-centered solution, the workflow process for both the therapist and patient is very challenging, especially within hybrid settings. The time, cost, and lack of expert availability makes it difficult for patients to ensure that they are receiving the proper care and for clinicians to provide that care. My proposed TCE methodology for more inclusive computationally enhanced healthcare can achieves three outcomes. First, it can help the patient to get better by allowing clinicians to focus on personalized treatment and assessment plans. Second, it can allow clinicians to spend more time focusing on patient outcomes and less time on administrative assessment needs. Third, it allows for machine learning systems to capture high-quality, and ultimately meaningful data. The TCE method can allow for all three outomes to be met simultaneously.

1.1 Research Questions Overview

Considering these outcomes, I frame my dissertation to address two critical questions:

1. What is a HCI design methodology that synergizes human intelligence with computational intelligence, makes tacit human knowledge explicit, turns that explicit human knowledge computable, and finally uses that computational power to empower the human?

2. How do I use this methodology to empower rather than replace the human in order to incentivize human participation?

1.2 Dissertation Overview

This dissertation is organized into seven sections: Introduction, Literature Review, Methods, TCE Methodology for Home Based Rehabilitation Training, CE Methodology for Standardized Clinical Assessment, Discussion, Conclusion, and multiple Appendixes. The outcomes of work described in this document was previously published at ACM CHI, PETRA, IMX, Frontiers in Neurology and in my VT Masters Thesis: [65], [40], [15], [41], [87], [88]. The studies described in this dissertation have been conducted at various universities and hospitals. The SARAH system was evaluated at the Emory Rehabilitation Hospital, and the ARAT system was evaluated at the Shirley Ryan Ability Lab in Chicago. The TCE method was co-designed with expert clinicians and patient, who also participated in the implementation and evaluation process.

Chapter 2: Literature Review Chapter 2 explores the research and work on embodied learning spaces in the wild. I primarily focus on Hutchins's and Schoen's work on the human-computer interface's embodied learning and reflective process. I then review literature discussing the timeline of User Experience to understand how User Experience has influenced past, present, and now future technology systems. Next, I review literature explaining the state-of-the-art for both in-home rehabilitation and clinical assessment. I end my literature review by exploring studies of the challenges of complex ML for embodied spaces.

Chapter 3: Methods Chapter 3 describes the process I used to develop my methodology. My dissertation uses a Participatory Design framework. I utilize qualitative and quantitative methods to understand the relationship between human and computational intelligence. Through ethnographic studies, I understand the varying user journeys in EL spaces that can be explicit, computable, and empowering.

I take on the role of a Human Factors Engineer and User Experience Researcher to explore

the best approach to merge HCI with ML for complex spaces. Through my research, we see that as humans and ML move towards these complex spaces, human and computational intelligence must be synergized to empower humans to create meaningful data and results.

Chapter 4: TCE Methodology for Home Based Rehabilitation Training In Chapter 4, I describe the TCE methodology for the design and development of a system for home based rehabilitation training. This system is called "Semi Automated Rehabilitation At Home (SARAH).

First, I focus on the approach I developed to understand how humans think in an implicit form in embodied spaces. I worked with thirteen clinicians for over four years to understand their thinking process for both in-home rehabilitation and clinical assessment. Through the various one-on-one interviews, focus groups, and co-design activities, I worked with the expert clinicians to develop a rubric that first translated their implicit thinking into explicit information. Then, I created a novel Video Application Tool that allows humans in embodied spaces to translate their tacit knowledge explicitly. Through this iterative process, I brought out explicit knowledge that can be computable for ML in these complex spaces. The findings from this chapter provide new levels of reflection for humans in these environments.

Second, I describe the process of creating computable models from explicit expert knowledge. This section describes the method I first used to understand both noisy variable data for a standardized approach to capture clean, low-cost, unobtrusive video data. I then begin to reveal how video data can be translated in a meaningful way for complex ML algorithms while still providing high value for the human. Finally, I explain how I aim to understand the relationship between Human and Computational Intelligence to reveal how both can mutually feed into each other throughout my TCE method. This approach empowers the human while still producing an ML model that overcomes the barriers of limited, noisy, variable data.

Third, I describe how this methodological approach feeds back into human intelligence. This method aims not to replace the human but instead to empower the human computational ML. I detail the process in which I combine ML data outputs to complex spaces where human intelligence cannot be computable.

Chapter 5: TCE Methodology for Standardized Clinical Assessment In this chapter, I describe how I adapt the TCE Methodology for use in the development of a standardized computational system for clinical assessment. I showcase how my methodology is generalizable within the area of healthcare. I report my work with nine expert therapists and two stroke survivors in capturing and assessing the Action Reach Arm Test (ARAT) test. Data was captured using a cyber-human hybrid workflow model developed by the larger multidisciplinary team at Virginia Tech. The four therapists used both the Ri and Ci through our in-person and remote studies.

Chapter 6: Discussion In Chapter 6, I describe the findings, insights, and broader implications of my research work and its impact for machine learning. I further describe the qualitative and quantitative feedback given by therapists, expert clinicians, patients, and even our interdisciplinary team. Within the discussion section, I explain the trade-offs, limitations, and future work of the TCE methodology, as well as its contributions to the HCI community.

Chapter 7: Conclusion I conclude my dissertation research in Chapter 7 by describing the broader impact of my findings and future work on understanding human intelligent systems and artificial intelligence. This chapter describes what occurs after the human in embodied spaces reaches new reflections and is empowered through computational intelligence.

Chapter 2

Literature Review

2.1 Literature Review Summary

In this chapter, I explore how my research questions fit into literature and past work conducted by researchers within the intersection of Human-Computer-Interaction and Machine Learning. Additionally, I review prior work performed in global and inclusive health and previous research at the Interactive Neurorehabilitation Lab. I focus my review more intently on understanding the shift of HCI into more complex spaces such as Machine Learning and Artificial Intelligence systems. To better understand this shift, I explore the evolution of the human experience in technology. I then discuss the tradeoffs, benefits, and shortcomings of the research conducted. Finally, through my survey of past literature, I explain my rationale for why I chose my research questions and how the answers and insights from my research questions will further advance the field of HCI for ML.

2.2 Human Experience and Technology

Don Norman formally introduced researchers and the world to the term User Experience [55]. User Experience [79]. By his definition, User Experience (UX) encompasses all areas of design, including physical and mental interactions, design systems, and other experiences that affect human behavior. However, even before Norman, examples of UX began as early

as Hippocrates' work of designing medical devices to Fedrick Winslow Taylor's research on understanding how humans interact with physical systems. Furthermore, companies such as Toyota and Disney also create a human-centered experience by focusing on humans' physical and mental experiences interacting with machine learning systems. With each advancement and progression with User Experience, researchers begin to take a Participatory Design (PD) approach to understand how to best design physical, mental, and interactive systems for the end-user. A PD approach is used for understanding how to create solutions with the end-user top-of-mind throughout the process. It involves bringing together all the stakeholders from the beginning of the design process [24]. With this approach, this design style does not emphasize the physical system. Instead, this approach focuses on understanding each stakeholder within the paradigm, process, and user journey flow to create the best experience within minimal barriers for the end-user(s).

As technology becomes more prominent in everyday life, the primary focus of UX has shifted from simply designing for a system or for one singular experience to now understanding the relationship between how technology and humans can be codependent throughout the journey [73], [27], [67]. Researchers such as [37] and [34] are beginning to discuss an approach that moves beyond making human central. [34] created a framework for analyzing human cognitive thinking for designing human systems. I built upon that framework to understand the different ecosystems needed for each stakeholder to achieve a successful user journey and experience. However, several vital challenges still occurred. As humans become more integrated within HCI systems, there is a growing need to understand how humans interact with Machine Learning applications.

2.3 Human-Computer Interaction for Machine Learning

HCI and ML are polarized in their objectives, approaches, and evaluation to solving complex computational problems. HCI takes a human-centered approach that uses human intelligence from a qualitative and quantitative framework to understand how an iterative approach can provide potential solutions. In contrast, ML processes focus solely on using quantifiable datasets that are standardized, non-variable, and pre-existing. However, authors such as [32], [74], [68] take a human-centric approach to understanding how to combine the processes of ML and HCI to create a process that is more human-oriented. Although these researchers sit at the intersection of these two polarized fields, there is currently no standardized methodological approach that allows for human intelligence to be at the forefront of the paradigm from both perspectives. Though their work produces human-centric, data-driven results, the data produced is often used in the ML learning process. As a result, human intelligence is often forgotten and lost in the training and learning process

Researchers such as [43],[53] use User Interface (UI) design for Interactive Machine Learning (IML). They define IML as Interactive Machine Learning: An interaction paradigm in which a user or user group iteratively trains a model by selecting, labeling, and generating information.[30]

In [37] work, for embodied learning spaces, there is a disconnect between HCI researchers and ML researchers understanding and evaluating human data. Additionally, the datasets and data analysis are assessed on different scales. HCI data evaluations are typically more human-centric, focused on understanding how the data can improve humans throughout the process. On the other hand, ML datasets are usually complex, heterogeneous, and large controlled environments. As a result, the ML approach and framework cannot be replicated

without impeding the human's journey for embodied learning spaces as shown in Fig. 2.1

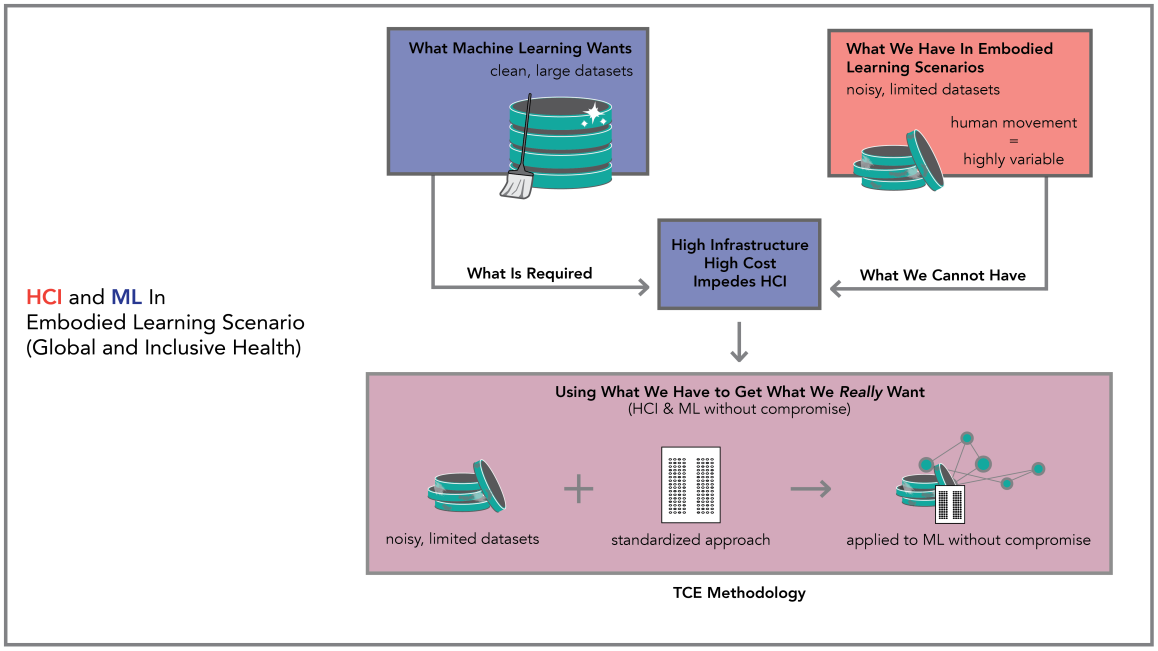

Figure 2.1: HCI and ML for embodied learning scenarios

Machine Learning and HCI Researchers have different goals, methods, and philosophies. For example, HCI emphasizes the end-user through an iterative process. In contrast, ML aims primarily to understand data and people from a quantitative standpoint, thus removing the why behind human decision-making.

Smith et al. and Dumitrache et al. [77], [22] addressed another critical issue HCI applications face for ML systems. As humans provide training data for machine learning systems, humans may be acting or delivering results that are inaccurate or uncertain, leading to training data that is inaccurate or inconsistent.For example, However, a potential solution is crowdsourcing [23], [21], [47] . Dumitrche et al. studied how crowd workers examine and label medical data to showcase distinctions between how humans and machine learning identify and label data. Researchers begin to explore the relationships between human and machine learning

intelligence through this distinction.

2.4 Human in the Loop

Human in the loop has emerged as a pivotal avenue not to replace human intelligence but to integrate human and computational intelligence. An iterative human-in-the-loop approach allows humans to understand their role and objective(s) for complex machine learning applications and systems. The fear, however, is that as these systems evolve, there will no longer be a need for the human within the process [78]. Therefore, a more active approach is needed for humans to ensure that the learning process in embodied spaces is never removed but rather empowered. This challenge is especially critical within global and inclusive health. Clinicians are hesitant to introduce ML systems into health practices, as the fear of being replaced rather than empowered [60].

2.4.1 Participatory Design

Participatory design (PD) situates the design process as a democratic social endeavor where all identified stakeholders are involved in the decision-making process[25]. Following the general evolution of perspectives in the design field, participatory design has expanded in scope and methods, gaining widespread acceptance as an approach to practice and research across various fields [10]. Building on a set of theories and methods first originating in Scandinavia when computers initially began to ”disrupt” the workers environment, PD is now encountered increasingly in a multitude of places, including our target areas of healthcare in the home. Participatory design tradition is defined as a perspective that always looks forward to the shaping of future situations and works for enabling participants to shape

their futures [75]

I chose to use a mixed-methods approach (see Figure 3.1) for three reasons: First, each method fits within the Participatory Design. This approach further allowed me to examine each of the interfaces I developed and the process of understanding humans in an embodied space to enhance design practices and build human knowledge throughout the design process [37].

2.5 Annotation Tools

The development of annotation tools has facilitated a computational approach for understanding and generating data that cyber-human systems can use. Tools such as [70], [43], [19], provide an application that enables people to label different human-movement activities. However, there are two issues with these systems. First, these tools focus on using a crowd-sourcing approach for the collection and annotation of web videos. This issue places a barrier within collecting and assessing human-movement exercise videos from patients due to the sensitivity and privacy of data. Second, each of these tools does not adequately provide a standardized approach to understanding or assessing each of the videos. Furthermore, we see a growing need for collecting and annotation videos that can be used for artificially intelligent agents. Annotation tools have been used to draw labels and focus attention to specific meta-features for understanding for decades[18].

Some authors have developed five possible approaches for annotation classification[31], including classification based on annotation modes, classification by predefined semantic categories, classification based on folksonomies, and classification based on ontologies. By surveying the state-of-the-art annotation tools, we begin to see the limitations and drawbacks of most of these tools. Each of these annotation tools lacks a focus on semantic classification,

which hinders meta-reflective thinking and educational learning. We begin to see the importance of shifting annotation to a broader form, especially within the classification through ontologies.

Reflection is a critical component of rehabilitation. OT and PT assess many patients, and through their assessment, therapists are focused on understanding different ways to produce high-yielding outcomes for patients. One challenge in this process is that since therapists are all trained differently, each has its reflection method that is not explicitly defined. However, Fleck and Fitzpatrick describe a reflective process that first defines reflections and creates a framework for assessing reflection. In their work, they present five different levels of reflection(RO-R4) [76]. Each level of reflection serves a different purpose for the user, and through that representation, a process can be created in which the user's behavior is understood. Our research focuses on translating each level of reflection for OT and PT therapists to reveal their expert knowledge by designing three iterations of the Video Application Tool that express each level of reflection.

2.6 Assistive Technology for Global and Inclusive Health

When creating healthcare systems, researchers [5], [57], [29] primarily focus on making the systems and applications first and then bring the primary stakeholder second in the design process. Previously, this order proved helpful because it allowed researchers to focus on creating effective computational systems that generated accurate results. However, bringing humans in second in the design process creates challenges and a steep learning curve for clinicians and patients in understanding how to integrate their experience with technology. This challenge is more prominent within healthcare, where clinicians and patients struggle to integrate new technologies such as smartwatches and tablets, challenging [56], [6], [2].

In addition, new technologies and systems introduce new complex relationships between stakeholders within the healthcare paradigm in the realm of rehabilitation and assessment.

Technology-assisted rehabilitation therapy has the potential to expand the ability of the therapist to impact patients significantly. There is a diverse variety of promising technology-assisted systems currently in development, including virtual and augmented reality systems [4, 38] robotic assist systems [63], tangible systems, wearable sensors [50], and computer vision techniques [71]. Several of these systems seek to address critical issues, including attempting to replicate the functions of the therapist in an automated system or implementing adaptive rehabilitation through low-cost systems. However, there is a lack of standardized understanding of how and why function changes limit the effectiveness of any approach in long-term adaptive therapy at scale [7, 62]. In addition, the complexity of the assessment problem indicates that technology alone cannot solve it. Instead, it requires a socio-technical approach driven by iterative, human-centered design such as provided by TCE.

Prior work shows that tracking and analyzing low-level kinematics during rehabilitation training in the clinic can result in an automated assessment of performance that correlates highly with expert assessment [13, 52, 81]. However, detailed tracking of movement through marker-based capture [26], or complex exoskeletons [61] is costly, complex, and intrusive. For example, data gloves may not fit those with hand contractures or joint inflammation, especially among stroke survivors (the population focused on in this study), and marks on hands can hinder the performance of detailed functional activities [36]. Furthermore, different low-level features are known to work for different types of movement quality assessment, and a combination of various measuring sensors may be needed to acquire accurate features. Finally, complex technological systems are challenging for therapists to use and are rarely part of their educational curriculum. Thus, they are not widely adopted. The TCE approach redresses this problem.

For several years, the relatively inexpensive Kinect camera tracking system supported multiple promising rehabilitation systems[65, 82] but with the manufacturing of the Kinect discontinued, together with other use-case criticisms, alternative low-cost movement capture approaches are needed. Tracking performance through a small array of video cameras (up to two cameras), as in the TCE approach discussed in this dissertation, is low cost, low effort, and unobtrusive. Still, the Kinect does not produce reliable tracking of all necessary low-level kinematics [65]. Similar issues with tracking reliability are shown in some tangible systems where each patient's unique movement profile creates challenges in detecting movement and counting task repetitions. Therefore, trade offs are required in developing a motion capture system that is effective, relatively inexpensive, and straightforward to install and operate.

Facing these different challenges, researchers have focused their attention on understanding and trying to restructure the current paradigm in healthcare into a framework in which technology and humans are codependent. The outcome of their research [3] proposed a paradigm of re-conceptualizing rehabilitation and placed the priority on creating digital tools and systems in place to promote growth, learning, and comfort for patients. There are two challenges with this approach. First, it negates the importance of clinicians for improving patient recovery. Second, the ML systems are not integrated to facilitate continuous growth and reflective thinking to all stakeholders [17].

2.7 Summary of Related Work

Although researchers have examined cyber-human frameworks for in-home therapy and clinical systems, none of the methods provides a solution in which the human synergizes with the ML systems and is empowered throughout the computational process. The trade offs are either that the clinician uses human intelligence to train the ML system and is then

faced with the possibility of being replaced, or the human learning reaches a plateau. The biggest challenge with understanding and computing tacit knowledge is the absence of data to collect and share,which is critical in a healthcare setting. Through frameworks such as [76] [72]. These framework present a qualitative way in which humans can understand their tacit knowledge and potentially turn that tacit knowledge into explicit knowledge through reflection. I leverage this approach for assessing reflection and provide a standardized process to understand tacit knowledge qualitatively while simultaneously transforming tacit knowledge into computable, explicit knowledge.

My dissertation offers a participatory design approach that leverages the needs of each stakeholder (patient, therapist, machine learning agent) to understand how to design an HCI methodological framework that aims to create a nonlinear, non-hierarchical paradigm for complex machine learning systems. This process allows for an asynchronous, standardized approach for capturing and assessing human data and knowledge for embodied spaces in the wild. ML has typically stayed away from HCI principles due to limited, noisy, variable datasets in embodied scenarios. However, through this dissertation, I contribute to the field of both HCI and ML by first developing a method0logical approach for understanding the relationship between human and machine intelligence for complex embodied spaces. Second, creating design process which encourages reflection for the end-user while increasing inter-rater reliability in complex noisy spaces. Finally, I create a standardized approach for both capturing and assessing human movement data for embodied scenarios.

Chapter 3

Methods

I designed and created the Tacit Computable Empowerment (TCE) methodology using a participatory design approach. Participatory design ensures the inclusion of all fundamental stakeholders (in this case patients, therapists, and machine learning developers/agents) in the design and decision making process.

I used a mixed-methods approach including remote one-to-one interviews, diary studies, quantitative task analysis, statistical analysis, and focus group discussions with all stakeholders (patients, therapists, ML developers/agents) [11]. Mixed methods approaches are well established in the fields of social science [28] and increasingly in HCI and healthcare [9], making the approach suitable for this inquiry context. Outcomes from qualitative inquiries informed the computational and quantitative approach developed by collaborating team members, and vice versa.

Figure 3.1: The various mixed-methods used with stakeholders throughout my dissertation.

Therapist	Role	Rating Model
Therapist 1	Occupational Therapist	VAT: F2MQ
Therapist 2	Occupational Therapist	VAT: F2MQ
Therapist 3	Occupational Therapist	VAT: F2MQ
Therapist 4	Physical Therapist	VAT: F2MQ, MQ2F, SDP
Therapist 5	Occupational Therapist	VAT: MQ2F, SDP
Therapist 6	Occupational Therapist	ARAT : SDP
Therapist 7	Occupational Therapist	ARAT : SDP
Therapist 8	Occupational Therapist	ARAT : SDP
Therapist 9	Occupational Therapist	ARAT : SDP

Table 3.1: The therapists involved in each iteration of our rating process their role in rehabilitation (Occupational Therapist (OT) or Physical Therapist (PT)).

Table 3.1 displays each of the qualitative and quantitative methods used throughout my dissertation and the varying tools and stakeholders involved.

3.1 Data Collection and Assessment

The two primary systems described in this dissertation are the Semi Automated Rehabilitation At Home (SARAH) system and the Action Research Arm Test (ARAT) Assessment system. Each system comprises two novel tools for capturing and rating human movement in rehabilitation training (SARAH) and assessment (ARAT) contexts. Chapters 4 and 5 describe the design methodology and computational development of these individual tools in detail and their shared goals of helping therapists standardize a movement assessment approach that relates functionality to movement quality.

For the SARAH system, we collected videos of nine stroke survivors performing 12 upper extremity generalizable training tasks that we established through our pilot research [65]. This study was conducted at Emory Rehabilitation Hospital in Atlanta Georgia in Fall 2017 and Spring 2018, and the study was approved by the Institutional Review Board (IRB) at

that institution. Participants in our studies were recruited through physical flyers posted in local hospitals and university training facilities, emails to relevant mailing lists, social media postings, and word of mouth via the clinicians on the team. The team compensated participants for their time with a $50 gift card). Two of the participants were categorized as moderate impairment (Fugl-Meyer score between 30 - 55), while seven participants presented with mild to moderate impairment after scoring (Fugl-Meyer score of greater than 55). The nine participants also had different specific movement challenges thus providing an even more varied data set. Each session was supervised by a physical therapist and/or a rehabilitation expert from the Emory Hospital. Each patient was asked to attempt each of the 12 tasks in the SARAH system four times each. The more severely impaired patients could not complete four iterations. The majority of patients could not perform the harder tasks (tasks 11 and 12). Three of the nine patients were asked to return for two repeat sessions so as to allow the therapist to explore differences within data sets of the same patients. The movement of the patients was recorded using two low-cost web cameras - one from a sagittal perspective of the torso and shoulders and one from a frontal view capturing arm and hand movement. In total, 15 data capture sessions produced 618 distinct activity attempts by the nine participants.

This SARAH data was assessed and labeled by 13 therapists in three in-person and online studies between 2018 - 2020 using three different iterations of our Video Application Tool (VAT). The assessment tools and the associated rating rubric were iteratively refined between studies based on insights from quantitative analysis of rating sessions, four follow up surveys, and twelve extended interviews and focus groups with the rating therapists [40].

For the ARAT system, we collected videos of two stroke survivors performing up to 19 of the upper extremity activities in the standardized ARAT movement assessment test. This pilot study was conducted at the Shirley Ryan Ability Lab (SRAL) in Chicago, Illinois in November, 2021, and was approved by the SRAL ethics research board. The participating

stroke survivors were invited to participate by our SRAL colleagues, in addition to the four active therapists. One participant was categorized as mild and had experienced two strokes in the previous 18 months. The other participant was seven years post-stroke and was classified as moderate. The data capture sessions were conducted by four therapists from the SRAL. The movement of the patients was recorded using four high resolution cameras (transverse, saggital-left, saggital-right, and back). In total 38 high resolution videos were captured. A larger study scheduled for January 2022 was postponed because of COVID-19 concerns at the SRAL.

This ARAT data was assessed and labeled by four therapists in one in-person (November 2021) and one online study (January, 2022) using an adapted version of the third iteration of the VAT. The therapists labeled 20 videos from the two patient study. Two in-person focus group sessions and one online interview session were carried out to capture additional feedback from participants on the pilot experience.

Chapter 4

TCE Methodology for Home Based Rehabilitation Training

4.0.1 SARAH System

The Semi Automated Rehabilitation At Home (SARAH) system is a collaborative research initiative led by the Interactive Neurorehabilitation (INR) Lab at Virginia Tech. As a member of the INR lab since 2017, my dissertation represents a significant research and application outcome for the collective. Early versions of the SARAH system are extensively described in my Virginia Tech Master's Thesis "Interactive Interfaces for Capturing and Annotating Videos of Human Movement Performance", which was awarded the 2019 William Preston Outstanding Thesis of the Year award for Virginia Tech. In addition, insights from this work have been published at the Conference on Progress in Clinical Motor Control: Neurorehabilitation and at ACM PETRA [87] , [88]. A summary of the early development, use, and pilot evaluation of the SARAH system is described below. Findings and outcomes from this earlier work influenced the development of the TCE methodology that forms a fundamental contribution of my dissertation.

The Semi Automated Rehabilitation At Home (SARAH) system combines computational intelligence, therapist expertise, and active patient engagement within a cyber human framework [64] aimed at improving human and cyber intelligence. The SARAH system can be

understood as operating within a broad and adaptive ecosystem, which contains diverse stakeholders, technologies, processes, resources, and environments [27]. The INR lab has conducted embedded ethnographic work with rehabilitation experts specializing in the therapy of the upper extremity of stroke survivors for over 12 years [12], [48] , [49], [39]. This observational work, conducted in hospitals, clinics, and research labs across multiple different projects, revealed that therapists use a limited set of training exercises and a limited set of training objects during therapy (e.g. reach for a cup, fold a towel, brush hair etc.). These exercises and objects are typically structured so as to generalize to many activities of daily living (ADLs).

In developing the SARAH system, the rehabilitation experts on our team proposed 12 movement activities ranging from reach an object; to reach and grasp an object; to reach and grasp and transport an object(s); and finally to reach and grasp and manipulate an object(s). The 12 activities were designed to map to important activities of daily living (ADLs) including eating, drinking, putting on clothes, personal hygiene, and typical household chores. Table 1 lists the ADL mappings for each of the training activities, which increase in complexity from 1 – 12. Fig 4.1 depicts the full range of objects, as well their functionality.

As shown in Fig 4.2, the SARAH system comprises two video cameras, a tablet computer, a flexible activity mat, and eight custom-designed 3D-printed objects. The two camera setup connects to a tablet computer that captures activities from angles representing where the therapists on our team typically like to stand or sit during therapy. One video camera positioned beside the participant captures a sagittal side-view of the shoulders and torso, while a table-mounted video camera focuses on the wrist and fingers during manipulation and transportation activities. The tablet is placed on a custom designed mat, with etchings on it denoting the range of the activity space. I led the development of an intuitive capture interface to deliver the activity protocol on the tablet to the patient and initiate recording

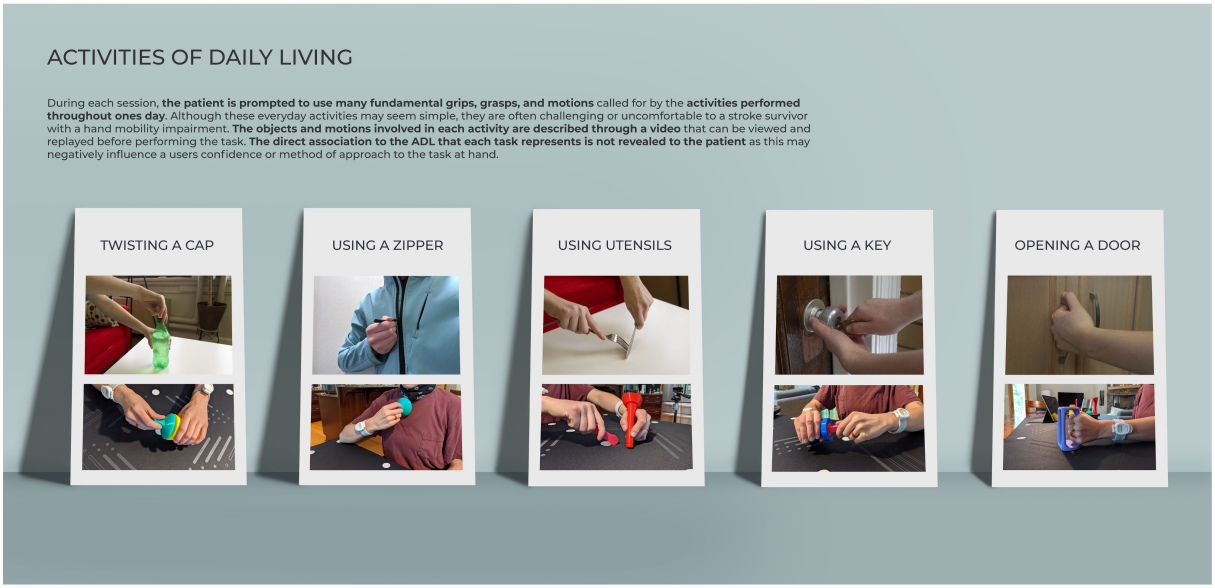

Figure 4.1: Activities of Daily Living (ADL) which represent performing movements using 3D objects that mirror activities such openining a door, or turning a key

on the cameras.

Figure 4.2: The Sarah System is an in-home based rehabilitation system which consists of a two-camera setup, tablet interface, a customizable activity space, and 3D-printed objects

I co-designed and evaluated the patient-centric movement capture interface with another researcher on our interdisciplinary team [15]. Fig 4.3 displays the patient's workflow process using this interface. The goal of the interface is to mirror the "voice" of a clinical therapist in a simulated rehabilitation session. The design and capture workflow process included collaboration between different parties in our team ranging from UX/UI designers, industrial

designers, occupational therapists, computer vision/machine learning experts, and computer scientists. Our primary objective in creating and (subsequently) standardizing this process, was to create a straight-forward, user friendly approach that allowed for high resolution video capture with little to no variance, while still not interrupting the workflow of the patient or therapist.

The most recent version of the SARAH system Fig 4.2 is optimized for daily at home therapy of stroke survivors with moderate and moderate-to-mild impairment. Stroke survivors who score above 30 on the Fugl Meyer test and can initiate even minimal movement into extension of the elbow, wrist and digits when discharged from the clinic post stroke, can show significant improvement in function through repetitive training lasting 2 weeks to 6 months (27– 29). Currently, third party insurance in the US provides financial support for up to 6 months of outpatient and/or in home therapy after release from the clinic. Patients with moderate impairment and a Mini Mental score >25 can follow instructions given by the SARAH system, use their unimpaired limb to control the SARAH system and can engage (at least partially) all SARAH objects (25). The SARAH system promotes active learning by the patient who is expected to interpret the coarse feedback provided by the system and to plan the next action so as to improve her performance. Furthermore, the generalizable design of the objects and tasks of SARAH encourages the patient to actively map the task to multiple ADLs. These active learning characteristics of the SARAH system, combined with the variety of included tasks in the system makes the system feasible for in-home rehabilitation lasting 2–8 weeks (25, 26).

SARAH Pilot Study

We installed the first version of the SARAH system [65], [40] at Emory Rehabilitation Hospital, where we recorded videos of nine stroke survivors(seven men and two women) per-

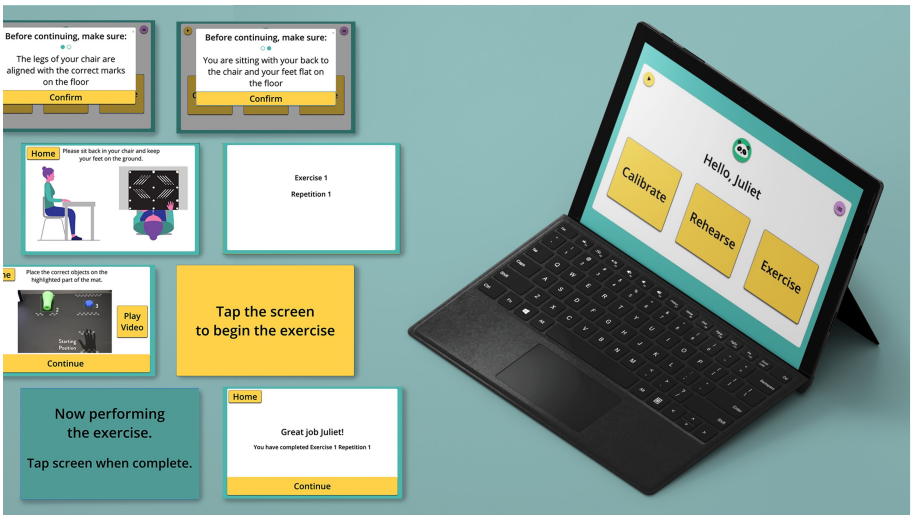

Figure 4.3: The capture interface is an assistive interface design to aid patients in performing rehabilitation exercises. This also allows the capture of high quality, standardized video data of patients performing each exercise through the use of calibration and patient setup process.

forming the first set of 12 SARAH tasks (see Methods section for more details). The nine stoke survivors had different levels of impairment ranging from mild/moderate to moderate impairment and had different types of movement challenges. The participants in the study were asked to attempt each of the first 12 SARAH tasks, repeating each task four times if possible. Most patients could not perform the final two of the 12 SARAH tasks since the last two are the most difficult.

The primary objectives of this first pilot study were 1) to gain insights and understand the implications of using the interactive movement capture tool from the patient and therapist perspective, particularly with respect to the relationship between the SARAH tasks and general activities of daily living; and 2) to capture data of stroke survivors performing each of the different activities to better understand various components of their movements, and to provide data to the machine learning experts on the INR team responsible for developing the computer vision and AI automated assessment and feedback components of the SARAH system.

Six of the participants were able to complete all four repetitions of each of the twelve activities, while three of the more moderately impaired participants struggled to complete the more complex movement activities. Different patient profiles (e.g. level of impaired sensation, increased spasticity, limitations in shoulder range of motion etc.) influenced the abilities of the participants to engage with some of the objects. In those cases, the therapist intervened in the recording session, and used the menu feature in the application to move to a different activity. Three sessions were also not fully completed because the capture process took longer than anticipated and the participants had to leave. Overall, we recorded 618 distinct activity attempts by the nine participants.

The participants had different tolerances for the duration of the training exercises with some of the milder impaired patients (P1, P2) indicating frustration with even the smallest delays in their execution of the task (as compared to what an unaffected limb could complete). Other participants (P3 and P8) persisted in completing the task, even if their ultimate execution time would be considered significantly too long by the therapist. Indeed, P3 noted that they enjoyed having no set task completion time, explaining "I wasn't being pressured by time. That was good. That was the best part about it really. If there was a time thing, I would have got nervous and stuff probably would have started falling on the floor".

All nine participants agreed that the training activities reminded them of activities they encountered in their daily lives, with two participants stating that they had done very similar movements earlier in the day. Six of the participants wanted to receive more detailed feedback from the system about their movement performance. Two of the participants were skeptical about the possible accuracy of the feedback, with both candidates expressing their displeasure (e.g. P8: "Nice try? Bullshit. I did it") or their ambivalence (e.g. P1: "Good job. Who knows?"). The therapist intervened on multiple occasions to assure the patient that their performance was better than the system had assessed. For example, after P6 received

several "nice try" assessments after doing the exercise, the therapist carefully stated: "So, before we move on there. It's saying "nice try" there (points at the interface) and part of what we're trying to align is that you're actually doing a great job of not leaning with your trunk which is exactly what you should be doing. So, it actually should be telling you good job or excellent".

The participants offered a rich variety of feedback and critique to the development team regarding additional activities to consider (e.g. more above the shoulder activities, and more bimanual tasks) and additional objects to include (e.g. eating utensils, keys and locks, personal hygiene artifacts). Several of the participants, particularly those with milder physical impairment, noted that while the activities were not physically challenging, they were cognitively challenging. P7 stated "It didn't challenge my strength but it did challenge my pea brain (laughs). Remembering what to do was - the further we got into it, the more I had to think, which is real good for me. It's the deeper I got into it, the more it went from the physical to the brain."

Based on our team's observations and the insights shared by the patients and the therapists, the iterative design and development of the SARAH objects and mat continued after this study and is described in an ACM TEI publication, led by the industrial design members in the INR lab [39]. The iterative design and development of the calibration interface, the patient interaction interface, and the physical setup of the SARAH system are also described in more depth in publications by our team including Juliet Clark's Virginia [14] Tech Masters' Thesis and an ACM IMX publication [15].

In order to satisfy the second goal of the pilot study - namely to collect human movement data to assist in the development of computational algorithms for a semi-automated system - I now needed to develop a process to better understand how therapists made decisions when observing and assessing patient movement. This required the development of an assessment

rubric and a rating approach that could annotated the captured data in a machine understandable way. This lead to the development of a key technical innovation in my dissertation - the Video Application Tool.

4.0.2 Video Application Tool

This section describes the three iterations and rating models assessed with expert clinicians. From a technical standpoint the Video Application Tool (VAT) was developed by using HTML, CSS, and Javascript. All patient information was stored on our encrypted Mongo Database and encrypted hard drives. The scores from each assessment were transferred, asynchronously, after each session for data processing and analysis. Finally, the processed scores and analysis were sent to the HBM model for computational analysis. In the following sections, I will present the results of each iteration of the rating interface. The primary objective within this phase is to understand through the use of the VAT how to transform clinician tacit knowledge into explicit knowledge.

Function to Movement Quality (F2MQ)

The first iteration of the tool presents our first attempt at understanding how therapists approach the relation of function to movement quality in rehabilitation movement. In our co-design with five expert therapists, each therapist described that a function to movement quality sequence would best resemble their assessment framework. The score of a task denotes the overall assessment of functionality while the movement features denoted as impaired by the therapist for each segment provide a more specific interpretive rationale detailing how and for what reason the therapist arrived at their overall score. Thus, the VAT first presents the therapist with videos depicting the entirety of the patient performing

Figure 4.4: The video annotation tool presents the therapist with both sagittal- and front-captured videos of the stroke survivors/patients attempting one of the 12 ADL tasks, in addition to an instruction video of an unimpaired person doing that particular ADL, which the patients had viewed before attempting the task. The goal of the tool is to assist the therapist in evaluating the quality of the patient movement performance.

the task. Next, the therapist must assign a discrete score and feature interpretation for each movement segment (e.g., initiate, progress, and terminate). The therapist moves in a linear progression scoring each of the segments, and they cannot move on from a segment until the rating for that segment is complete. Once they finish rating each of the segments, they are once again presented with the videos depicting the entirety of the patient performing the task, which they must rate again. This feature was implemented at the request of the rubric and rating assessment development team as they were interested in discovering if the process of reflecting on and assessing the movement in discrete segments might compel the therapist to change their overall score from their initial first assessment. Two therapists rated 72 tasks using this version of the VAT.

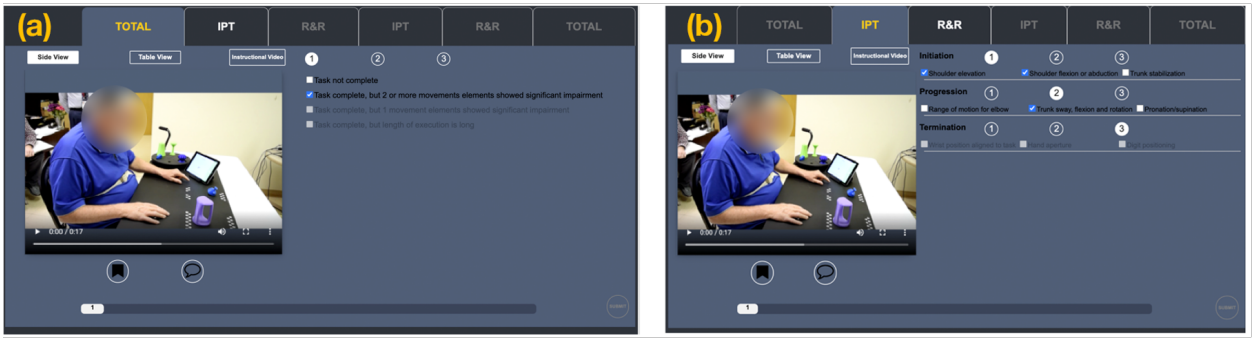

Figure 4.5: VAT function to movement quality interface. (a) Overall video assigned rating of 1, with interpretive feature "Task complete, but 2 or more movement elements showed significant impairment; (b) Initiation segment video assigned rating of 1 with interpretive features"shoulder elevation" and "shoulder flexion or abduction" selected, while Progression is rated 2 with interpretive feature "Trunk sway,flexion and rotation" selected. Termination segment is rated 3 with no features selected

The VAT "F2MQ" model interface is displayed in Fig 4.5, depicting what the therapist sees when viewing the overall ADL task videos. They can choose to view the instruction video or the side or table view in the large video panel. The top of the screen presents the linear series of tabs that the therapist selects to move from the overall/total activity video through the four segments comprising the complete activity before returning to the overall/total activity

video again. When rating tasks the therapists selects one of the rating buttons to provide an overall score. When rating segments, the therapist selects one of the rating buttons and then denotes composite movement features as impaired (using the checkbox next to each featured) to provide interpretive (movement quality) context for the segment rating. In addition, the therapists can use the "comment" button beneath the main video panel to further annotate their assessment with commentary about the specific video they are viewing. Similarly, the therapists can use the "flag" button beside it to send a message to our development team regarding a technical problem with the interface or an issue with the displayed video. Finally, the progress bar at the bottom of the screen displays where they are within a typical rating "session."

The rubric and rating assessment development team gave us clear guidance about the approximate time it would take to complete the rating and interpretation of one ADL video. They assessed that a skilled therapist could complete a full ADL video (overall movement and all movement segments) in no more than five minutes. They agreed that a therapist would most likely reliably (and enjoyably) rate no more than 12 full ADL videos in one sitting/session. Thus, the interface allows the therapist to move through 12 full videos at a time per session, although they have the option of beginning a new 12 video session upon completion if they wish.

Movement Quality to Function (MQ2F)

The second iteration of the tool and our rating model begins to challenge the therapist's methods of assessing rehabilitation movement while exploring further the conditioning effect of movement quality observations on functionality assessment. In this version, the overall segment and task scores are computationally generated based on the composite movement feature impairment observations of the therapist. We customized the interface for this exper-

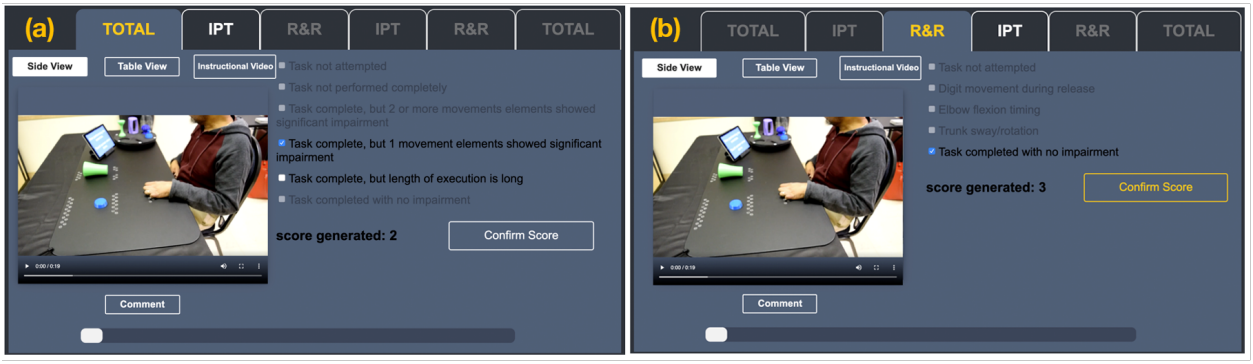

Figure 4.6: VAT movement quality to function interface. (a) Overall video assigned rating of 1, with interpretive feature "Task complete, but 2 or more movement elements showed significant impairment; (b) Initiation segment video assigned rating of 1 with interpretive features"shoulder elevation" and "shoulder flexion or abduction" selected, while Progression is rated 2 with interpretive feature "Trunk sway,flexion and rotation" selected. Termination segment is rated 3 with no features selected.

iment though a co-design process with three therapists. We moved the main video panel to the left of the screen with the video perspective selection buttons placed above. This change opened additional screen real-estate to the right where we could more carefully and prominently display the composite movement features for therapists to focus more intently. We also removed the technical flag button as the therapists preferred to email/text their issues or wait for an in-person meeting to discuss. Fig 4.6 below displays the interface presented to therapists when asked to rate the overall/total patient performance.

When assessing a segment, a therapist is first asked to check the composite movement features that show a level of impairment that can influence function/task execution. If only one feature is checked as impaired then a segment score of 2 is automatically generated. If 2 or more features are checked as impaired then a segment score of 1 is automatically generated. The therapist can influence the segment score by denoting more or less composite movement features as impaired. When training therapist to use this version of the rating interface we cycled each therapist through the relationships between the meaning of each movement feature and its relationship to the score generated by the VAT. Figure XX displays the

interface presented to therapists when asked to rate the Release and Return movement segment of a task. Two therapists rated 185 tasks for this version of the tool.

After rating all segments, the therapist moves on to rate the overall patient performance of the task. Here the therapist can select one of six possible interpretations: "Task not attempted" to "Task completed without impairment."

1. Initiation
2. Progression
3. Termination
4. Manipulation (Bimanual) Transport
5. Complex Bimanual Manipulation
6. Release and Return

These six phrases were co-designed with expert therapists so as to assist the therapist in connecting their functionality assessment, denote through the task performance rating, to movement quality issues identified at the segments level. In the situation depicted in Fig 4.5, the therapist has selected "Task complete, but one movement element showed significant impairment," which generates a score of 2. The feature "Task complete, but the length of execution is long" is also bolded in this situation as it is the one other option that could also generate a score of 2. If the therapist agrees with the score generated based on their interpretive assessment, they can select the confirm button and move to the next segment. It is important to note here that this version of the interface introduces the possibility of a performance score of zero. We had overlooked this in our previous version of the rubric/assessment rating/annotation tool combined. During the presentation session with

the clinical team at the end of iteration one, it became clear that performance events were not well described using our 1 - 3 rating schema, and non-performance needed to be considered.

Structured Decision Process (SDP)

In our final iteration, the VAT guides the rating therapist though a structured decision tree process. The binary decision points and their sequence were again co-designed with expert therapists and informed by validated clinical measures for the standardized assessment of upper extremity therapy of stroke survivors [86]. The process aims to reveal underlying assumptions made by therapists when assessing movement and help the rating therapists explicitly connect assessment of functionality and movement quality. Additionally, in this iteration, we wanted to analyze if the process of using the VAT helped each therapist reflect on their practice beyond the use of the VAT.

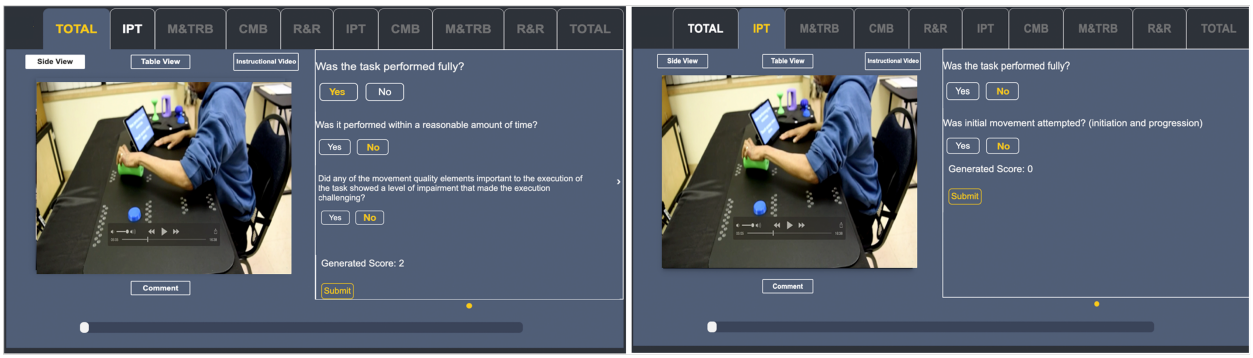

Figure 4.7: VAT structured decision process interface. (a) Overall video generated a score of 2 after therapist answer yes to task performed fully, and no to both whether it was performed within a reasonable time or if any of the movement qualities made the task execution challenging. (b) IPT segment generates a score of 3 after therapist selects "Task completed with no impairment"

Therapists were asked to answer binary yes/no questions at the level of the task and segment: successful completion of task/segment, within reasonable amount of time, without movement impairment, check impaired composite features. A yes answer to the first two questions

automatically assigned a score of 3 to the task/segment. A yes to the first, with a no to the second provided a 2. A yes to the first, with a no to the third provided a 2 and opened up the annotation interface for composite features where at least one composite feature needed to be checked as impaired. A no to the 1st question provided a score of 1 and opened up a yes/no question regarding completion of the initiation stage. If a no was provided to the initiation the score became a 0. Fig 4.7 shows an example. The therapist is first asked if the task has been performed fully. A positive answer generates the next question, which inquires if it was performed in a reasonable amount of time. In this instance, the therapist answers no, which prompts a question as to whether any movement quality element might have impacted the task execution. Again the therapist answers no to this question which generates an overall score of 2 for the task. Figure 3b depicts the "get-to-the-point" outcome from the questions about the IPT movement segment further on in the activity. Again, the therapist answered negatively to both preliminary questions, resulting in a generated score of 0. With this iteration, we allowed the interface to give the therapists more freedom in assessing the rating as we did in iteration 1 but still generated the score to keep a level of consistency for the therapist. Three therapists rated 150 tasks for the SDP version of the tool.

4.0.3 Movement Taxonomy and Rating Rubric

The development of the movement taxonomy and assessment rating rubric is described more fully in my Virginia Tech Masters Thesis, and related prior publications at CHI in 2020 [40]. The full taxonomy and rubric is presented also in Appendix A. Briefly summarized, our team of physiatrists and four collaborating therapists (two OT, two PT) developed an initially limited set of five movement segments that together represent the sum of all possible movements within the 12 ADLs attempted by the stroke survivors video database. These

segments relate to activities including:

1. Initiate, progress, and terminate (three segments in one)
2. Manipulate and transport
3. Manipulate and bimanual transport
4. Complex bimanual manipulation
5. Release and return

Segments 1 and 5 describe movements where the patient either reaches their arm and hand out and away from the body or brings their arm and hand back towards the body. Segments 2 and 3 describe movements where the patient picks up and moves one or two objects. In comparison, Segment 4 relates to movements where the patient combines two objects in a complex set of movements, such as screwing two objects together. Having defined the taxonomy, the team next proposed an initial draft of the assessment rubric, identifying what they considered to be the most significant movement features for each of the movement segments (no more than four to reduce confusion). The segments and the movement features are depicted in Fig 4.9).

The clinical team also developed a corresponding ruleset to rate the performance of each ADL movement segment and the overall movement of the entire ADL. The segments and overall ADL can be stored in a simple 0 to 3 rating, with 0 meaning ”movement was not attempted” and three meaning ”movement was completed (fully with multiple attempts as needed) within a reasonable time and with minimal movement impairment.” Figure three depicts the movement segment and overall movement rating scales. While 0 - 3 represents the current rating scale, our first version of the tool used a 1 - 3 rating scale, a difference

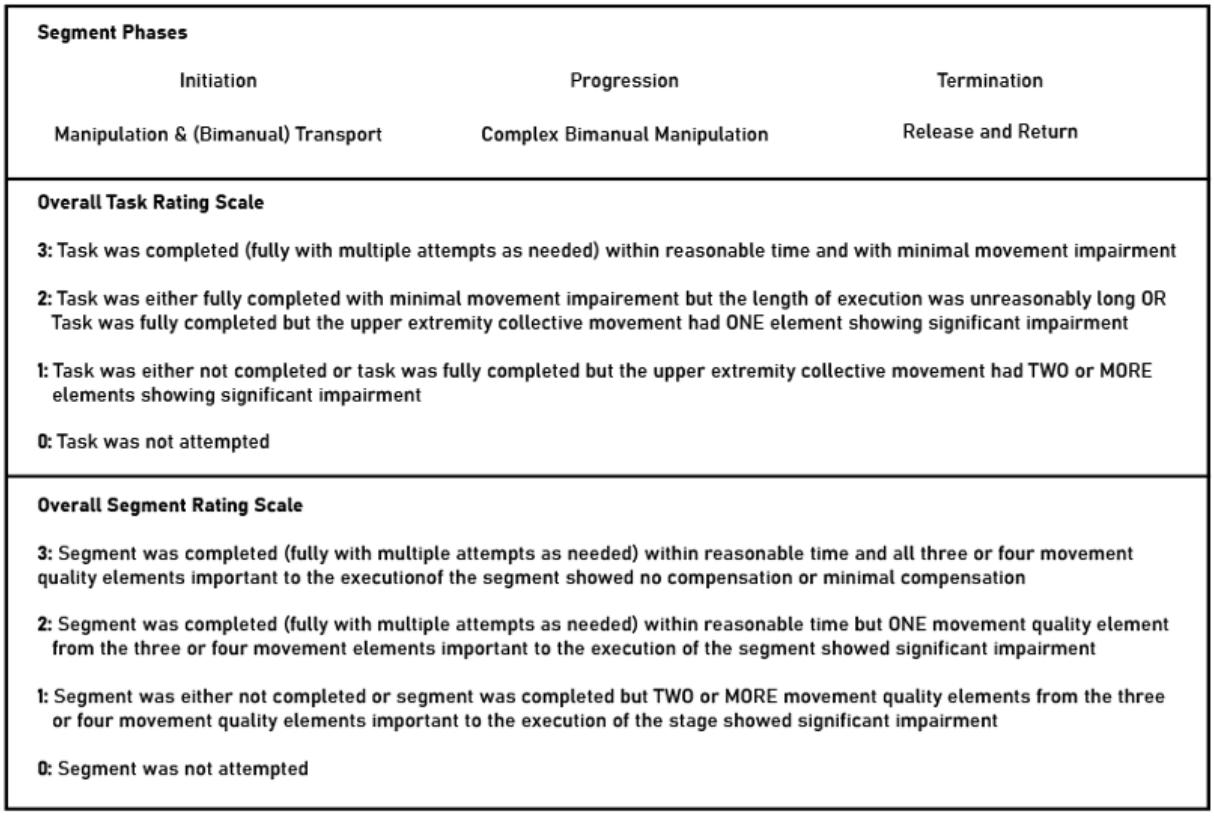

Figure 4.8: The segments relate to the following movements 1) Initiate,progress and terminate; 2) Manipulate and transport; 3) Manipulate and bimanual transport; 4) Complex bimanual manipulation; and 5) Release and return. Segments 1 and 5 describe movements where the patient either reaches their arm and hand out and away from the body, or brings their arm and hand back towards the body. Segments 2 and 3 describe movements where the patient picks up and moves one either one or two objects respectively, while segment 4 relates to movements where the patient combines two objects in a complex set of movements, such as screwing two objects together.

Key SARAH components/movement features to be reviewed per movement segment and rating score definitions and use

MOVEMENT SEGMENTS

- *Initiation*
  - Inappropriate shoulder elevation
  - Inappropriate shoulder flexion or abduction
  - Inappropriate trunk stabilization
- *Progression*
  - Inappropriate range of motion for elbow
  - Inappropriate trunk sway, flexion and rotation
  - Inappropriate pronation/supination
- *Termination*
  - Wrist position not aligned to task
  - Inappropriate hand aperture
  - Inappropriate digit positioning

- *Manipulation & Transportation (MTR) or Manipulation & Bimanual Transportation (MTRB)*
  - Inappropriate finger positioning and orientation
  - Inappropriate limb trajectory (smoothness and accuracy)
  - Inappropriate limb orientation (supination/pronation)
  - Inappropriate trunk movement and position

- *Complex Manipulation & Transportation (CMTR) or Complex Bimanual Manipulation (CMB)*
  - Inappropriate finger positioning
  - Inappropriate finger motion after positioning
  - Inappropriate limb motion following finger positioning
  - Inappropriate limb trajectory (accuracy)

- *Release and Return (R&R)*
  - Inappropriate digit movement during release
  - Elbow flexion timing not reasonable
  - Inappropriate trunk sway/rotation

Figure 4.9: Five SARAH segments and corresponding significant movement features per segment"

that we explain in more detail in section 4.2. In the following sections, we introduce the results from rating sessions using different versions of the Video Annotation Tool by teams of therapists.

4.0.4 Video Application Tool Results

In the following section, I describe each of the qualitative and quantitative results from using the Video Application tool across the three iterations of the system. First I describe qualitatively how VAT enabled levels of reflection for the therapists across each iteration. Then, I describe the quantitative results per the Inter-Rater Reliability, Rater Consistency, and Time Per Session.

Workshop, Interview and Focus Group Outcomes

The results from our first VAT pilot rating study with four therapists using the Function to Movement Quality (F2MQ) interface are described in-depth in the CHI 2020 paper [40] and in my Masters Thesis, but key points are surfaced in this section. Commentary on the patient experience is noted above. From the therapists perspective, it is important to note that over a four week period, four therapists used the first version of the VAT tool to rate 72 patient activities for a total of 240 ratings. While the quantitative data is presented in detail below, we also observed several items of note. It is possible to discern the signature styles of the therapists from the rating data. T1 rated the movement quality in a relatively lenient way compared to the other therapists. The ratings by T1 were higher than the mode rating for an activity in nine of the 48 instances the therapists were asked to rate. T2, the therapist with the most experience, was very consistent in their rating and disagreed with the mode in only two of the 48 activities. Similarly, T3, the next most experienced therapist

only disagreed with the mode in four of the 48 instances. T4 was the youngest therapist and their ratings were more mixed, demonstrating variance from the mode both over and under on multiple occasions.

To better understand situations where therapists disagreed, we created a multi-part online survey for the four therapists and our rehabilitation expert, with sections on rating patient/activity order, impact of patient impairment profiles, interpretation of the disagreement in rating scores, and impact of therapist training and experience. The survey also presented the therapists with examples of the split (2X2) and 1, 2, 3 inter-rater disagreements for comment. The therapists were asked if they expected or were surprised about the increased level of disagreement when rating segments and when rating different participants, and asked for suggestions as to how to optimize the order and presentation of the activity videos for improving inter-rater reliability.

T1 noted

> "It was difficult and somewhat unnatural as a therapist to break down movement to the degree that we were, especially between the initiation of movement to the progression of movement. It is very difficult to determine when initiation stops and progression begins, so it would not surprise me to have much variability there."

The challenge of rating segments was corroborated by similar comments by the other therapists. This expert input, along with the observed drop in inter-rater consistency at the segment level, confirms a core assessment challenge presented in our introduction. Even expert therapists are not able to consistently observe all detailed movement parameters of an activity performance [84].

In the section on patient impairment profiles, the respondents all agreed that the participants

in the mild to moderate impairment range (as opposed to only mild or only moderate) were the most challenging to evaluate, with T1 noting that

> "the patients in the middle are always the hardest to judge because their movement patterns are likely more varied. In other words, a person may be considered moderately impaired overall, but will have elements of mild and severe movements intermixed."

These comments conform with patterns observed in health related ratings and expert ratings in other fields. "Easy" instances are the ones at the edges of a continuum (not impaired, very impaired) and thus further away from decision boundaries [42], [45], [46].

For our second VAT rating study, we invited two occupational therapists (one from iteration one) to use a more refined version of VAT to rate a further 175 videos over three months. One therapist had approximately 14 years of experience, while the other had over 40 years. The study took place over three months and the therapists completed the studies outside of their normal work hours. They were compensated at a rate of $50 an hour. It was challenging to get the therapists to complete their work in a timely manner as other work and life commitments naturally interfered. For example, at one stage, T1 noted:

> "We've had a little trauma at home– one of our dogs passed away unexpectedly yesterday, so I hope my focus is still OK. If you notice any major discrepancies between this session and others, that may have something to do with it."

Having used the tool for an extended period, T1 also noted how the instrumentation of the rubric through the interface helped them reflect on their assessment and rating approach, both in practice and as experienced in training:

> So, but breaking it down into very small parts, like the tool does it does it make

> you look at the components of a movement more closely, whereas I might can tend to just look Look at the entire movement. And then well how functional did that work? You know, is that really a useful thing for that person to be able to do, as opposed to? Well, you know, he's lacking 15 degrees, flexion 10 degrees abduction, which is what you learned to do in school, you know, these minute little things. But as time goes on, and you're actually a clinician, you tend to look at overall function as opposed to each individual component."

This version of the tool used the Movement Quality to Function (MQ2F) interface, and T2 noted how it was easier for them to use this assessment approach than having to come up with a score themselves and then rationalize:

> But I like to look at it and click off the the impairments. And then I'd look at it again. And then I'd look at it again and make changes if I need to do..(it) is not very often that I would change something because they didn't get better. But um, then just having the score come up automatically was easier than me trying to decide... I mean, it was easier for me to pick out the impairments and have the score pop up, then pick a score and then try to fill it up with stuff."

Based on the input, results analysis, and recommendations from the participating therapists in study two, we developed the Structured Design Process version of our VAT interface, which we subsequently assessed in a third pilot observational and interview study with one occupational therapist. This therapist rated 136 videos over a three month period.

Inter Rater Reliability

Table 4.10 showcases the Inter-Rater Reliability for the task, segment, and composite feature sets across the three inteface studies. The score distribution from each study is also included

below. Our goal is to have high inter-rater consistency across all three levels (IRR>60%).

Study 1: Function to Movement Quality (F2MQ) We produced a good IRR by definition of the Cohen Coefficient for the task level at 61%. However we see a drop at the segment level, and a score of N/A on the composite feature set. For Study 1, we observe that when the therapists did not have to provide the meaning behind a score (i.e. which impaired movement feature they were using to make a judgement), we lost granularity across the segment and composite level.

Study 2: Movement Quality to Function (MQ2F), For this version of the interface, we observed that by flipping the approach, we shrunk the interpretation space of the therapists. As noted several times already, therapists are trained to assess by focusing primarily on movement functionality. By amending the VAT to allow therapists to focus primarily on the meaning behind the score, the therapists skewed towards one direction in their analysis. For study 2, we had a strong IRR across the task level at 83%. However across the segment and composite, the IRR was 43% for the segment and 46% for the composite. Though through this approach, we began to gain some insight into the composite feature sets, it was not enough to accurately and efficiently provide accurate results for the algorithmic models.

Study 3: Structured Decision Process This interface approach produced the most consistent inter-rater reliability on the task, segment, and composite level. There was an IRR of 64% on the task level, and 68% for the segment and composite level. By calculating the IRR using the Cohen Coefficient, 61-80% shows a substantial agreement amongst raters. The scores were also well distributed on the 0-3 scale as compared to the second MQ2F study (see Fig 4.10). The rating process, when applied to a larger data set, could allow the data driven delineation between movement impairments that interfere with the execution of a task or segment (thus resulting in a 1 or 0) and the ones that may hinder or slow down the execution but do not deter it. However, the detailed quantification of these impairments (i.e.

	Study 1: **Score-to-Meaning**	Study 1: **Meaning-to-Score**	Study 1: **Socratic Approach**
Inter-Rater Reliability (IRR) - **Task**	61%	83%	64%
Inter-Rater Reliability (IRR) - **Segment**	45%	43%	68%
Inter-Rater Reliability (IRR) - **Composite**	N/A	46%	68%
Score Distributions	**2 Therapists:** 1 = 48% 2 = 42 % 3 = 10% **4 Therapists:** 1 = 40% 2 = 39% 3 = 16%	**2 Therapists:** 1 = 72% 2 = 20 % 3 = 8%	**2 Therapists:** 1 = 31% 2 = 67 % 3 = 2%

Figure 4.10: Inter-Rater Reliability and Score Distribution Across Study 1.

the exact amount of hand digit openness that allows for the execution of a task) requires the integration of the therapist ratings achieved through this version of the rating interface with computationally analyzed kinematic data since the therapists don't provide detailed kinematics review.

Rater Consistency

Both T1 and T2 had high rater consistencies for the second MQ2F. T1 rater consistency was 98.2% and T2 was 94.734%. This model of the rating interface produced high intra-rater consistency: the rater consistency of T1 and T2 were above 80% with T1 averaging 85% and T2 averaging 90% across 5 sessions. However, though T1 rater consistency 85% ,it was a small drop from the MQ2F approach at 98%. The drop in rating quality was primarily due to external factors which affected the therapists ability to be consistent in their rating (see

the death of a pet quote above).

Time per Session

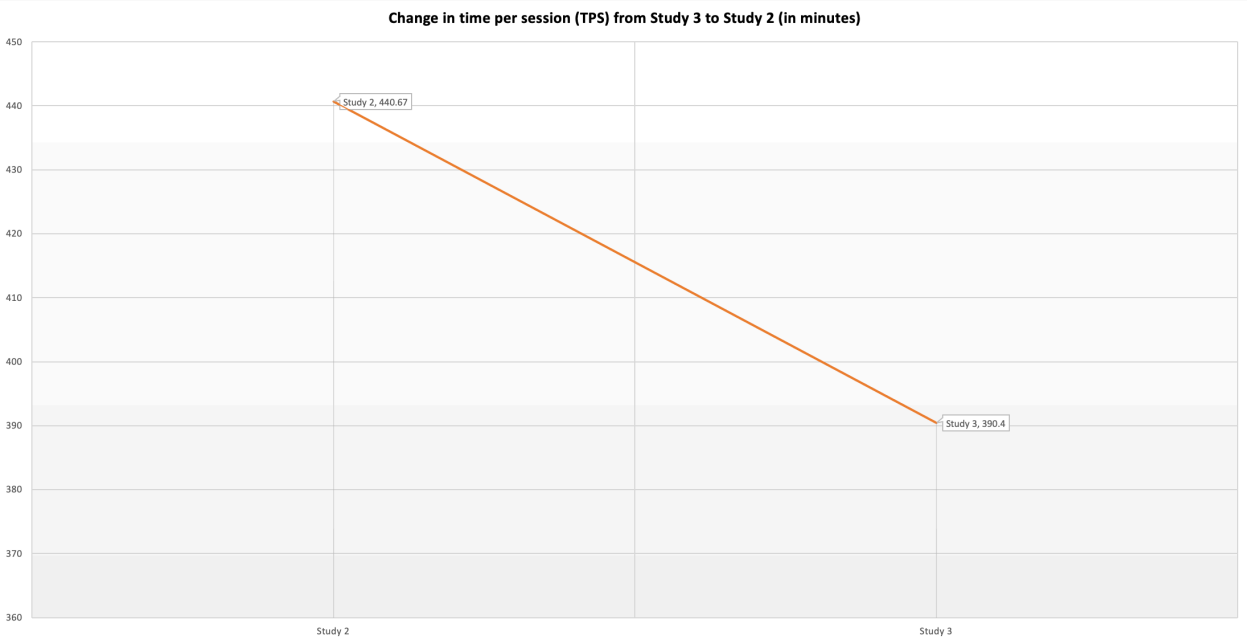

Figure 4.11: Change in Time Per Session across Study 2 and 3. In Study 2, it took T1 and T2 an average of 441 minutes to complete a session. However, in Study 3 we see a drop of 11% to 390.4. per session (25 videos per session).

The therapists learned to use the interface faster when the rating process was guided/structured by the interface (MQ2F and SDP). More experienced therapists learned faster and had higher consistency as shown in Fig 4.11.

For T1, there was a steady decrease in annotation session completion time across the rating sessions. For the first set of 37 videos, T1 spent three hours and 25 minutes completing their assessment and interpretation of the 37 ADL videos. By session five however, T1 completed their assessment of a similar group of 37 ADL videos in two hours and 37 minutes. T2 also decreased time spent on task, but in a somewhat more sporadic fashion across the middle sessions. That said, from an initial first session time of four hours and 41 minutes, T2

completed their final session of 37 videos in a time of three hours and 29 minutes. It is important to note that T1 had been involved in the VAT studies from the first day and had significantly more input into the tool development and indeed, experience with the tool. T2, while initially slower with their manipulation of the tool, nonetheless demonstrated growing fluency and efficiency over time.

Throughout the completion of the three studies (which were greatly extended over time owing to the pandemic), we observed multiple instances where improved design of our physical and digital systems could better assist participants in making their knowledge explicit and in parallel, help make that knowledge computable and useful for an automated system. Elaborations on these improvements are presented in the next section.

4.1 Explicit to Computable - SARAH

Our objective within this phase was to understand how to translate explicit therapist knowledge into a computable model for machine learning systems. The three VAT studies helped identify and standardize how therapists could consistently assess patient movement using the SARAH system. With this knowledge at hand, consideration for additional stakeholders also needed to be reckoned with including patient and caregiver experience, and the needs of the machine learning components in our system which are vital in terms of making the system provide accurate and meaningful feedback to the patient in the absence of the therapists. The data collection process in the pilot study was not standardized and produced high variant data which made analysis very challenging. The camera(s) were not placed in a standard or consistent position, producing high variability in the captured videos in terms of location of the patient, activity space in the image, lighting, camera height, and viewing angle. The surface markings on the mat helped the patient to position the objects, but were

not useful to the computer vision members of our team with regards to assisting with the identification of movement boundaries and/or transitions between movement segments. In addition, the computer vision system regularly "lost" the objects owing to too much similarity in color and/or to patient clothes. The research team were also present during the pilot data collection, meaning members of the team were responsible for the entire system setup and calibration of the system. This is not feasible or likely for a home-based system, meaning that considerable changes needed to be made to our system in order to ensure ease of use and accuracy and consistency of the data collection in the absence of the research team. Ultimately, the goal with this work was to understand how to first setup and then capture high-quality, non-variant videos of patient movement in an embodied space without intruding on both the patient and/or therapist workflow.

Standardizing Data Capture

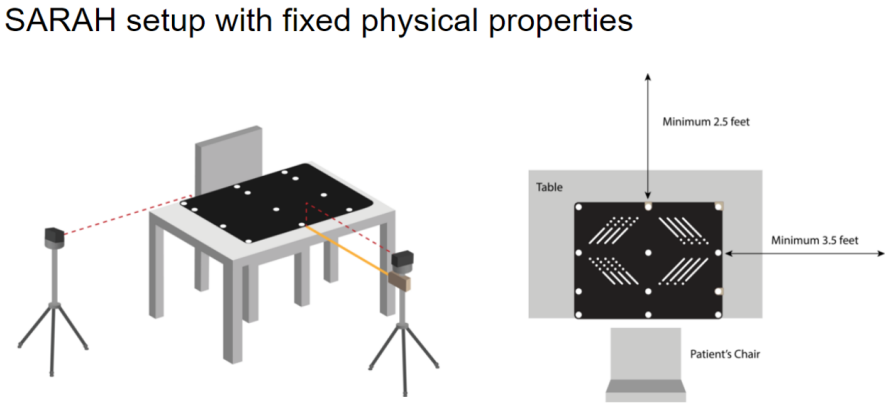

Figure 4.12: SARAH system setup with calibration and camera placement standardization

Through learning from our initial pilot studies, we developed a more complete and robust SARAH system to encompass the needs of patients, therapists, and machine learning agents. This involved standardizing and improving the SARAH objects, mat, camera placement,

setup instructions, calibration process, and patient/therapist interface. Our primary objective in creating and standardizing this process was to create a straight-forward, user friendly approach that allowed for high resolution video capture with little to no variance, while still not impeding in the workflow of the patient or therapist.

The camera used in the Emory capture study regularly "lost" the objects when the patients moved them as the computer vision system was impeded by non-standardized camera angles and confusion with patient's clothing. Similarly, some of the smaller objects were not reliably detected due to occlusion by the patient's fingers, hands, or limbs when they were moved. The participants in the study (patients and supervising therapists) also pointed out several important gestures and activities not addressed by our approach, including tasks related to eating, hygiene (brushing hair or teeth), and manipulating keys and locks. The industrial design team modified and extended the set of objects to address the identified human and machine concerns, as shown in Fig 4.13. The green hourglass object became two red objects that can be screwed and unscrewed (adding another task dimension), with the wider unscrewed piece now serving as a "hairbrush" for an above the shoulder grooming task, while the narrower piece can be used with the "key" object for a fork and knife cutting type activity. The soft doorbell object is now fitted in the bottom with an internal mechanism that mimics a lock that can be "opened" with the new grey "key" object.

The orange button press object was modified with a small internal mechanism to create a more satisfying force-feedback experience for the patient when it is depressed. With respect to the computer vision system, the designers re-conceived the set of objects to ensure that each object is unique in terms of height, width, and color, which when assessed together, greatly improves the accuracy of the object detection approach. Smaller objects, such as the grey "key" and green "screw top" were either lengthened or extruded in diameter, to mitigate issues of occlusion.

Figure 4.13: Sarah system setup with calibration and camera placement standardization

As noted, for the Emory study, the design and engineering team setup and calibrated the system in the clinic before each patient visit. However, this approach is unrealistic for proposed home based use at scale. In this context, the therapist and even perhaps the patient caregiver needs to be able to efficiently and accurately setup the physical components of the system and calibrate the cameras with regards to the home space and to the patient. To address this, the industrial designers, the user experience designers, the computer vision developers, and the two occupational therapists on our team worked together to conceive a redesign of the staging mat (shown in an overhead view with the printed objects in Fig 4.14) and the creation of an accompanying set of analogue and digital instructions to support the therapist in preparing the system for patient use.

The redesigned mat is a modified large format black mousepad, screen-printed with high-contrast guidance lines for the patient, indicating the four primary therapy activity performance spaces through increasingly colored lines. In order to ensure the correct alignment of the two cameras, the designers created three 3D fiducial markers which are aligned with the circles on the mat and used to assist with mat detection. The fiducial marker to the side

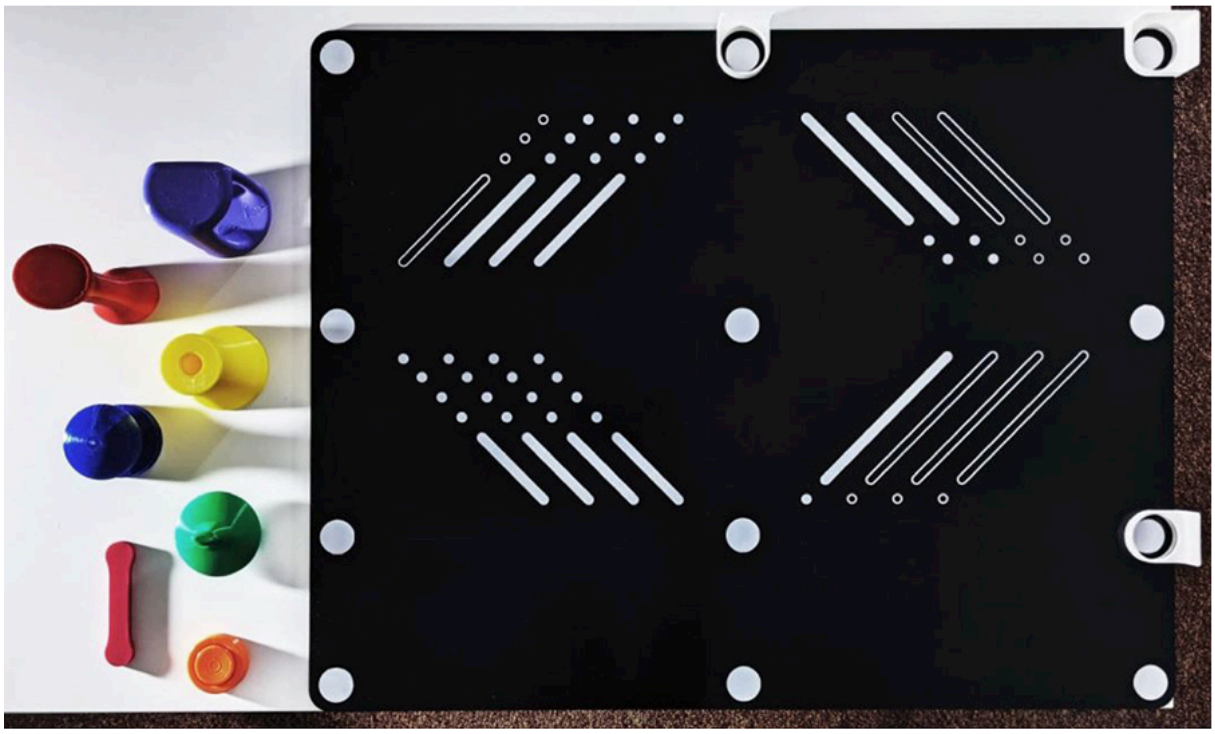

Figure 4.14: Sarah system setup with calibration and camera placement standardization

of the patient needs to line up with the center of the camera capture frame, while a second fiducial marker at the top right-hand corner of the mat assists with the correctly aligning the mat on the table. The third marker is used to correctly align the front facing camera (see Fig 4.15) The four rows of circles on the mat assist the computer vision system with boundary detection, which is needed to consistently analyse the patient activities.

The calibration component of the SARAH system was conducted by fellow researchers at the INR lab and findings from the work are published in Juliet Clark's VT master's thesis [14] and in a peer reviewed publication at ACM IMX [15]. The IMX paper describes the design and evaluation of the setup and calibration system. The goal of the study was to assess if non-expert users could efficiently and accurately set up the system using the interactive instruction manual provided and calibrate the cameras both for the space and for a "patient" (one of the research team members filled in for this role in the study). The therapists were

Figure 4.15: Sarah system setup with calibration and camera placement standardization

given the tablet interface to access the instructions and received no other prompts from the research team. Prior to the study, the tablet interface was instrumented to record the time participants spent on each stage of the calibration process, as well as the number of times they clicked on any of the diagrams and videos. Four of the therapists completed the system setup in less than three minutes and twenty seconds, with two of the therapists taking almost seven minutes. All of the therapists took less time to set up the (second) front camera than the side camera, and a few of the participants also volunteered that if they were asked to conduct the entire process a second time, they would be able to move through it considerably faster. Quantitative analysis of the calibration processes across the participants was also promising, as deviations along the x- and y- axis were minimal.

From a computer vision perspective, standardizing the captured video frame through the reduction of frame invariance is significant. The cameras must be positioned so that the patient and the staging mat itself can be recorded with minimal deviation across different patients, setup equipment, and home environments. The less variance there is, the more

accurately the machine learning approach can assess the patient movement and offer appropriate feedback to the patient. The therapist uses custom software on a tablet interface to progress through the setup and calibration activities. This calibration software provides feedback to the therapist in the form of text cues (e.g. “Move the camera to the left”) in response to the computer vision analysis. Once the cameras are setup and aligned with the mat, the therapist can then ask the patient to come to the table for the final step of the process, whereby the system must detect the patient.

After the study at Emory Hospital, significant changes were made to the capture interface. From our observation, interviews, and participants’ feedback, the design of the instructional interfaces were amended to focus on the needs and limitations of an older population. The median age for a stroke survivor is 63, meaning that the design needs to focus more specifically on addressing that demographic’s needs. During the study, we observed patients struggle to read the interface instructions without assistance from the therapists. In response, we increased font sizes throughout, shortened instructions, and reduced the amount of ”clicks” need to progress through the activity session. A back button was introduced into the interface to allow patients to revisit prior instructions without feeling the need to have to remember each set of instructions. During the first pilot study, each of the patients stated they found it ranging from difficult to challenging to progress through the training without viewing prior instructions.

A demonstration or rehearsal session was introduced in response to observations of patient anxiety about ”getting the task right” on the first attempt. This design was implemented to alleviate performance anxiety and provide patients with the capability to view how the system and instruction format would be organized for the rest of the assessment. In addition, by allowing them to rehearse, the patients could better familiarize themselves with the system in a low-pressure way. From the therapists perspective, the addition of a menu button served

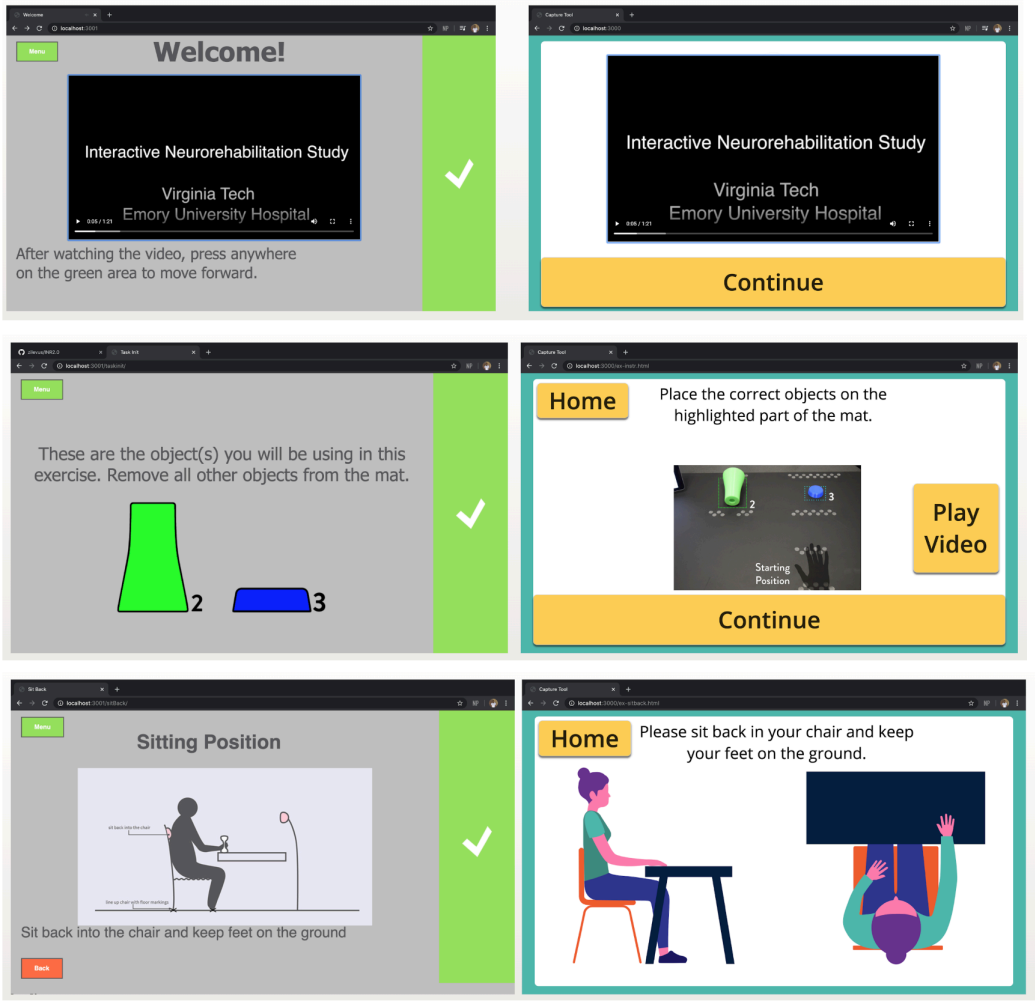

Figure 4.16: On the left is the preliminary version of the interface. On the right, the newer version of the interface based on the patient and therapist feedback.

as a way for therapists to have the ability to adjust the activity protocol based on their observations of the patient (which could be used when therapists are present one day a week in the home).

Knowledge Based Data Driven Approach

The capture of standardized videos of stroke survivor training activities with the updated SARAH system will result in more accurate labeling of activity for automated assessment, even in diverse contexts and with limited datasets. The leveraging of expert assessment of task performance can inform the computational rating of movement using "low-cost, limited, noisy, and variable kinematic data" [1]. The outcomes of this process need to dovetail with typical therapist assessment approaches so that they can use generated summaries to virtually observe patient progress and consequently adapt the therapy protocol. The therapists on our team used the captured videos to assist with the identification of well-defined sub-spaces that are drawn as bounding boxes on the video data as shown in Fig 4.17

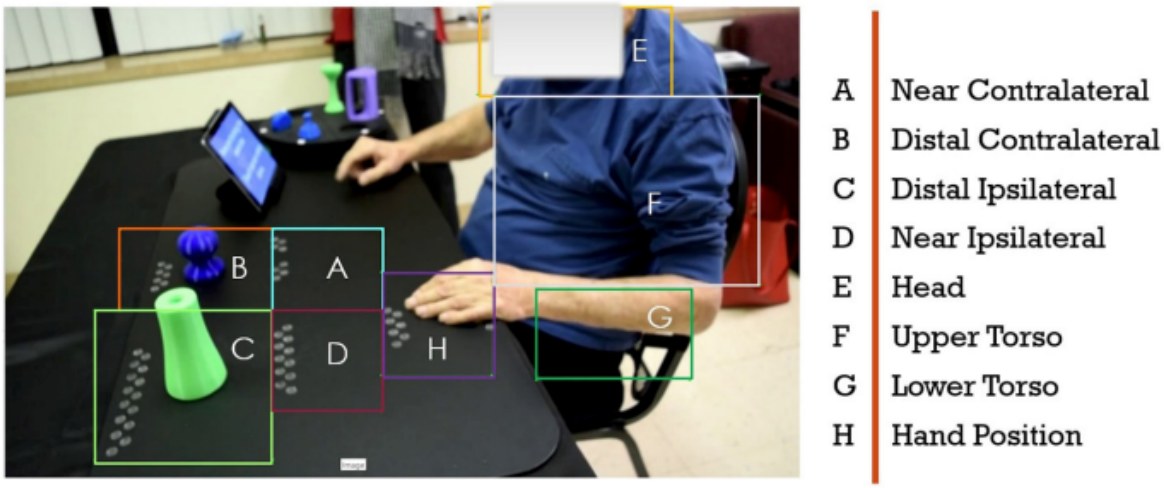

Figure 4.17: On the left is the preliminary version of the interface. On the right, the newer version of the interface based on the patient and therapist feedback.

The ratings provided by the therapist can ultimately serve as the ground truth for assessing the effectiveness of the automated computational approach. Recent findings by members of the INR lab published in Frontiers in Neurology describe how the "hierarchical model fusing expert knowledge based approaches with data-driven techniques can produce robust results in segmentation and task completion assessment in rehabilitation even when utilizing low quality and high variability data. [1]" With over a %90 performance rate in assessment of segment completion and task completion, semi-automated rehabilitation at the home using the SARAH system is considered feasible.

4.2 Empowerment for Humans - SARAH

In Phase 3, we focus on how to leverage human intelligence with computational intelligence to empower humans. In this case, I aim to understand how to take the high-quality video capture data produced by patients, along with the explicit therapeutic rating to better understand how to create a feedback and summarization process that could provide feedback with the assistance of a machine learning approach. The goal is to understand and reveal meaningful summary data of value to both the therapist and the patient.

In Figure 4.18, the top showcases the composite feature sets in which the HBM model generated through the use of the videos recorded with the capture tool. These features shows what the HBM model believes to be the key feature sets that would impact impairment. The bottom half of 4.18, displays the percent of total observed features over agreement with the feature sets the therapists used with the rating interface. When focusing on quadrant 4 for both figures, we begin to see the correlation between the feature sets in which the computer agreed (HMM) with the therapists.

Therapists can readily embrace and utilize a top-down hierarchical approach to rating that

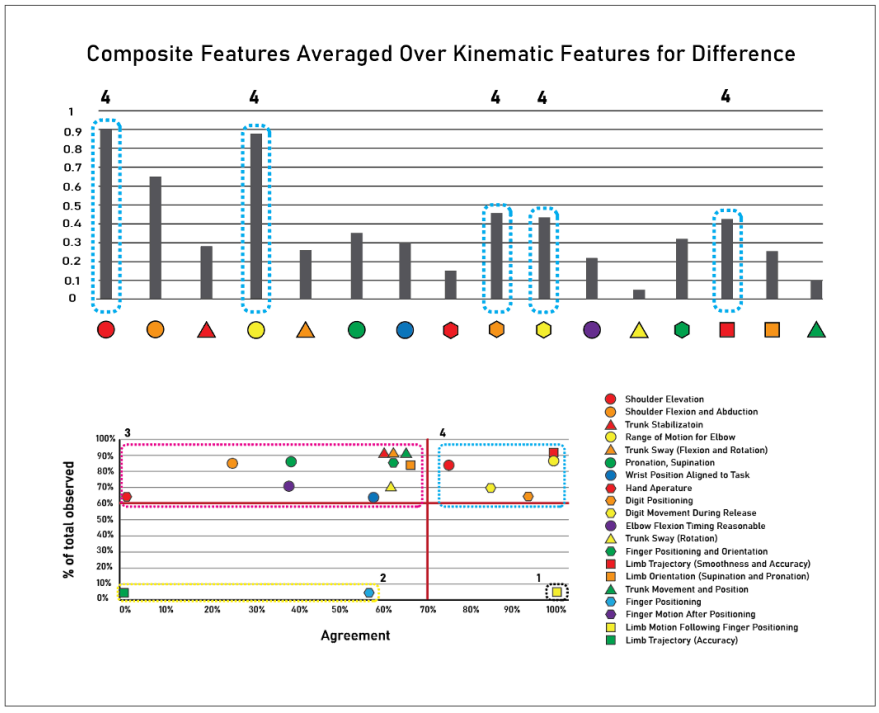

Figure 4.18: Percentage of agreement vs percentage of observation per feature using the two therapist ratings from the Structured Decision Process. In this case we have set the threshold based on the x axis meaning when both therapist check the same feature (they agreed these features are influencing functionality)

consists of three layers: task layer (which is strongly associated with assessment of function), the segment layer (partly associated with function and partly with movement quality) and the composite feature layers (associated with movement quality). This may denote that therapists already use tacit and personalized hierarchical approaches to rating. Therapists can consistently separate even slightly impaired movement from unimpaired movement. But even when using a highly structured approach, therapist approaches to the impact of movement quality on functionality have similarities and differences. Using data to map the level of disagreement across the continuum from severely impaired movement (score of 0) to unimpaired movement (score of 3) both at the task and segment layer will produce a normal distribution (movement impairment on X, disagreement on Y (0 unit disagreement, 1 unit, 2 units). Mapping the number of composite feature disagreement will also produce a normal distribution.

From Phase 1 and 2, we can more clearly observe the relationship between functionality and movement quality. The correlation between the agreement vs percentage observed and the composite features averaged over kinematic features for differences 4.18 begin to show statistical and probabilistic approach to understanding movement quality to functionality. The Structured Decision Process method showcases the optimal process in increasing the granularity of observation to further explore the relationship between movement quality and functionality. The rating and standardization process now provide a statistical hierarchy for understanding rehabilitation movements. Through the use of the HMM and computational analysis, we can begin to provide results to assist in the automation of hierarchical assessment and create a computational approach to understand how movement quality and functionality is assessed. From a machine learning approach, our larger interdisciplinary team aims to develop a model that can integrate raw features and kinematics extracted computational through both the patient and video application tool to create automated

interpretable analysis of human movement performance.

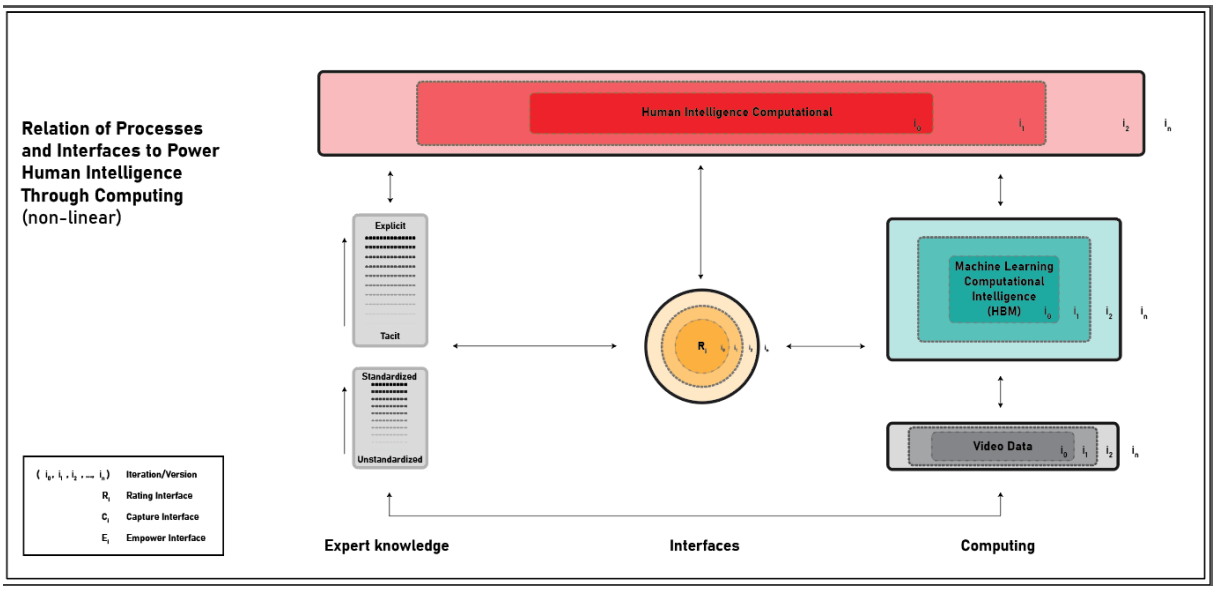

Figure 4.19: Phase 1: Phase 1 describes the process of transforming tacit knowledge into explicit knowledge through the use of the rating interface (Ri). The Ri allows therapists to qualitatively reflect upon their practice while quantitatively facilitating a computable approach for ML systems to increase computational human intelligence.

4.3 Summary of TCE SARAH Outcomes

4.3.1 Phase 1: Tacit to Explicit

In Phase 1, as shown in Figure 4.19, the rating interface is used as a tool to facilitate the translation of clinician tacit knowledge into explicit knowledge. This interactive computing experience allows clinicians to engage in a a process of reflection over time while providing more granular level feedback about their assessment model. This process to move from tacit to explicit knowledge assists the machine learning approach by allowing the computer to understand not only the top-level interpretation of an task but also the decision tree by which that score is derived

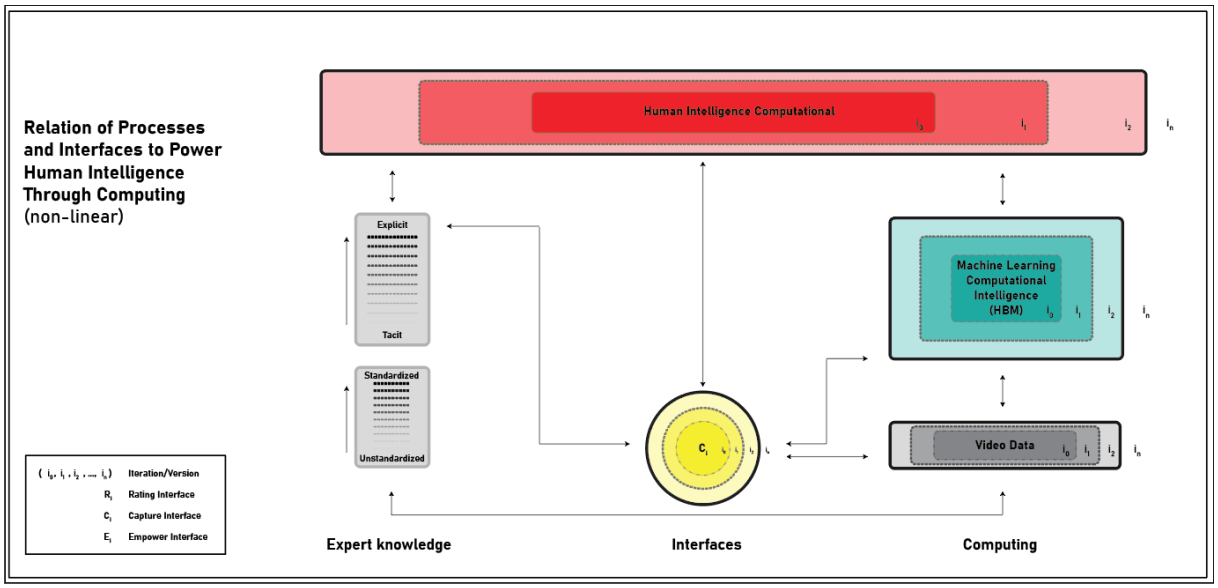

Figure 4.20: Phase 2: Phase 2 focuses on moving from a non-standardized video capture process to a standardized process through the use of the capture interface (Ci). The video data generated from the Ci is fed into the ML system to detect the human key points during each exercise.

4.3.2 Phase 2: Explicit to Computable

Phase 2 focused on transforming the explicit knowledge from phase 1 into a computable model. The goal of phase 2 was to generate data and create a model that machine learning systems could understand. In phase 2, we observe the development of a standardized framework that is unobtrusive for both clinicians and patients for in-home rehabilitation and clinical assessment. The findings from this stage, combined with phase 1, allowed for computational human intelligence to increase.

4.3.3 Phase 3: Empowerment for Humans

In Phase 3, we focused on using computational power to empower the human. The process involved using a non-standardized to standardized process, a tacit to explicit process, the

HBM model and video data to power human intelligence and create summaries that iterate through each step to make human intelligence better.

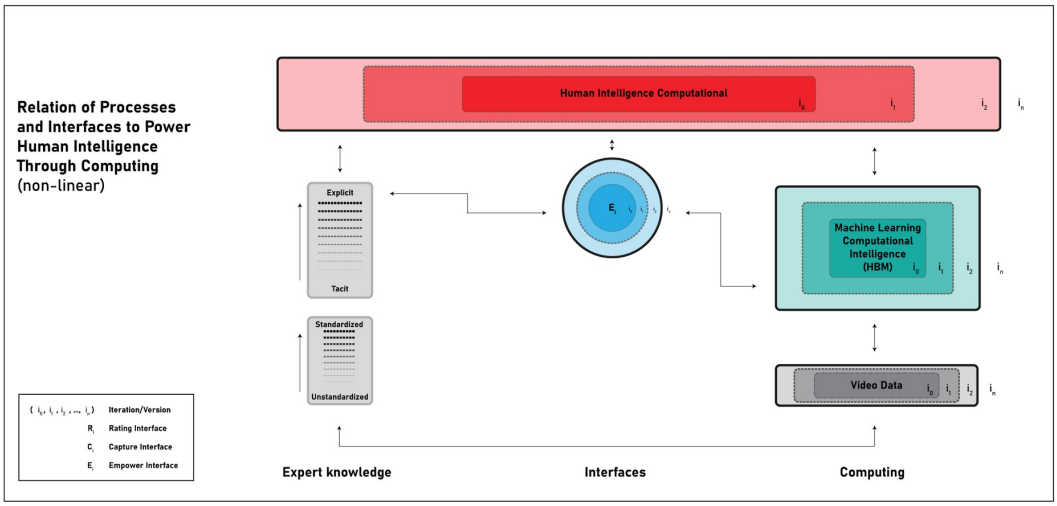

Figure 4.21: Phase 3: Phase 3 combines the data gathered from the capture and rating interfaces and the HBM model to to empower the human. The process includes using both the capture and rating data to begin to understand how to make therapy recommendations for both the patient and therapist.

Chapter 5

TCE Methodology for Standardized Clinical Assessment

One key aspect of my research is understanding if my methodology and approach can be generalized to other application spaces such as clinical assessment. I applied my method to the ARAT (Arm Reach Assessment Test) to assess my approach and evaluate my reflective design process in this different context. I utilized adapted versions of the design methods and approach used in the development of the SARAH System. A key difference between the SARAH system and the ARAT system is that the ARAT is a standardized instrument used by clinicians and therapists, with a known and familiar rating system. However, the ARAT rating rubric, as explained below, is relatively coarse as it uses numbers only (no elucidation of movement features). A second difference between the systems is that the ARAT is administered by the therapist on a tablet/laptop, meaning that the capture interface is controlled by the therapist and not the patient. From a development perspective, I also adapted both the capture and rating interfaces for the ARAT simultaneously, building on insights described in the previous chapter. This efficiency also meant that the capture and rating interfaces could be co-designed and adapted with the ARAT administering team simultaneously. The rating approach therefore used was the Structured Decision Process (SDP) as that was the most successful approach from the SARAH studies.

5.1 ARAT System

The Action Research Arm Test (ARAT) is a standardized upper extremity performance assessment for individuals experiencing hemiparesis after a stroke. The ARAT is rated as having excellent test-retest reliability and inter-rater reliability, moderate burden overall to complete, and moderate construct validity and responsiveness [86]. For individuals six months post-stroke with moderate impairment, the ARAT is the recommended measure of activity and participation [54], [86], [16]. This outcome measure is highly recommended for use by the American Occupational Therapy Association and the American Physical Therapy Association. Currently, clinicians use the ARAT to assess patients' function and impairment post-stroke, to inform treatment, and to monitor progress. The ARAT takes approximately 10-15 minutes for therapists to complete with their patients.

This ARAT instrument measures 19 functions divided into four sub-tests (grasp, grip, pinch, and gross arm movement) which assess upper limb functioning using observational methods [51]. Clinicians rate the performance of each task on a 4-point ordinal scale ranging from "0" (can perform no part of the test), to "3" (performs the test typically). To complete the assessment, patients are first asked to perform the most difficult task of each sub-scale with their less impaired limb for clinicians to gauge an understanding of "normal" or "near-normal" patient movement (some patients demonstrate movement impairment in both limbs). After performing the most challenging task in each sub-scale with their unaffected limb, the patient begins the assessment test by attempting tasks in each subscale. The clinician uses the Lyle Decision Rule, whereby if a patient scores a 3 on the hardest task of each subscale, the patient will receive a score of 3 for the remaining tasks in the subscale. If a score of zero is given to the first and second tasks of each subscale (the first being the hardest task and the second being the easiest in the subscale), the patient will receive a zero

for the remaining tasks within that subscale. The same logic is applied throughout the first three sub-scales. However, in the Gross Sub-Scale, the first task is deemed the easiest, so if a score of zero is given the patient will receive a score of zero for the remaining sub-scales.

5.1.1 ARAT Stakeholder Meeting

In November 2020, a stakeholder meeting was convened online between the development team at Virginia Tech and a diverse group of participants from the Shirley Ryan Ability Lab. Participating stakeholders consisted of (1) an occupational therapy clinical practice leader, (2) an inpatient occupational therapist, (3) an occupational therapist/research scientist, (4) a physiatrist, (5) two patients with stroke, and (6) members of the VT research team. The objective of the meeting was to identify factors that might "contribute to the success of implementing innovative technology measures into clinical practice." The VT team presented a virtual prototype design of the capture ARAT system to the stakeholders in a 10 minute presentation. During the meeting, the potential barriers and facilitators of using the new technology in clinical practice were discussed and are described below.

The clinicians noted two key issues with the administration of the ARAT at the SRAL:

1. The ARAT is not consistently used in inpatient settings, due to factors including patient appropriateness (i.e. floor effect).

2. The ARAT is often captured by a number of different clinicians for the same patient rather than a consistent rater at SRAl [83]

Following the presentation of the virtual ARAT capture prototype, the stakeholders discussed several potential key benefits, in particular the use of the system for standardization purposes. The stakeholders believed that having "a highly standardized measurement and

rating system would be beneficial for newer, less skilled clinicians to use for training purposes, as well as for re-education and ongoing training of current clinicians". The two participating patients stated that they would find viewing video results of their performance during assessment useful in helping them understand how to correct their compensatory movements. The potential of the system with regards to supporting clinical research was also discussed, with clinicians noting that "(1) the video assessment would allow for more sensitivity to capture changes in impairments;" and "(2) the standardized system could allow for non-clinician researchers to assess patients." [83]

Several potential barriers for adoption and use were also noted in the stakeholder meetings. A primary issue echoed concerns noted by the clinicians and therapists with the SARAH system regarding the typical therapist focus on function over movement quality. The clinicians noted that "many occupational therapists are more concerned with function, rather than impairment, so they value their observational assessment of the functional tasks performed during the ARAT greater than the potential benefit of having more precision with the video-assessed impairment measure." [58]. The issue of time was also described as paramount with regards to both efficiency and cost. The clinicians described how even with the traditional analog ARAT setup, they often didn't use the ARAT as it took time away from therapy and cost one billable unit to administer. They thought that this could be an issue with regards to implementation and acceptance. Table XX below summarizes the key pros and cons as revealed by the clinical team. Findings from this early workshop greatly influenced the approach adopted in conceptualizing how to best provide augmented value for the collaborating clinicians. Working with other members of the INR lab, I created a cyber-human hybrid workflow system to help standardize the ARAT assessment. As a UX Researcher, I worked with expert clinicians, industrial designers, and computer vision experts to develop a capture system that contains a video capture process to detail the key movements of a

patient's performance. One key component of the capture system is to cater to the needs of both clinician therapists and the computer vision system to produce accurate results that are meaningful for both stakeholders. The key components of the ARAT capture and rating system include:

1. A rig set up to hold the cameras at standardized positions.
2. A labeled staging mat, shelf, and set of ARAT objects.
3. A checkerboard and fidicial marker calibration set.
4. A capture interface allowing the therapists to control the cameras while administering the test.
5. A rating interface including emphasis on patient movement function and quality.

One of the significant challenges of the ARAT system is maintaining the test's existing standards while incorporating a new physical camera architecture. I communicated with the SRAL clinical therapists experienced in conducting the ARAT to reveal the optimal camera viewpoints for assessing of patient activities. The INR team designed the architecture of the activity space and the ARAT objects in accordance with the standardized specifications of ARAT [44]. In the next section, I briefly describe the design, development, and pilot testing of the preliminary ARAT physical and digital system.

5.1.2 ARAT Design and Pilot Study

The goal of the ARAT system is to allow for the detailed capture of patient activities during standardized movement assessment. The use of the system during assessment, as per our stakeholder findings, needs to fit with the typical therapist workflow in terms of time taken

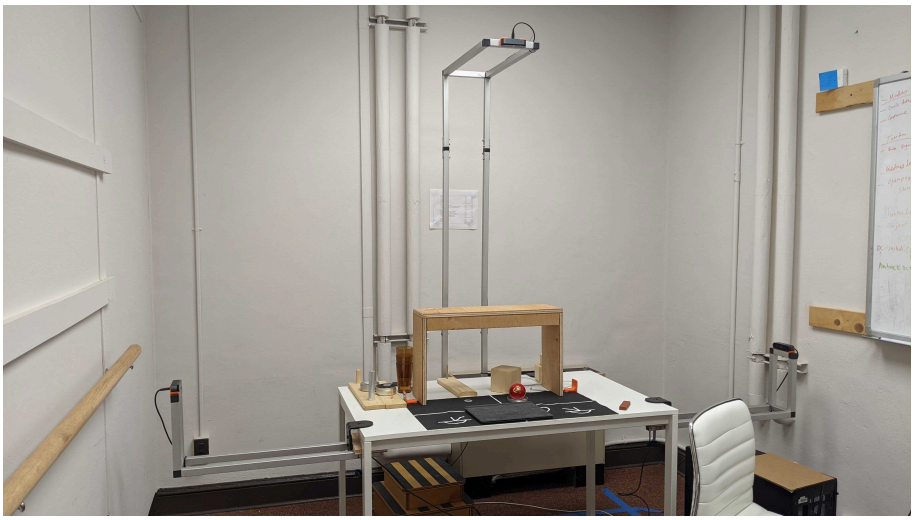

Figure 5.1: The ARAT system embodies a human-in-the-loop architecture to capture therapist knowledge and make recommendations to assist clinicians with movement assessment. It contains a 4 camera setup.

to administer and accuracy of the assessment output. In addition, the use of the system must offer an additional benefit to the patient and the therapist above and beyond that offered by the traditional experience. To limit workload burden on the therapist, we devised a physical infrastructure to be semi-permanently installed by our team, comprised of a camera truss, cameras, tripod, and labeled mat (see Figure XX below)

We developed a camera calibration process (checkerboard pattern to collect multiple pairs of images) aimed at creating a quick, accurate and easy to use way of checking the cameras prior to assessment. Fiducial markers on the table also assist the therapists in ensuring that the physical rig has not moved between tasks/patient. The SARAH capture interface was amended by simplifying and orienting the capture process for a therapist user administering the ARAT.

While we ensured keeping in step with the standardized ARAT assessment setup, some adjustments were also made to some components to simplify the setup process and improve compliance to standardized factors. Our team decided to use 1080p webcams as our primary

Figure 5.2: Components and system setup for ARAT Captures

Figure 5.3: The Capture Interface for the ARAT

capture method while also exploring the integration of additional unobtrusive capture sensor including the Leap sensor (such efforts were ultimately abandoned as the results were not helpful). Camera placement was initially distributed by anticipated usefulness to the clinical therapist for later assessment of movement quality and optimization of framing for later computer vision analysis. A high-powered tower machine hosts the capture tool interface and the computer vision and computational assessment models. The tower computer acts as a base for running all computational operations of the system, including the user interface, video cameras, and the database for collecting the captured data. A screen sharing approach is considered from the tower machine to the tablet interface, allowing the therapist to access the capture tool. The convenience of a remote interface lets the therapists freely move around the activity space during the assessment.

Our team conducted a pilot study using unimpaired subjects in a simulated clinical setting. The purpose of the study was to stress-test the physical infrastructure, determine the usability of the digital interface, assess the appropriateness of the camera viewpoints, and collect videos of unimpaired subjects completing ARAT activities for subsequent review and feedback by the SRAL clinicians.

The prototype for the ARAT capture rig was assembled and positioned around the workspace as shown in 5.1. The ARAT evaluation equipment mat component was placed centered along the proximal edge of the tabletop workspace. The rig's frame was attached firmly to the lateral edges of the table, allowing us to fasten and connect each camera quickly. Checkerboard calibration occurred amongst the three tabletop-focused camera perspectives. Each participant was seated centrally to the workspace with approximately 15 cm between the anterior trunk and the proximal edge of the table. A facilitator used the tablet interface to initiate calibration checks, navigate the interface, and enter information as prompted. The components of the first activity were arranged, and the subject was verbally instructed on how to perform the specific ARAT task. The capture was initiated and terminated through the interface, and the videos were downloaded upon termination. The subject was assigned a score, and the process was repeated for the remainder of tasks in the ARAT assessment using the capture interface as depicted in Fig5.3.

Through observation of the pilot study, our team identified several areas of improvement. The rear side camera, used for evaluation of the fourth subscale, needed to be detached from the rig system as its positioning relative to the mat was not a significant factor. The structure also needed to be more rigid and condensed to occupy a smaller footprint. We concluded that our team needed to create more consistently visible markings to ensure that the calibration checks occurred without interference from the assessment kit. The wired connections to the cameras also needed to be color-coded to ensure each device could be assigned an appropriate label within the interface.

5.2 Movement Taxonomy and Rating Rubric

The development of the movement taxonomy and rating rubric for the ARAT system was initiated online between the rehabilitation experts at Emory Rehabilitation Hospital, led by Dr. Steve Wolf, and the rehabilitation experts and therapists at the Shirley Ryan Ability Lab, let by Dr. Zev Rymer and Caitlin Newman. While the ARAT is a standardized instrument for movement assessment, it is a relatively coarse measurement tool, using numerical ratings only to describe patient movement. Between 2020 and 2021 (a somewhat extended timeline due to COVID-19), the clinician team developed a rubric and rating manual for the ARAT system (full manual is presented in Appendix X). As with the SARAH rubric, this approach serves to make the tacit knowledge of the therapist explicit by using a standardized set of features by which to describe observation of human movement. The purpose of the ARAT rating manual is to "provide a segmentation and rating approach (rating rubric) for rating videos of upper extremity movement of stroke survivors performing the ARAT. The proposed approach aims to make the rating of movement performance computable (i.e. can drive a machine learning algorithm) and feasible from a therapist point of view (i.e. a group of therapists can be trained to use it in a consistent manner)." The rubric provides therapists with a set of five standardized segments observable across all ARAT tasks and up to four movement elements per segment that the therapist must observe and rate. The manual states: "This approach makes consistent expert rating more feasible. The rubric thus advances a uniform approach to the rating of the execution of all segments and tasks which allows the development of a consistent hierarchy of rating that can inform the standardization of rating by experts and the use of standardized expert rating for the development of automated analysis and rating algorithms." The segments and associated movement elements are depicted in Fig5.4.

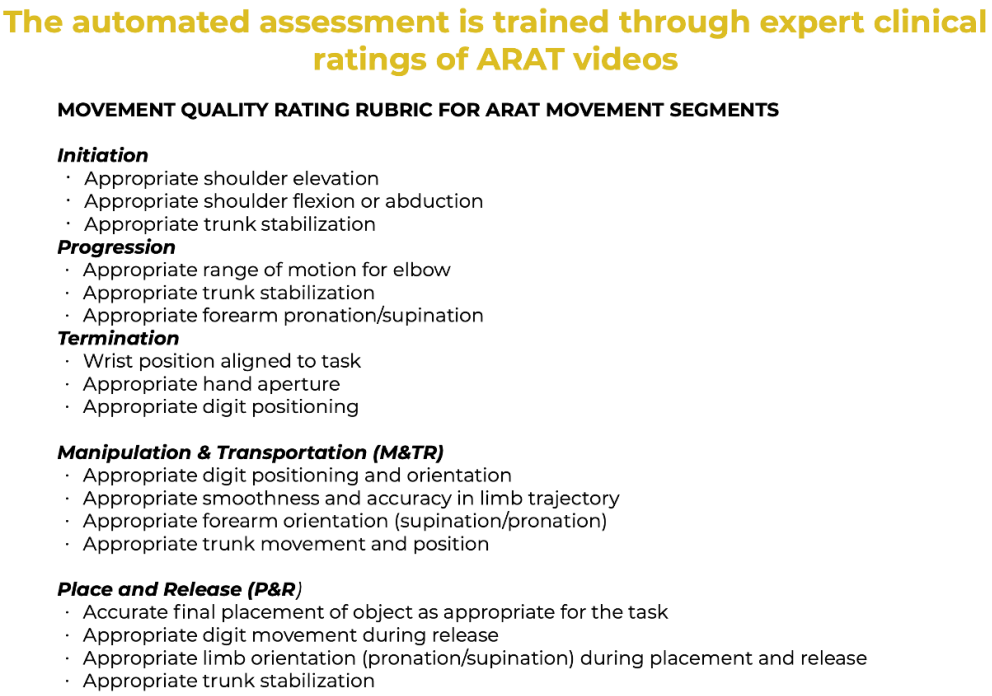

Figure 5.4: Diagram depicting the five potential movement stages for the training activities in our system

5.3 Video Application Tool II

The video application rating tool uses an identical question based approach as the VAT used in the Structured Decision Process model in the SARAH system. The primary difference is the addition of the Place and Release segment and associated movement features which are important components in the ARAT assessment. This version of the Video Application Tool was temporarily installed at the Shirley Ryan Ability Lab in Chicago in November, 2021. Over a two day period, stakeholders were introduced to and/or used the system. The objective of this proof-of-product stage of development was aimed at identifying measurement-related factors that may contribute to the success of implementing innovative measures into clinical practice in the future. Key clinical concerns such as setup, instructions, measurement displays, and recommendations for data interpretation were discussed. The first stakeholder meeting consisted of: (1) two participants with stroke, (2) members of the clinical translation team core, and (3) members of the research team. The second stakeholder meeting consisted of: (1) four occupational therapists, (2) members of the clinical translation team core, and (3) members of the research team. The occupational therapists received a brief training session on use of the ARAT capture session as they observed members of the research team use the system. Following this, each therapist participated in the capture of two patients completing the ARAT (two assessment sets each). As each task was completed, the capture interface streamed, recorded, and downloaded high-quality videos of each exercise from four angles. The research team segmented the captured video that evening and inserted the video into the ARAT rating interface. The following day, the four therapists participated in a preliminary rating session with the data and a follow up discussion session.

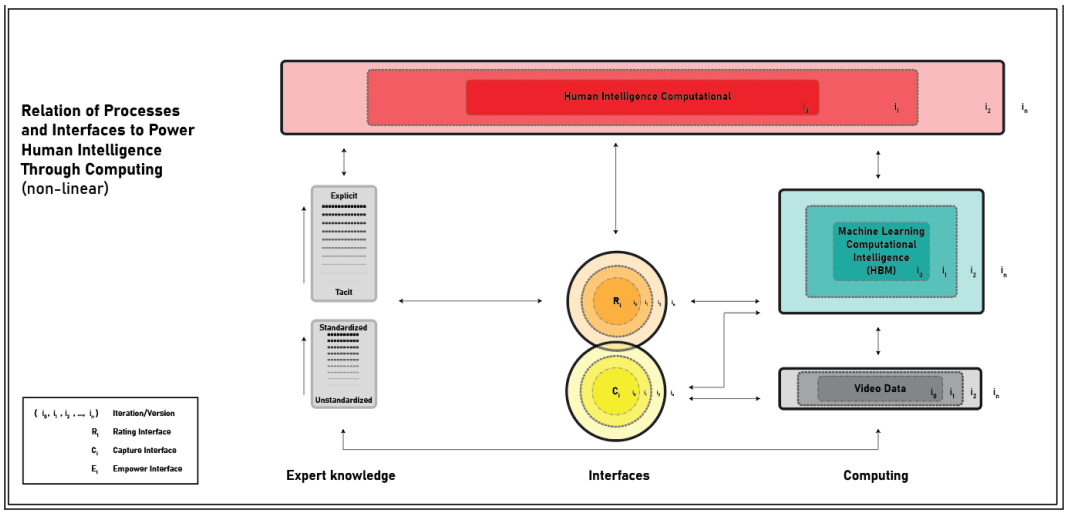

Figure 5.5: The TCE Methodology which includes both the Rating and Capture interfaces used at the Shirley Ryan Ability Lab.

5.4 Video Application Tool Results II

5.4.1 In Person Workshop Outcomes

The four therapists successfully completed the setup procedures and capture of the data sessions. The therapists were able to maneuver around the patient and activity space in a composed way while manipulating the capture interface on the laptop. The internal focus group report notes the following important observations: "During the stakeholder panel, clinicians reiterated a common theme that oftentimes clashes with engineering perspectives of rehabilitation. Therapists focus on function, rather than impairment, during clinical treatment. The goal of inpatient rehabilitation, for example, is not to decrease shoulder hiking or trunk lean during reaching tasks. Rather, the goal is to help patients improve their participation in their day-to-day activities (eg. dressing, cooking). A crucial point was relayed during the meeting that while therapists are more concerned with function over impairment, capturing precise impairment metrics could add value to understanding and diagnosing movement problems in adults with an upper extremity neurological injury.

Understanding precise kinematics may be particularly useful in a moderate scoring range, where individuals can complete most tasks, but with compensatory movements (score 2 out of 3). [59]"

Participants in both discussion panels reported that the video component captured smaller, refined movements and changes over time. Additionally, the computational system could allow improved feedback to be incorporated into electronic medical records, provided through an app or just a printout. The crucial element of the feedback mentioned during the panel meetings was how the information is relayed to patients and clinicians in a user-friendly way. The feedback should use language that the clinician already knows and does not need additional training to understand. This feedback will facilitate how the clinician translates the computational feedback as a part of their treatment session and patient education. Although the video capture system can provide a plethora of impairment measures, clinicians may lack the time and bandwidth to interpret more than a few new variables. It would be helpful for the computer to provide a simple feedback process that prioritizes key kinematic impairments and provides a composite kinematic score. One suggestion was to develop a composite "kinematic score" expressed as a percentage that represented all the compensatory movement errors the patient exhibited in each ARAT task, each ARAT subscore, or the total ARAT. Clinicians and patients also suggested that it would be helpful to link ARAT scores and impairments to the salient daily activities associated with each task. This could be useful for patients to have activities they could practice at home without having them practice test items. Including these functional activities in a report would increase saliency for clinicians and the patients.

Based on the feedback from the therapists and clinicians in the in-person study, the rating interface is currently adapted to support the playback of selected videos at half speed to allow therapists to more precisely observe movement quality within each task. A zoom option is

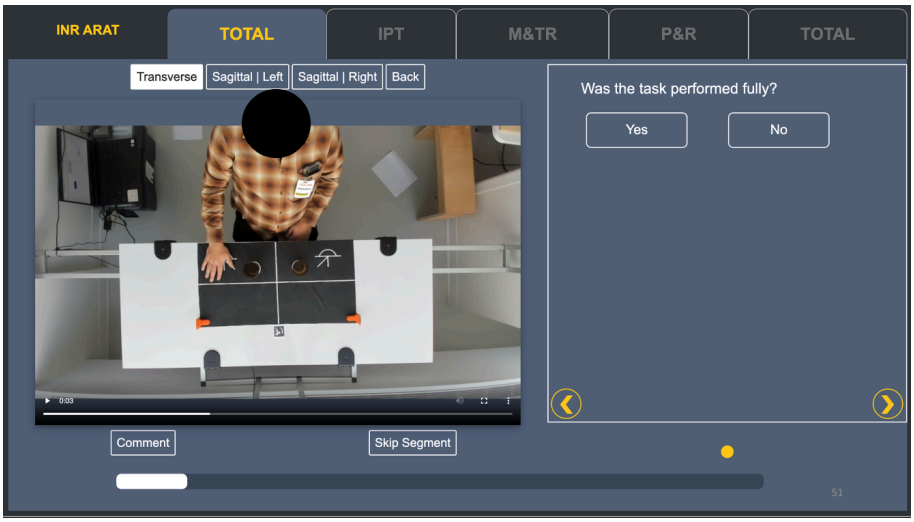

Figure 5.6: ARAT Video Application Tool: The figure above depicts the Video Application Tool redesigned for the ARAT assessment. The Video Application Tool includes the same logic (SDP), however the segments and camera views are modified to meet the needs of the clinicians who administer the ARAT assessment.

also introduced to allow therapists to view particular segment movements at a closer level. For example, for task 11: the marble bearing, the patient must pick up a small marble with their middle finger. From the focus group discussion, the therapists voiced how having the ability to not only zoom in, but also control the rate at which the video was played would allow them to better assess using the rating interface. Furthermore, through our facilitated discussions, it was clarified that the therapists wanted to rate the tasks in the same sequential order as they would enact during in-person assessment.

Based on the discussions at the in person workshop, we made some changes to the rating interface. Unfortunately we were unable to run the rating workshop in person in January 2022, as COVID restrictions at SRAL were updated based on rising infection levels, meaning all research projects were postponed until at least March 2022. Nonetheless, we worked on hosting an online rating workshop with the four clinicans from the November 2021 study. During this event, the clinicians were asked to rate 20 impaired videos of two patients performing the ARAT assessment. Each clinician was given a dedicated laptop with the corresponding

videos and the updated VAT rating interface preinstalled. The goal of this workshop was to understand, even given limited training, if the rating interface could qualitatively and quantitatively provide results that could support the TCE methodology and reorient the relationship between movement functionality and movement quality.

5.4.2 Online Rating Workshop Outcomes

	Workshop Session Part 1	Workshop Session Part 2
Inter-Rater Reliability (IRR) - **Task**	89.97%	80.58
Inter-Rater Reliability (IRR) - **Segment**	94.35%	87.80
Inter-Rater Reliability (IRR) - **Composite**	78.74%	68%

Figure 5.7: Inter-Rater-Reliability across ARAT Rating with four clinicians

Figure 5.7 summarizes the quantitative assessment of the rating workshop. Across the four clinicians, the results of the rating workshop produced high inter-rater reliability across all three layers. The rating workshop (1-10) produced an IRR of 89.97% on the task level, 94.35% for segment, and 78.74% for composite. For the second set of tasks, the IRR was 80.58% on the task level, 87.80% for the segment, and 68% for composite. Across all ratings the significant majority ($> 96\%$) of rating disagreement between therapists is only by one

step. These disagreements are fairly evenly split between 1 and 2 (one therapist gives a 1 and the other therapist(s) give a 2) and 2 and 3.

> "Even for clinicians, like I think we struggle with that gap between a score of two and a score of three. So I mean, eventually, that might be something that's useful."

The clinicians voiced the usefulness of the tool. They viewed the system and the rating interface as a useful learning tool that helps standardize their mental model. One expert clinician stated:

> "I think it would be helpful. So I really view this system as like a learning tool for newer clinicians that are trying to get better at rating, movement, performance, movement capacity, and trying to understand where the breakdowns are. So I think if I was a really green clinician, and as I was seeing a video and having to read it, it would help me understand, you know, is the is the individual breaking down with, you know, releasing an item, right, and, and help me think through Okay, well, why is that release really challenging?...So I really am viewing this and I think kind of the mentality that I'm taking into the, the ARAT project to is like, how could we use these types of systems to provide training and education, consistency of care amongst clinicians, and, you know, reliability of assessment ratings, right, so I could rate something and then a new therapist could rate something. And then that could give us if the ratings were different, it would give us an opportunity to discuss and kind of problem solve the why those ratings were different".

Additionally, the rater consistency for the ARAT ratings were above or equal to 95%. T1,

T2, and T3 all had a rater consistency of 100% for the 20 videos rated, and T4 had a rater consistency of 95% as shown in Fig 5.8

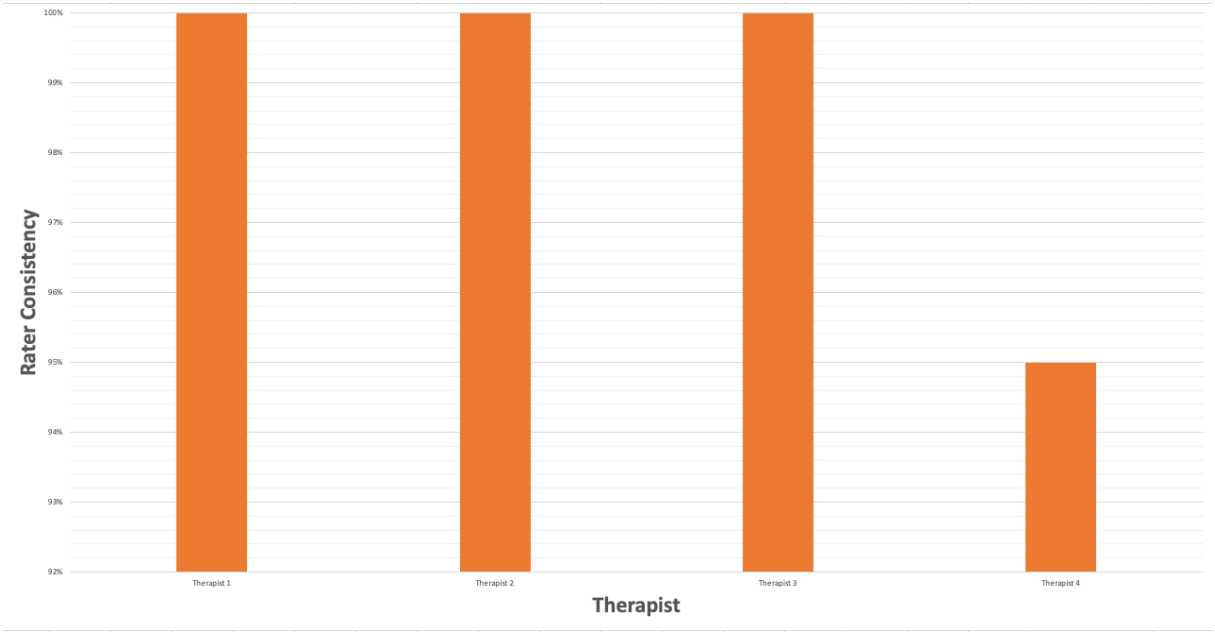

Figure 5.8: Rater Consistency of Therapist Rating

The clinicians suggest that the computational ARAT report should include:

1. The total ARAT score, domain sub-scores, and item scores.

2. Total time to complete the ARAT.

3. Top 3 movement impairments of the patient overall and in each ARAT domain.

4. Kinematic score for each ARAT domain.

5. Exercise Page | ten recommendations related to the patients' impairments exhibited during testing administration.

5.5 Empowerment for Humans - ARAT

The addition of technology into a widely used upper extremity assessment opens up new avenues for precision measurement and sending feedback to patients and therapists within the domain of in-home rehabilitation (SARAH System) and clinician assessment (ARAT). Clinicians begin to see added precision to have value, particularly around the moderate impairment range in patients who can score 2s on most items, but with compensatory movements that will prevent them from (ever) scoring a 3. The system can provide a more detailed assessment, mainly if a patient's overall score stays the same. Still, their kinematics improves, or their time to complete tasks improves. In addition, the system can capture precise information about compensatory movements that a clinician cannot keep track of over time which will help educate the patient on movement quality and performance. Finally, providing a printout of exercise recommendations or functional tasks to practice based on impairments observed in the test can improve the saliency of the measurement system. It is important to clinicians and patients to relate movement quality and impairments to functional tasks.

Clinicians noted that a highly standardized measurement and rating system would be beneficial for newer, less skilled clinicians to use for training purposes and re-education and ongoing training of current clinicians. Allowing patients to view their video results would serve as a feedback tool to facilitate understanding and correct compensatory movements. The new system could also be helpful for clinical research for a couple of reasons: (1) the video assessment would allow for more sensitivity to capture changes in impairments; (2) the standardized system could allow non-clinician researchers to assess patients. Clinicians also indicated some potential barriers to incorporating the new system into clinical care. Primarily, many occupational therapists are more concerned with function rather than impairment,

so they value their observational assessment of the functional tasks performed during the ARAT greater than the potential benefit of having more precision with the video assessed impairment measure. Another known barrier to incorporating new technology into clinicians' regular practice is the issue of time. Currently, many clinicians frequently choose not to administer the ARAT because of the time it takes from therapy (10-15 minutes, or one billable unit), so the addition of setting up the new system will be a barrier to implementation.

5.5.1 Informing the HBM Model

From the focus group discussion with clinicians, one key recommendation in which they suggested would make their assessment model easier would be the ability to view the videos at a slower frame rate in order to fully see the patients movement. Therapist 3 stated:

> "What's really cool about this descaling is this issue of appropriateness, fingers digits, what they're intended to do and what they really do do the single frame allows you to look very careful what persons doing with their hands. Do they really have their digits on it, they competent around her and sometimes we move quickly."

With the introduction of the 'half speed' option as shown in Fig 5.9, clinicians can now begin to use the rating interface as to more accurately delineate between the different composite feature sets as patients perform exercises which by the human eye alone, would be difficult to distinguish.

Additionally insights from the focus group discussion, allowed for the team to better understand how to better capture the full scope of the movement elements. For the therapists seeing the release of of an object, and the patient's hand is a very critical part of their

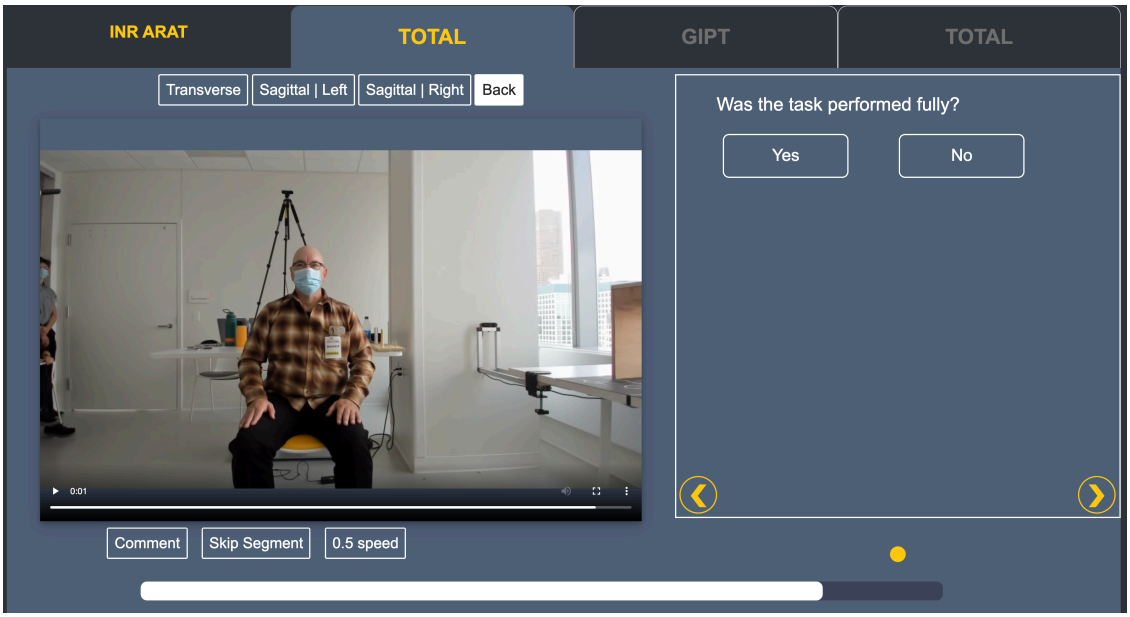

Figure 5.9: Through focus group discussion, the VAT introduced the feature of 'half speed' to allow therapists the ability to watch each video at a slower frame rate in order to see the movement quality elements of each exercise.

assessment. Therapist 2 stated:

> "And even termination is tricky. But I think he got it terminated correctly. But when it came to picking it up and manipulating it, that's where he doesn't have it. But I don't think it's he doesn't have control over what goes on after the control that's in the image er, but in the termination is like, well, he had control. But when he went to lift off, that's what he lost. So then, again, but I think perfection was there, because he got there, right? You progression, an appropriate range of motion is trapped in a move in the right spot. The termination was the hard part where he faltered."

Based on their feedback, I redesigned the VAT to include termination as its own individualized segment that captures the full release with a zoomed in and 3D option as shown in fig5.10. Through this, computationally we can better understand how the patient's hand

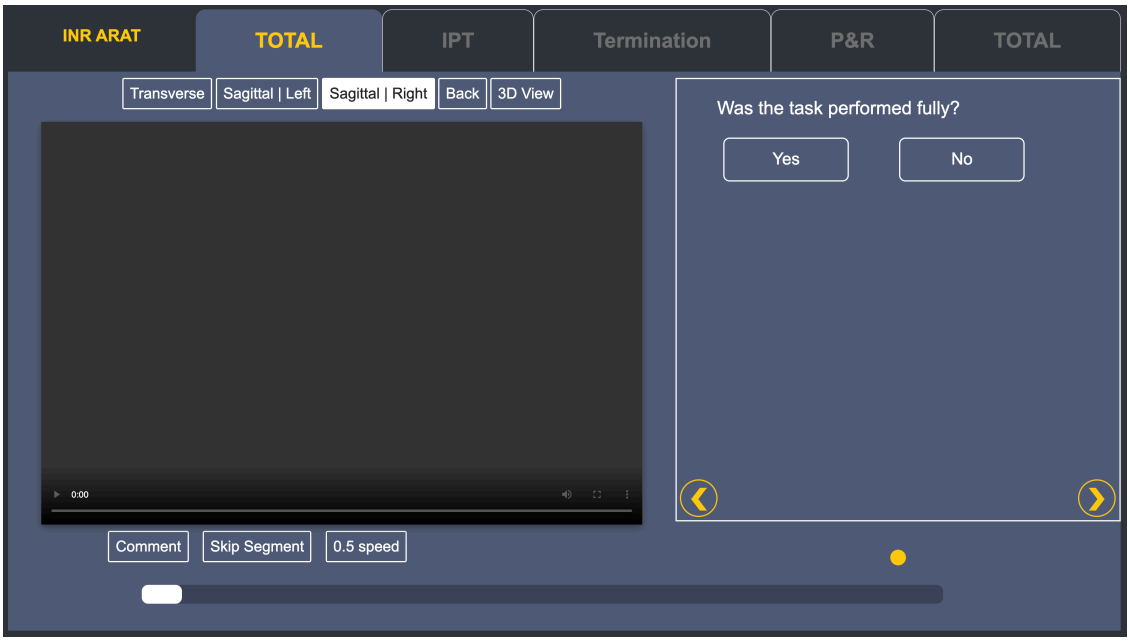

Figure 5.10: Video Application Tool which includes Termination as its own individual segment apart from IPT.

engages with the object (which is focused on the termination) and I automatically default to the zoom camera because it allows them to see the details of what is happening with the hand.

This integration also informs the computational model, by integrating across the layers in which the computational model cannot fully understand or replicate yet. The integration of the zoom-in and half speed can begin to inform the rating and allow the computational model to understand which movement qualities are most prominent for therapist throughout their assessment.

Clinicians can separate functional movement with even slight impairment from non-impaired functional movement (as clearly shown by the low number of 3s in our experiments). However, therapists vary in their assessment of the impact of movement quality on functionality. They partly vary on the movement impairments they focus on during assessment and partly on whether the observed impairments would result in minimal influence on functionality

(i.e., a three at a task level), moderate influence (i.e., a two), or significant impact (i.e.a one). More extensive use of the Socratic approach of the rating interface would produce more samples and improve our statistical understanding of the effect of movement quality on functionality. Therefore, we are expanding and extending this version of the rating interface. However, the detailed quantification of the relation of movement quality to functionality (i.e., the exact amount of hand digit openness that allows for the daily execution of a task) may require integrating the therapist ratings achieved through this version of the rating interface with computationally analyzed kinematic data.

	Study 1: **Score-to-Meaning**	Study 2: **Meaning-to-Score**	Study 3: **Socratic Approach**
Inter-Rater Reliability (IRR) - **Task**	61%	83%	64%
Inter-Rater Reliability (IRR) - **Segment**	45%	43%	68%
Inter-Rater Reliability (IRR) - **Composite**	N/A	46%	68%
Score Distributions	**2 Therapists:** 1 = 48% 2 = 42 % 3 = 10%	**2 Therapists:** 1 = 72% 2 = 20 % 3 = 8%	**2 Therapists:** 1 = 31% 2 = 67 % 3 = 2%
Time Per Session (minutes)	N/A	440.67	390.4

Figure 5.11: Inter-rater reliability and score distribution across study 1

Therapists can readily embrace and utilize a top-down hierarchical approach to a rating that consists of three layers: task layer (which is strongly associated with the assessment of function), the segment layer (partly associated with function and party with movement quality),

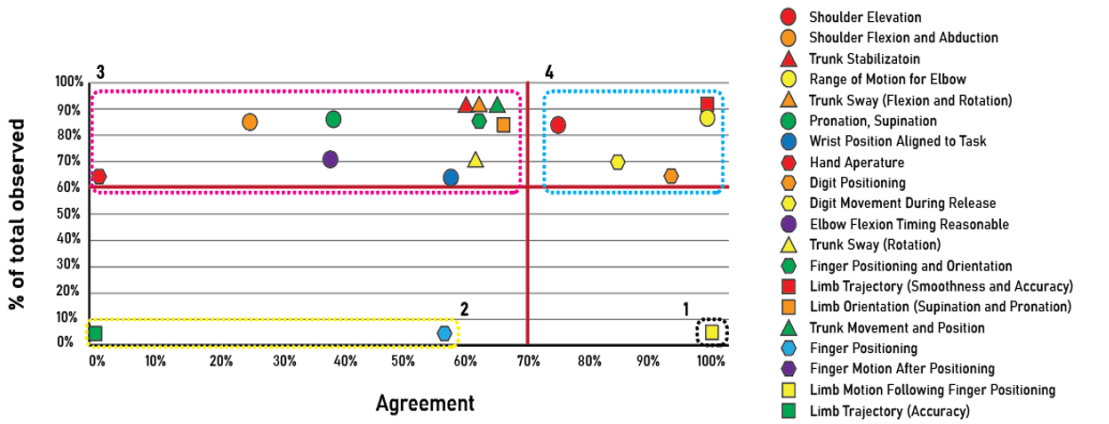

Figure 5.12: Percentage of agreement vs percentage of observation per feature using the two therapist ratings from Experiment 1c: Socratic Approach. In this case we have set the thresholding based on the x axis meaning when both therapist check the same feature (they agreed these features are influencing functionality)

and the composite feature layers (associate with movement quality). This may denote that therapists already use tacit and personalized hierarchical approaches to rating. Therapists can consistently separate even slightly impaired movement from unimpaired movement. But even when using a highly structured approach, therapeutic approaches to the impact of movement quality on functionality have similarities and differences.

Using data from the HBM model as shown in Fig 5.13 into map the level of disagreement across the continuum from severely impaired movement (score of zero) to the unimpaired movement (score of three) both at the task and segment layer will produce a normal distribution (movement impairment on X, disagreement on Y (zero unit disagreement, one unit, two units). In addition, mapping the number of composite feature disagreements will also produce a normal distribution.

There are four possible rating (zero, one, two, three) in which the therapist can give. From figure 5.13, when the rating is a score of two, there is high level of disagreement amongst the therapist, as shown by the corresponding colors. The dataset in which the therapists used

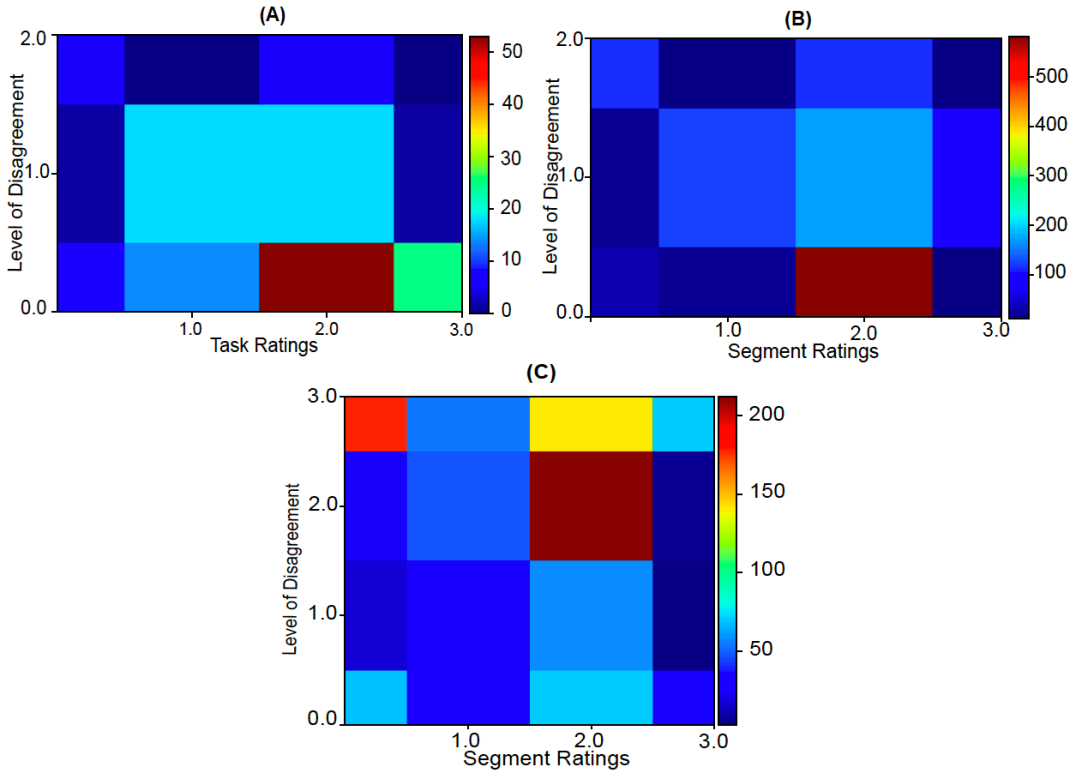

Figure 5.13: Rating scores vs level of disagreement between two therapists on the task, segment and composite feature level; on the x-axis ratings (0,1,2,and 3) are shown and on the y-axis the level of rating disagreement (0,1, and 2 unit) are shown; the color bars show the instances of any particular case;

from Chapter 4 included patients with mild, and moderate impairment. The disagreement is highly due therapists stating the difficulty of distuginishing between one and two or three and a two since the variation depends on the therapist practice and intuition. What we see from the research I conducted in this dissertaion is that we can successful understand and reveal the different combination sets in which therapists use to understand and reveal there tacit knowledge.

When a task is rated as a two as shown in Fig 5.12, expert therapists select a fixed number of combinations as shown in the top-right quadrant. The machine learning model is easily

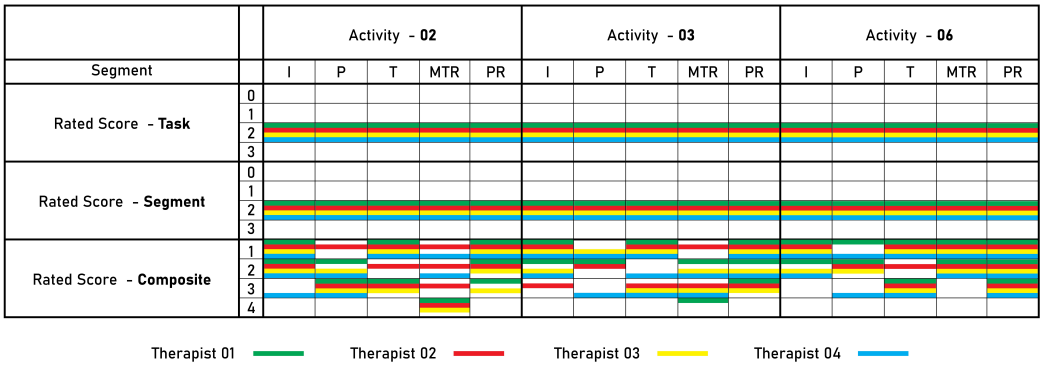

Figure 5.14: The IRR across thee task, segment and composite level for the videos (v=20) expert clinicians rated

able to distinguish between a score of one and a score of three, however a computational model finds it very difficult to understand a score of two. In order to understand the possible combinations, the computational model needs to understand the therapist expertise when deciding to give a patient a score of two. From Fig 5.12 and 5.13, we can see that the relationship between what therapists deem as two vs. a one or three is probabilistic, not deterministic. By the use of the Ri and understanding the therapist's knowledge of the relationship between movement functionality and movement quality, the computational model can solve for this to understand how to generate a score of two by analyzing the expert therapist combinations vs. trying to understand the nine million possible combinations that would generate a score of two. By using the data generated from the Ri, the computational model can begin to understand the prominent feature sets that are more likely to be used for a set of tasks and begin to combine the therapist expert knowledge with the raw kinematic sets. Through this syngerizing of human intelligence and computational intelligence, we can begin to quantitatively visualize tacit knowledge (what therapists are observing in practice) with what the human eye cannot see, to begin to produce results that can empower the therapist with their clinical assessment.

This analysis and generalizability is further proven through the research conducted with ARAT. With the ARAT, I worked with expert clinicians who were not trained on the use of the Ci and Ri. However, by using a participatory design approach and by re-engineering the Ri, we see that across all three layers (task, segment, and composite) there is a probabilistic pattern to understanding and revealing therapist knowledge. In Fig 5.14, I visualize this structure, and the results are replicated just as they did in Chapter 4.

Thus, we propose that relations of movement quality to functionality are statistical and probabilistic. We can continue to inform and reveal these statistical relations by collecting more data through the Ri. Increasing the granularity of observation can also further inform exploring these relations. We propose integrating the statistical hierarchy of therapist ratings with computational analysis of the kinematics and raw features of rehabilitation movement through an HBM. We propose using the results to automate a hierarchical movement assessment that produces interpretable results, where the functionality score is correlated with specific movement quality issues.

Chapter 6

Contributions

Automated assessment of rehabilitation movement during therapy would allow remotely supervised therapy at the home. We are creating an interactive therapy system (SARAH) [40] where a remote therapist can use the system to assign the sequence of tasks to be executed per training sessions at the home and receive automated summaries of performance. Therapists can use automated summaries of therapy task performance at the home along with quantitatively identification of relations between movement changes and function for remote decision making in structuring therapy and remote feedback to the patient. We are planning to test the SARAH system in the home extensively thus generating more data for informing the HBM.

Automated assessment in the clinic can release more time for therapists to focus on delivering therapy. The quantification of movement quality changes and their effect on functionality can help support the therapist decision processes for structuring therapy. In partnership with the Shirley Ryan Ability Lab (SRAL), we have adapted the SARAH system to be used for automated assessment of the ARAT instrument in the clinic using four cameras. We have began the capture of over 100 patients performing the ARAT. The captures will generate standardized and invariant data of upper extremity rehabilitation movement while performing functional tasks. The captured data will further inform the HBM and test the transferability of our automated assessment approaches across the home and the clinic. Results from this work will be discussed in an upcoming publication.

Because our 15 training tasks (and their generalizable segment vocabulary) map well to ADLs, we expect that the relations established through our system between movement changes and task performance can reveal the relation between movement changes and overall daily life function. But the relation between the task layer of our hierarchy and the overall daily life functionality layer needs to be quantitatively developed and verified in a similar manner as the relation of the other layers. Some data for the ADL layer is already capturable through questionnaires as well as simple wearables (like activity monitoring applications on a smartphone). IMU based tracking of the affected limb, hand and torso using methods being developed by our lab [69] and other research [63] can significantly increase observability of the daily activity. We are integrating unobtrusive IMUs units in the SARAH system [69]. The IMUs will be worn on the two wrists, index finger of the affected limb and waist throughout the day including the training sessions at the home. The cross training of the IMU data with the video data and the expert ratings could allow us to identify particular distributions of IMU based kinematics that can be used to recognize particular segment blocks during ADLs. We could then adapt our decision tree algorithm to use these segment information to estimate types of tasks being performed.

6.1 Contributions to HCI Community

Through low cost capture of clinic and home based upper extremity therapy sessions, and the subsequent rating of these sessions by therapists through our intuitive interfaces, we can produce data that increases the observability of the probabilities of our HBM and our ability to train computational algorithms for automated assessment. We have created an interactive therapy system (SARAH) where a remote therapist can use the system to assign the sequence of tasks to be executed per training sessions at the home. The patient can

use the system to start and stop the video capture of each attempted task thus segmenting the video stream of a therapy session into tasks. Knowing the task number gives us the related transition matrix of segments, the appropriate decision tree tuning for assisting in task completion assessment and the conditional probabilities for task, segment, composite feature and kinematic relations. We use these elements to constraint our computational hierarchy so as to provide automated interpretable assessment of performance to patients during home rehabilitation. The automated assessment could provide segment and task scores and denote to patients which composite features of their movement they should focus on during particular segments of particular tasks. This assessment would be driven by previously collected data from sessions of other patients but as we collect data specific to this patient we could tune the probability calculation to weight more kinematics information specific to each patient. Automated assessment of movement during therapy would allow remotely supervised therapy at the home and give more time for therapists to focus on structuring therapy. Therapists could use automated summaries (and their ability to quantitatively identify relations between movement changes and function) for their decision making in structuring therapy, the training of novice therapists and the delivery of home therapy.

The TCE Methodology provides a step-by-step reflective process to create a human-centric approach to building machine learning models. Through my dissertation, I have six contributions to the field of HCI.

1. A reflective thinking and design process to reveal tacit human knowledge in complex sociotechnical contexts.

2. A process by which explicit knowledge is represented through a computable model that enhances human and machine learning intelligence through a reflective process.

3. An HCI methodology integrating human intelligence into complex machine learning

spaces through an iterative design approach that empowers and incentivizes human participation.

4. Developed a methodological approach for the relationship between human and machine intelligence for complex embodied spaces.

5. Understood how to use interface design to encourage reflection while increasing inter-rater reliability in a complex noisy space.

6. Created a standardized approach to capture and assess human movement data for embodied learning scenarios.

Bibliography

[1] Tamim Ahmed, Kowshik Thopalli, Thanassis Rikakis, Pavan Turaga, Aisling Kelliher, Jia-Bin Huang, and Steven L. Wolf. 2021. Automated Movement Assessment in Stroke Rehabilitation. *Frontiers in Neurology* 12 (2021). https://doi.org/10.3389/fneur.2021.720650

[2] Adegboyega Akinsiku, Ignacio Avellino, Yasmin Graham, and Helena M. Mentis. 2021. *It's Not Just the Movement: Experiential Information Needed for Stroke Telerehabilitation.* Association for Computing Machinery, New York, NY, USA, Chapter 3. https://doi.org/10.1145/3411764.3445663

[3] Adegboyega Akinsiku, Ignacio Avellino, Yasmin Graham, and Helena M. Mentis. 2021. It's Not Just the Movement: Experiential Information Needed for Stroke Telerehabilitation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 661, 12 pages. https://doi.org/10.1145/3411764.3445663

[4] SaiDhiraj Amuru, Cem Tekin, Mihaela van der Schaar, and R. Michael Buehrer. 2016. Jamming Bandits—A Novel Learning Method for Optimal Jamming. *IEEE Transactions on Wireless Communications* 15, 4 (2016), 2792–2808. https://doi.org/10.1109/TWC.2015.2510643

[5] Lesley Axelrod, Geraldine Fitzpatrick, Jane Burridge, Sue Mawson, Penny Smith, Tom Rodden, and Ian Ricketts. 2009. The reality of homes fit for heroes: Design challenges for rehabilitation technology at home. *Journal of Assistive Technologies* 3 (07 2009), 35–43. https://doi.org/10.1108/17549450200900014

[6] Belén Barros Pena, Rachel E Clarke, Lars Erik Holmquist, and John Vines. 2021. *Circumspect Users: Older Adults as Critical Adopters and Resistors of Technology*. Association for Computing Machinery, New York, NY, USA, Chapter 3. https://doi.org/10.1145/3411764.3445128

[7] Sara L. Beckman and Michael Barry. 2009. Design and Innovation through Storytelling. *International Journal of Innovation Science* 1, 4 (2009), 151–160.

[8] Tarazi W Bosworth A Ruhter J, Samson LW Sheingold S Taplin C and Zuckerman R. 2020. Beneficiary Use of Telehealth Visits: Early Data from the Start of COVID-19 Pandemic. https://ASPEaspe.hhs.gov

[9] A. E. Bryman. 2008. 6 Why do Researchers Integrate/Combine/Mesh/Blend/Mix/Merge/Fuse Quantitative and Qualitative Research?

[10] R Buchanan. 1992. *How Design Thinking Tools Help To Solve Wicked Problems*.

[11] Pascale Carayon, Sarah Kianfar, Yaqiong Li, Anping Xie, Bashar Alyousef, and Abigail Wooldridge. 2015. A Systematic Review of Mixed Methods Research on Human Factors and Ergonomics in Health Care. *Applied Ergonomics* 51 (11 2015). https://doi.org/10.1016/j.apergo.2015.06.001

[12] Y. Chen, M. Baran, H. Sundaram, and T. Rikakis. 2011. A low cost, adaptive mixed reality system for home-based stroke rehabilitation. *Annu Int Conf IEEE Eng Med Biol Soc* 2011 (2011), 1827–1830.

[13] Yinpeng Chen, Weiwei Xu, Hari Sundaram, Thanassis Rikakis, and Sheng-Min Liu. 2007. Media Adaptation Framework in Biofeedback System for Stroke Patient Rehabilitation. In *Proceedings of the 15th ACM International Conference on Multimedia*

(Augsburg, Germany) *(MM '07)*. Association for Computing Machinery, New York, NY, USA, 47–57. https://doi.org/10.1145/1291233.1291248

[14] Juliet Clark. 2021. *Designing Telehealth Rehabilitation Systems For Diverse Stakeholder Needs*. Master's thesis. Virginia Tech.

[15] Juliet Clark, Setor Zilevu, Tamim Ahmed, Aisling Kelliher, Sai Krishna Yeshala, Sarah Garrison, Cathleen Garcia, Olivia C. Menezes, Minakshi Seth, and Thanassis Rikakis. 2021. Hybrid Workflow Process for Home Based Rehabilitation Movement Capture. In *ACM International Conference on Interactive Media Experiences* (Virtual Event, USA) *(IMX '21)*. Association for Computing Machinery, New York, NY, USA, 241–246. https://doi.org/10.1145/3452918.3465499

[16] Earllaine Croarkin, Jerome V Danoff, and Candice Barnes. 2004. Evidence-based rating of upper-extremity motor function tests used for people following a stroke. *Physical therapy* 84 1 (2004).

[17] Yngve Dahl and Dag Svanæs. 2020. *Facilitating Democracy: Concerns from Participatory Design with Asymmetric Stakeholder Relations in Health Care*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi-org.ezproxy.lib.vt.edu/10.1145/3313831.3376805

[18] Marcello N. de Amorim, Ricardo M.C. Segundo, Celso A.S. Santos, and Orivaldo de L. Tavares. 2017. Video Annotation by Cascading Microtasks: A Crowdsourcing Approach. In *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web* (Gramado, RS, Brazil) *(WebMedia '17)*. Association for Computing Machinery, New York, NY, USA, 49–56. https://doi.org/10.1145/3126858.3126897

[19] Augusto Dias Pereira dos Santos, Lian Loke, and Roberto Martinez-Maldonado. 2018. Exploring Video Annotation as a Tool to Support Dance Teaching *(OzCHI '18)*. Asso-

ciation for Computing Machinery, New York, NY, USA, 448–452. https://doi.org/10.1145/3292147.3292194

[20] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017).

[21] Anca Dumitrache, Lora Aroyo, and Chris Welty. 2018. Capturing Ambiguity in Crowdsourcing Frame Disambiguation. In *HCOMP*. Association for Computing Machinery, New York, NY, USA, 157–170.

[22] Anca Dumitrache, Oana Inel, Benjamin Timmermans, Carlos Martinez Ortiz, Robert-Jan Sips, Lora Aroyo, and Chris Welty. 2018. Empirical Methodology for Crowdsourcing Ground Truth.

[23] Serge Egelman, Ed Chi, and Steven Dow. 2014. *Crowdsourcing in HCI Research*. Association for Computing Machinery, New York, NY, USA, 267–289. https://doi.org/10.1007/978-1-4939-0378-8_11

[24] Pelle Ehn. 2008. Participation in Design Things. In *Proceedings of the Tenth Anniversary Conference on Participatory Design 2008* (Bloomington, Indiana) *(PDC '08)*. Indiana University, USA, 92–101.

[25] Pelle Ehn. 2008. Participation in Design Things. In *Proceedings of the Tenth Anniversary Conference on Participatory Design 2008* (Bloomington, Indiana) *(PDC '08)*. Indiana University, USA, 92–101.

[26] Henry Etzkowitz, Carol Kemelgor, and Brian Uzzi. 2000. *Athena unbound : the advancement of women in science and technology*. Cambridge University Press, Cambridge ; New York. 282 p. pages.

[27] Shelley Evenson and Hugh Dubberly. 2010. *Designing for Service: Creating an Experience Advantage*. Introduction to Service Engineering, USA, 403 – 413. https://doi.org/10.1002/9780470569627.ch19

[28] Michael D. Fetters, Leslie A. Curry, and John W. Creswell. 2013. Achieving Integration in Mixed Methods Designs—Principles and Practices. *Health Services Research* 48, 6pt2 (2013), 2134–2156. https://doi.org/10.1111/1475-6773.12117 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-6773.12117

[29] Rebecca Fiebrink and Marco Gillies. 2018. Introduction to the Special Issue on Human-Centered Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 7 (jun 2018), 7 pages. https://doi.org/10.1145/3205942

[30] Félix Buendía García. 2016. Design of an Annotation Tool for Educational Resources. In *Design of an Annotation Tool for Educational Resources* (Salamanca, Spain) *(TEEM '16)*. Association for Computing Machinery, New York, NY, USA, 1005–1009. https://doi.org/10.1145/3012430.3012639

[31] Joaquín Gayoso-Cabada, Antonio Sarasa-Cabezuelo, and José-Luis Sierra. 2018. Document Annotation Tools: Annotation Classification Mechanisms. In *Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality* (Salamanca, Spain) *(TEEM'18)*. Association for Computing Machinery, New York, NY, USA, 889–895. https://doi.org/10.1145/3284179.3284331

[32] Thomas Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing. *International journal of human-computer studies* 43 (1995), 907–928.

[33] Mostafa Haghi, Kerstin Thurow, and Regina Stoll. 2017. Wearable Devices in Medical Internet of Things: Scientific Research and Commercially Available Devices. *Healthcare Informatics Research* 23 (2017), 4 – 15.

[34] James Hollan, Edwin Hutchins, and David Kirsh. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.* 7, 2 (2000), 174–196.

[35] Lars Erik Holmquist. 2017. Intelligence on Tap: Artificial Intelligence as a New Design Material. *Interactions* 24, 4 (jun 2017), 28–33. https://doi.org/10.1145/3085571

[36] H. Huang, Y. Chen, W. Xu, H. Sundaram, L. Olson, T. Ingalls, T. Rikakis, and J. He. 2006. Novel design of interactive multimodal biofeedback system for neurorehabilitation. *Conf Proc IEEE Eng Med Biol Soc* 1 (2006), 4925–8. https://doi.org/10.1109/IEMBS.2006.260409

[37] Edwin Hutchins. 1995. *Cognition in the wild*. MIT Press, Cambridge, Mass. xviii, 381 p. pages.

[38] Jonas Ivarsson and Todd Nicewonger. 2016. Design Imaginaries: Knowledge Transformation and Innovation in Experimental Architecture. *Mind, Culture, and Activity* 24 (05 2016), 1–14. https://doi.org/10.1080/10749039.2016.1183131

[39] Aisling Kelliher, Andrew Gibson, Eric Bottelsen, and Edward Coe. 2019. Designing Modular Rehabilitation Objects for Interactive Therapy in the Home. In *Proceedings of the Thirteenth International Conference on Tangible, Embedded, and Embodied Interaction* (Tempe, Arizona, USA) *(TEI '19)*. Association for Computing Machinery, New York, NY, USA, 251–257. https://doi.org/10.1145/3294109.3300983

[40] Aisling Kelliher, Setor Zilevu, Thanassis Rikakis, Tamim Ahmed, Yen Truong, and Steven L. Wolf. 2020. Towards Standardized Processes for Physical Therapists to Quantify Patient Rehabilitation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376706

[41] Aisling Kelliher, Setor Zilevu, Thanassis Rikakis, and Steve Wolf. 2019. Towards the development of semi-supervised rehabilitation systems for the home, In Towards the development of semi-supervised rehabilitation systems for the home. *technology,mind, society* 5, 3, 1–4.

[42] Faisal Khan, Bilge Mutlu, and Jerry Zhu. 2011. How Do Humans Teach: On Curriculum Learning and Teaching Dimension. In *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.), Vol. 24. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2011/file/f9028faec74be6ec9b852b0a542e2f39-Paper.pdf

[43] Sabrina Kletz, Andreas Leibetseder, and Klaus Schoeffmann. 2019. A Comparative Study of Video Annotation Tools for Scene Understanding: Yet (Not) Another Annotation Tool. In *A Comparative Study of Video Annotation Tools for Scene Understanding: Yet (Not) Another Annotation Tool* (Amherst, Massachusetts) *(MMSys '19)*. Association for Computing Machinery, New York, NY, USA, 133–144. https://doi.org/10.1145/3304109.3306223

[44] Chia-Lin Koh, I-Ping Hsueh, Wen-Chung Wang, Ching-Fan Sheu, Tzu-Ying Yu, Chun-Hou Wang, and Ching-Lin Hsieh. 2006. Validation of the Action Research Arm Test using item response theory in stroke patients. *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine* 38 (12 2006), 375–80. https://doi.org/10.1080/16501970600803252

[45] J. W. Krakauer. 2005. Arm function after stroke: from physiology to recovery. *Semin Neurol* 25, 4 (Dec 2005), 384–395.

[46] Gert Kwakkel, Boudewijn J Kollen, and Eline Lindeman. 2004. Understanding the

pattern of functional recovery after stroke: facts and theories. *Restorative neurology and neuroscience* 22 3-5 (2004), 281–99.

[47] Edith Law, Krzysztof Z. Gajos, Andrea Wiggins, Mary L. Gray, and Alex Williams. 2017. Crowdsourcing as a Tool for Research: Implications of Uncertainty. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1544–1561. https://doi.org/10.1145/2998181.2998197

[48] N. Lehrer, Y. Chen, M. Duff, S. L Wolf, and T. Rikakis. 2011. Exploring the bases for a mixed reality stroke rehabilitation system, Part II: design of interactive feedback for upper limb rehabilitation. *J Neuroeng Rehabil* 8 (Sep 2011), 54.

[49] N. Lehrer, Y. Chen, M. Duff, S. L Wolf, and T. Rikakis. 2011. Exploring the bases for a mixed reality stroke rehabilitation system, Part II: design of interactive feedback for upper limb rehabilitation. *J Neuroeng Rehabil* 8 (Sep 2011), 54.

[50] Haghi M. and R. Thurow K, Stoll. 2017. Wearable Devices in Medical Internet of Things: Scientific Research and Commercially Available Devices. *Healthc Inform Res* 23, 1 (2017), 4–15. https://doi.org/10.4258/hir.2017.23.1.4

[51] Michelle McDonnell. 2008. Action Research Arm Test. *The Australian journal of physiotherapy* 54 (2008), 220. https://doi.org/10.1016/S0004-9514(08)70034-5

[52] Matthew J Meyer, Shelialah Pereira, Andrew McClure, Robert Teasell, Amardeep Thind, John Koval, Marina Richardson, and Mark Speechley. 2015. A systematic review of studies reporting multivariable models to predict functional outcomes after post-stroke inpatient rehabilitation. *Disability and rehabilitation* 37, 15 (2015), 1316–1323.

[53] Elaheh Momeni, Claire Cardie, and Nicholas Diakopoulos. 2015. A Survey on Assessment and Ranking Methodologies for User-Generated Content on the Web. In *A Survey on Assessment and Ranking Methodologies for User-Generated Content on the Web*, Vol. 48. Association for Computing Machinery, New York, NY, USA, Article 41, 49 pages. https://doi.org/10.1145/2811282

[54] Åsa Nordin, Margit Alt Murphy, and Anna Danielsson. 2014. Intra-rater and inter-rater reliability at the item level of the Action Research Arm Test for patients with stroke. *Journal of rehabilitation medicine* 46 8 (2014), 738–45.

[55] Donald A. Norman. 2002. *The design of everyday things* (1st basic paperback. ed.). Basic Books, New York. xxi, 257 p. pages. Contributorbiographicalinformationhttp://www.loc.gov/catdir/enhancements/fy0830/2003269451-b.htmlPublisherdescriptionhttp://www.loc.gov/catdir/enhancements/fy0830/2003269451-d.html

[56] Carolyn Pang, Zhiqin Collin Wang, Joanna McGrenere, Rock Leung, Jiamin Dai, and Karyn Moffatt. 2021. *Technology Adoption and Learning Preferences for Older Adults: Evolving Perceptions, Ongoing Challenges, and Emerging Design Opportunities.* Association for Computing Machinery, New York, NY, USA, Chapter 3. https://doi.org/10.1145/3411764.3445702

[57] Nicola Plant, Clarice Hilton, Marco Gillies, Rebecca Fiebrink, Phoenix Perry, Carlos González Díaz, Ruth Gibson, Bruno Martelli, and Michael Zbyszynski. 2021. *Interactive Machine Learning for Embodied Interaction Design: A Tool and Methodology.* Association for Computing Machinery, New York, NY, USA, Chapter 12, 4. https://doi.org/10.1145/3430524.3442703

[58] Miriam Rafferty, Laura Stoff, and Celian. [n.d.]. "Stakeholder Discussion Panel Report. ([n. d.]).

[59] Miriam Rafferty, Laura Stoff, and Celian. [n.d.]. "Stakeholder Discussion Panel Report. ([n. d.]).

[60] Vishwajith Ramesh, Andrew Nguyen, Kunal Agrawal, Brett C. Meyer, Gert Cauwenberghs, and Nadir Weibel. 2020. *Assessing Clinicians' Reliance on Computational Aids for Acute Stroke Diagnosis.* Association for Computing Machinery, New York, NY, USA, 146–155. https://doi-org.ezproxy.lib.vt.edu/10.1145/3421937.3422019

[61] D. J. Reinkensmeyer and M. L. Boninger. 2012. Technologies and combination therapies for enhancing movement training for people with a disability. *J Neuroeng Rehabil* 9 (2012), 17. https://doi.org/10.1186/1743-0003-9-17

[62] David J. Reinkensmeyer, Etienne Burdet, Maura Casadio, John W. Krakauer, Gert Kwakkel, Catherine E. Lang, Stephan P. Swinnen, Nick S. Ward, and Nicolas Schweighofer. 2016. Computational neurorehabilitation: modeling plasticity and learning to predict recovery. *Journal of NeuroEngineering and Rehabilitation* 13, 1 (2016), 42. https://doi.org/10.1186/s12984-016-0148-3

[63] Bodine C Reinkensmeyer DJ, Blackstone S. 2017. How a diverse research ecosystem has generated new rehabilitation technologies: Review of NIDILRR's Rehabilitation Engineering Research Centers., In How a diverse research ecosystem has generated new rehabilitation technologies: Review of NIDILRR's Rehabilitation Engineering Research Centers. *J Neuroeng Rehabil* 4, 4, 1. https://doi.org/10.1186/s12984-017-0321-3

[64] T. Rikakis. 2011. Utilizing media arts principles for developing effective interactive neurorehabilitation systems. *Annu Int Conf IEEE Eng Med Biol Soc* 2011 (2011), 1391–1394.

[65] Thanassis Rikakis, Aisling Kelliher, Jinwoo Choi, Jia-Bin Huang, Kris Kitani, Setor Zilevu, and Steven L. Wolf. 2018. Semi-Automated Home-Based Therapy for the Upper Extremity of Stroke Survivors. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference* (Corfu, Greece) *(PETRA '18)*. Association for Computing Machinery, New York, NY, USA, 249–256. https://doi.org/10.1145/3197768.3197777

[66] Thanassis Rikakis, Aisling Kelliher, Jia-Bin Huang, and Hari Sundaram. 2018. Progressive cyber-human intelligence for social good. *Interactions* 25 (06 2018), 52–56. https://doi.org/10.1145/3231559

[67] Joy Robinson, Candice Lanius, and Ryan Weber. 2018. The past, present, and future of UX empirical research. *Commun. Des. Q. Rev* 5, 3 (2018), 10–23. https://doi.org/10.1145/3188173.3188175

[68] Md Abdus Salam, Mary E. Koone, Saravanan Thirumuruganathan, Gautam Das, and Senjuti Basu Roy. 2019. A Human-in-the-Loop Attribute Design Framework for Classification. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1612–1622. https://doi.org/10.1145/3308558.3313547

[69] Anik Sarker, Don-Roberts Emenonye, Aisling Kelliher, Thanassis Rikakis, R. Michael Buehrer, and Alan T. Asbeck. 2022. Capturing Upper Body Kinematics and Localization with Low-Cost Sensors for Rehabilitation Applications. *Sensors* 22, 6 (2022). https://doi.org/10.3390/s22062300

[70] Aziret Satybaldiev, Peter Hevesi, Marco Hirsch, Vitor Fortes Rey, and Paul Lukowicz. 2019. CoAT: A Web-Based, Collaborative Annotation Tool *(UbiComp/ISWC*

'19 Adjunct). Association for Computing Machinery, New York, NY, USA, 5 pages. https://doi.org/10.1145/3341162.3345592

[71] Richard L. Schreiner and Niceta C. Bradburn. 1988. *Care of the newborn* (2nd ed.). Raven Press, New York. xv, 197 p. pages.

[72] Nikola Serbedzija. 2010. Reflective Assistance for Eldercare Environments. In *Proceedings of the 2010 ICSE Workshop on Software Engineering in Health Care* (Cape Town, South Africa) *(SEHC '10)*. Association for Computing Machinery, New York, NY, USA, 104–110. https://doi.org/10.1145/1809085.1809099

[73] Amit Sheth. 2010. Computing for Human Experience: Semantics-Empowered Sensors, Services, and Social Computing on the Ubiquitous Web. *IEEE Internet Computing* 14, 1 (2010), 88–91.

[74] Herbert A. Simon. 1987. *Models of man, social and rational : mathematical essays on rational human behavior in a social setting*. Garland Pub., New York. xiv, 287 p. pages.

[75] Jesper Simonsen and Toni Robertson. 2013. *Routledge international handbook of participatory design*. Vol. 711. Routledge New York.

[76] Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective Practicum: A Framework of Sensitising Concepts to Design for Transformative Reflection. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2696–2707. https://doi.org/10.1145/3025453.3025516

[77] Linda B. Smith and Lauren K. Slone. 2017. A Developmental Approach to Machine Learning? *Frontiers in Psychology* 8 (2017), 15. https://doi.org/10.3389/fpsyg.2017.02124

[78] Chaehan So. 2020. Human-in-the-Loop Design Cycles – A Process Framework That Integrates Design Sprints, Agile Processes, and Machine Learning with Humans. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings* (Copenhagen, Denmark). Springer-Verlag, Berlin, Heidelberg, 136–145. https://doi.org/10.1007/978-3-030-50334-5_9

[79] Cham Springer. 2018. *The Engineering of Experience.* HCI Series, USA. https://doi.org/10.1007/978-3-319-6823-6_18

[80] Niya Stoimenova and Rebecca Price. 2020. Exploring the Nuances of Designing (with-/for) Artificial Intelligence.

[81] Vinay Venkataraman, Pavan Turaga, Nicole Lehrer, Michael Baran, Thanassis Rikakis, and Steven Wolf. 2014. Decision support for stroke rehabilitation therapy via describable attribute-based decision trees. *Conf Proc IEEE Eng Med Biol Soc* 2014 (2014), 3154–9. https://doi.org/10.1109/EMBC.2014.6944292

[82] David Webster and Ozkan Celik. 2014. Systematic review of Kinect applications in elderly care and stroke rehabilitation. *Journal of neuroengineering and rehabilitation* 11, 1 (2014), 1–24.

[83] et al Wolf. [n.d.]. "Model for ARAT Movement Ratings Project. Internal Document". ([n. d.]).

[84] S. L. Wolf, C. J. Winstein, J. P. Miller, P. A. Thompson, E. Taub, G. Uswatte, D. Morris, S. Blanton, D. Nichols-Larsen, and P. C. Clark. 2008. Retention of upper limb function in stroke survivors who have received constraint-induced movement therapy: the EXCITE randomised trial. *Lancet Neurol* 7, 1 (Jan 2008), 33–40.

[85] Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018).

[86] Nuray Yozbatiran, Lucy Der-Yeghiaian, and Steven C. Cramer. 2008. A Standardized Approach to Performing the Action Research Arm Test. *Neurorehabilitation and Neural Repair* 22, 1 (2008), 78–90. https://doi.org/10.1177/1545968307305353 PMID: 17704352.

[87] Kobla Setor Zilevu. 2019. *Interactive Interfaces for Capturing and Annotating Videos of Human Movement Performance*. Master's thesis. Virginia Tech.

[88] Setor Zilevu, Thanassis Rikakis, Aisling Kelliher, Jinwoo Choi, Eric Bottlesen, Jia-Bin Huang, Sarah Garrison, and Steve Wolf. 2018. A Machine Learning Approach for the Quantitative Assessment of the Upper Extremity Movement in Stroke Survivors.

Appendices

Appendix A

First Appendix

MODEL FOR SARAH MOVEMENT RATINGS VERSION 2.0

Introduction This document provides a segmentation and rating approach for rating videos of upper extremity movement of stroke survivors during training tasks. These tasks map well to Activities of Daily Living (ADLs) and are the core training tasks of our Semi Automated Rehabilitation At the Home system (SARAH). The proposed approach aims to makes the rating of movement performance computable (i.e. can drive a machine learning algorithm) and feasible from a therapist point of view (i.e. a group of therapists can be trained to use it in a consistent manner).

The rating rubric uses a small set of standardized segments across all tasks and establishes a maximum of four movement elements per segment that the therapist needs to observe and rate. This make consistent expert rating more feasible. The rubric uses a consistent approach to the rating of the execution of all segments and tasks which allows the development of a consistent hierarchy of rating that can inform computation (automated segmentation, labeling and rating). In an upcoming paper we propose facilitation of computation through a Hierarchical Bayesian Model with four levels (from top to bottom): overall movement impairment (as captured through validated clinical instruments), overall task rating by therapist, segment rating by therapist, raw kinematics assessment through computational means.

The rating system of the segment and overall tasks proposed in this document is aligned

with the ARAT rating approach thus allowing use of a standardized approach across training with SARAH and assessment through ARAT. Rated videos across SARAH and ARAT can also provide consistent training data for the machine learning algorithms.

Movement Segments and Tasks All tasks for SARAH can be segmented using four types of generalizable segments: IPT: Initiation + Progression + Termination (a threes stage segment) M&TR: Manipulate & Transport MTRB is for bimanual CM: Complex Manipulation CMB stands for bimanual manipulation R
&R: Release and Return

All 15 tasks in the SARAH system can be represented by one possible multi-segment path as shown in diagram 1. The multi-segment paths per task are in table 1. The IPT stages are interconnected parts of the starting segment of every reach and grasp/touch task used in SARAH and are thus always grouped together as one segment.

Diagram 1

Table 1 – segment paths per Task

Task 1: I+P+T, R

&R, I+P+T, R

&R I = Initiation Task 2: I+P+T, R

&R, I+P+T, R

&R P = Progression Task 3: I+P+T, M

&TR, M

&TR, R

&R T= Termination Task 4: I+P+T, M

&TR, M

&TR, R

&R M

&TR= Manipulation

& Transportation Task 5: I+P+T, M

&TR, M

&TR, R

&R M

&TRB= Manipulation

& Bimanual Transport. Task 6: I+P+T, M

&TR, R

&R, M

&TR, I+P+T, R

&R CMB= Complex Bimanual Manipulation Task 7: I+P+T, M

&TR, R

&R, I+P+T, M

&TR, R

&R R

&R= Release

& Return Task 8: I+P+T, M

&TR, M

&TR, M

&TR, R

&R Task 9: I+P+T, I+P+T, R

&R Task 10: I+P+T, M

&TR, R

&R, I+P+T, M

&TR, R

&R Task 11: I+P+T, M

&TRB, CMB, R

&R, I+P+T, CMB, M

&TRB, R

&R Task 12: I+P+T, M

&TRB, M

&TRB, R

&R, I+P+T, M

&TRB, M

&TRB, R

&R Task 13: Task 14: Task 15:

Rating the Movement quality of segments The movement quality of each movement stage and/or segment is rated using a set of 3-4 key movement elements per segment as shown below:

Initiation Appropriate shoulder elevation Appropriate shoulder flexion or abduction Appropriate trunk stabilization Progression Appropriate range of motion for elbow Appropriate trunk sway, flexion and rotation Appropriate pronation/supination Termination Wrist position aligned to task Appropriate hand aperture Appropriate digit positioning

Manipulation

& Transportation (M

&TR) or Manipulation

& Bimanual Transportation (TB) appropriate finger positioning and orientation limb trajectory with appropriate smoothness and accuracy limb orientation (supination/pronation) appropriate trunk movement and position

Complex Manipulation

& Transportation (CM

&TR) or Complex Bimanual Manipulation

& T (CMB

&T) appropriate finger positioning appropriate finger motion after positioning appropriate limb motion following finger positioning limb trajectory with appropriate accuracy

Release and Return (R

&R) Appropriate digit movement during release Elbow flexion timing reasonable Appropriate trunk sway/rotation

The Operational Definitions of terms used to evaluate movement quality and inform rating are as follows: Appropriate: The range, direction and timing of the movement component for the task compared to that expected for the less impaired upper extremity Digit Positioning: The volitional placement of relevant digits is representative of what would normally be expected for the task Trunk sway: The forward (translational) movement of the torso is appropriate for the task Aperture: The positioning of separation between the thumb pad and index finger pad is what would be normally expected for the task Jitter: Lack of fluidity of movement especially as the wrist and hand approach or (when required) release the object The movement element being rated must be noticeably different from the same element in unimpaired movement to be marked as impaired. All elements that show such difference need to be marked as impaired.

Rating the execution of the segments and tasks Segment rating 3: segment was completed within reasonable time and all three or four movement quality elements important to the execution of the stage showed no compensation or minimal compensation 2: stage was completed (fully with multiple attempts as needed) but either: execution took too long one or more of the key movement element (s) showed a level of impairment that made

the execution challenging 1: segment was not fully executed because one or more of the key movement elements showed a level of impairment that impeded the full and proper execution significant compensation was used that limited or negated the use of the affective limb in the execution (i.e. segment was primarily executed through use of unimpaired limb) fatigue or other reason (please provide comment for other reason) 0: movement was not attempted or segment could not be initiated (i.e. patient reached object at end of IPT but could not grab object to initiate M

&TR) Overall task rating 3: Task was completed within reasonable time and with minimal movement impairment 2: Task was completed (fully with multiple attempts as needed) but either: the length of execution was unreasonably long upper extremity collective movement showed a level of impairment that made the execution challenging 1: Task segment was not fully executed because upper extremity collective movement showed a level of impairment that impeded the full and proper execution significant compensation was used that limited or negated the use of the affective limb in the execution (i.e. task was primarily executed through use of unimpaired limb) fatigue or other reason (comment for other reason) 0: movement was not attempted or task could not be initiated (patient could not reach or grab object so as to start execution) Unreasonably long is considered a time of execution past the first standard deviation of time of execution of mildly impaired patients

Rating Interface

Currently the rating is done using a custom system our team has developed. We segment each video of a task into the segments given in table 1 (first by hand and gradually automatically) and load it into the rating system. For most movements we provide two angles (upper body from the side and from the front). All rating pages also provide comment space at the bottom. (Please note that the stages I+P+T are always shown as one segment by the system – the separation into three separate stages for rating purposes needs to be done by

the rating therapist).

The interface shows the whole task first and asks for an overall rating using a multiple choice table that utilizes the standardized rating approach outlined in this document. We then show each segment (I+P+T, manipulation transportation, complex manipulation, release and return) and ask for the therapist to review the key movement components per movement stage and rate them as impaired or not impaired using the standardized rating approach outlined in this document. The therapist can check which movement elements show significant impairment by clicking on a button next to the element (s). We then show the whole task again and ask for a revised total rating (if the revised total rating is different from the original we ask for a reason).

A.1 Section one

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor

sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Appendix B

Second Appendix

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

RATING RUBRIC MODEL FOR the Action Research Arm Test (ARAT)

I. Introduction This document provides a segmentation and rating approach (rating rubric) for rating videos of upper extremity movement of stroke survivors performing the ARAT. The proposed approach aims to make the rating of movement performance computable (i.e. can drive a machine learning algorithm) and feasible from a therapist point of view (i.e. a group of therapists can be trained to use it in a consistent manner).

The rating rubric allows therapists to rate both performance of segments of movement and overall task performance. The rating happens offline using a new custom interface being developed for this project. The interface allows therapists to utilize 3 points of view (from 3 different cameras) to view the movement, view each segment separately and repeat viewing of tasks and segments as needed. The therapists provide ratings through an intuitive multiple choice approach.

The rating rubric establishes a small set of standardized segments across all ARAT tasks and establishes a maximum of four movement elements per segment that the therapist needs to observe and rate. This approach makes consistent expert rating more feasible. The rubric thus advances a uniform approach to the rating of the execution of all segments and tasks which allows the development of a consistent hierarchy of rating that can inform the standardization of rating by experts and the use of standardized expert rating for the development of automated analysis and rating algorithms. Rated segments are also easier to connect to raw assessment of kinematics through computational means, thus allowing for the development of a robust hierarchical model of rating (task, segment, kinematics) that combines expert and computational intelligence. The resulting computational rating tools will assist the therapist in assessment thus allow therapists to focus more of their effort on therapy. The tools can also be used for training of therapists learning the assessment.

The rating rubric proposed in this document is aligned with the rating approach being used in our Home Based Rehabilitation System (SARAH), thus allowing use of a standardized approach across training with SARAH and assessment through ARAT. Rated videos across SARAH and ARAT can also provide consistent training data for the machine learning algorithms.

II. Movement Segments and Tasks The goal is to segment all ARAT assessment tasks using three (or four) types of generalizable segments: IPT: Initiation + Progression + Termination (a three-stage segment) Gross IPT, GIPT: for tasks 17-19 M&TR: Manipulate & Transport P&R: Place and Release

Table 1 – segment paths per Task

GRASP SUBSCALE Task 1: I+P+T, M&TR, P&R (Block 10 cm) Task 2: I+P+T, M&TR, P&R (Block 2.5 cm) Task 3: I+P+T, M&TR, P&R (Block 5 cm) Task 4: I+P+T, M&TR,

P&R (Block 7.5 cm) Task 5: I+P+T, M&TR, P&R (Cricket Ball) Task 6: I+P+T, M&TR, P&R (Sharpening Stone) GRIP SUBSCALE Task 7: I+P+T, M&TR, P&R (Glass-Water) Task 8: I+P+T, M&TR, P&R (Tube 2.25 cm) Task 9: I+P+T, M&TR, P&R (Tube 1 cn) Task 10: I+P+T, M&TR, P&R (Washer) PINCH SUBSCALE Task 11: I+P+T, M&TR, P&R (ball bearing between ring finger and thumb) Task 12: I+P+T, M&TR, P&R (marble between index finger and thumb) Task 13: I+P+T, M&TR, P&R (ball bearing between middle finger and thumb) Task 14: I+P+T, M&TR, P&R (ball bearing between index finger and thumb) Task 15: I+P+T, M&TR, P&R (marble between ring finger and thumb) Task 16: I+P+T, M&TR, P&R (marble between middle finger and thumb) GROSS MOVEMENT SUBSCALE Tasks 17, 18, 19: GIPT

III. Rating the Movement quality of segments The movement quality of segments is rated using a set of 4 key movement elements per segment as shown below. Each of the three stages of IPT is rated using only 3 movement elements for each Initiation + Progression+Termination · Initiation o Appropriate shoulder elevation o Appropriate shoulder flexion or abduction o Appropriate trunk stabilization · Progression o Appropriate range of motion for elbow o Appropriate trunk stabilization o Appropriate forearm pronation/supination · Termination o Wrist position aligned to task o Appropriate hand aperture o Appropriate digit positioning – for GIPT replace with Appropriate location on head

Manipulation & Transportation (M&TR) o Appropriate digit positioning and orientation o Appropriate smoothness and accuracy in limb trajectory o Appropriate forearm orientation (supination/pronation) o Appropriate trunk stabilization Place and Release (P&R) o Accurate final placement of object as appropriate for the task o Appropriate digit movement during release o Appropriate limb orientation (pronation/supination) during placement and release o Appropriate trunk stabilization

The Operational Definitions of terms used to evaluate movement quality and inform rating

are as follows: Appropriate: The range, direction and timing of the movement component for the task compared to that demonstrated or expected for the less impaired upper extremity (demonstrated denotes that the task has been performed by the less impaired limb) Appropriate Digit Positioning: The volitional placement of relevant digits is representative of what would normally be expected for the task Appropriate digit positioning for pinch subscale requires the use of specific fingers (given in task instructions) as well as use of pads of indicated thumb/fingers Appropriate hand aperture: The positioning of separation between the thumb pad and index and middle finger pads is what would be normally expected for the task Trunk Stablilization: the forward (translational) and rotational movement of the torso is appropriate for the task

The movement element being rated must be noticeably different from the same element in the less impaired arm movement to be marked as impaired. All elements that show such difference need to be marked as impaired.

IV. Rating the execution of the segments and tasks Segment rating 3: segment was completed within reasonable time and all key movement quality elements important to the execution of the segment showed no or minimal compensation (reasonable time for segment performance means that the segment duration was such that would allow for the completion of the whole task within 5 seconds) 2: segment was completed (fully with multiple attempts as needed) but either: execution took too long (timing of segment was such that would not allow for the execution of the overall task within 5 seconds) one or more of the key movement element (s) showed a level of impairment that made the execution challenging 1: segment was not fully executed because one or more of the key movement elements showed a level of impairment that impeded the full and proper execution significant compensation was used that limited or negated the use of the affective limb in the execution (i.e. segment was primarily executed through use of torso compensation) fatigue or other reason (please

provide comment for other reason) 0: movement was not attempted or segment could not be initiated (i.e. patient reached object at end of IPT but could not grab object to initiate M&TR) Overall task rating 3: Task was completed within 5 seconds and with minimal movement impairment 2: Task was completed (fully with multiple attempts as needed) but either: o the length of execution was unreasonably long (between 5 and 60 seconds) (see Yozbatiran et al) o upper extremity collective movement showed a level of impairment that made the execution challenging

1: Task was not fully executed Tasks 1-16 (first three subscales): IPT segment was completed but the M&TR and/or P&R segment of the task could not be completed because upper extremity collective movement showed a level of impairment that impeded the full and proper execution of M&TR and/or P&R significant compensation was used that limited or negated the use of the affective limb in the execution (i.e. task was primarily executed through use of torso compensation) fatigue or other reason (comment for other reason) Tasks 17-19 (gross motor subscale): IPT segment was initiated but could not be completed because upper extremity collective movement showed a level of impairment that impeded the full and proper execution significant compensation was used that limited or negated the use of the affective limb in the execution (i.e. task was primarily executed through use of torso compensation or head compensation) fatigue or other reason (comment for other reason) 0: movement was not attempted or task could not be initiated (i.e. patient could not reach or grab object, could not complete IPT, so as to start execution)

V. Rating Interface Currently the rating is done using a custom system our team has developed. We segment each video of a task into the segments given in table 1 (first by hand and gradually automatically) and load it into the rating system. (Please note that the stages I+P+T are always shown as one segment by the system – the separation into three separate stages for rating purposes needs to be done by the rating therapist). The interface provides

three viewing angles to the rating therapist: hand – contralateral view of impaired limb head, torso and arm (profile) – ipsilateral view, back of body For tasks 16-19 two viewing angles are provided (front and side). The interface shows the whole task first and asks for an overall rating using a multiple choice table that contains the standardized rating approach outlined in this document. We then show each segment (I+P+T, manipulation & transportation, place & release) and ask for the therapist to review the key movement elements per movement segment and rate them as impaired or not impaired using the standardized rating approach outlined in this document. The therapist can check which movement elements show significant impairment by clicking on a button next to the element (s). We then show the whole task again and ask for a revised total rating (if the revised total rating is different from the original we ask for a reason).