# Survey of Domain-Specific Performance Measures in Assistive Robotic Technology

Katherine M. Tsui and Holly A. Yanco
University of Massachusetts Lowell
Department of Computer Science
1 University Avenue
Lowell, MA, USA
{ktsui, holly}@cs.uml.edu

David J. Feil-Seifer and Maja J. Matarić
University of Southern California
Department of Computer Science
941 West 37th Place
Los Angeles, CA, USA
{dfseifer, mataric}@usc.edu

## ABSTRACT

Assistive robotics have been developed for several domains, including autism, eldercare, intelligent wheelchairs, assistive robotic arms, external limb prostheses, and stroke rehabilitation. Work in assistive robotics can be divided into two larger research areas: *technology development*, where new devices, software, and interfaces are created; and *clinical application*, where assistive technology is applied to a given end-user population. Moving from technology development towards clinical applications is a significant challenge. Developing performance metrics for assistive robots can unveil a larger set of challenges. For example, what well established performance measures should be used for evaluation to lend credence to a particular assistive robotic technology from a clinician's perspective? In this paper, we survey several areas of assistive robotic technology in order to demonstrate domain-specific means for evaluating the performance of an assistive robot system.

## Categories and Subject Descriptors

A.1 [**Introductory and Survey**]

## General Terms

Performance measures

## Keywords

Assistive technology, human-robot interaction, robotics, end-user evaluation

## 1. INTRODUCTION

Assistive robotics may have therapeutic benefits in domains ranging from autism to post-stroke rehabilitation to eldercare. However, it can be challenging to transition an assistive device developed in the lab to the target domain. This problem can occur even when the device was designed with a specific end user in mind. Römer et al. provided guidelines for compiling a technical file for an assistive device for transfer from academic development to manufacturing [52]. Their guidelines state that documentation of an assistive device must include its "intended use, design specifications, design considerations, design methods, design calculations, risk analysis, verification of the specifications, validation information of performance of its intended use, and compliance to application standards" [52]. Academic and industrial research labs are the piloting grounds for new concepts. However, due to the institutional separation between the research environment and end-users, special care must be taken so that a finished project properly addresses the needs of end-users. As such, it is imperative for the development of assistive robotic technologies to involve the end-user in the design and evaluations [28]. These end-user evaluations, with the proper performance measures, can provide the basis for performance validation needed to begin the transition from research pilot to end product.

Does there exist a ubiquitous set of performance measures for the evaluation of assistive robotic technologies? Time to task completion or time on task are common measures. Römer et al. propose an absolute measure for time to task completion, where the time is normalized with an able-bodied person's performance [52]. Task completion time fits many robotic applications, such as retrieving an object with a robotic manipulator. However, it may not suit other applications, such as a range of motion exercise in the rehabilitation of an upper limb. Römer et al. also acknowledge other factors in determining performance measures, namely "user friendliness, ease of operation, [and] effectiveness of input device" [52].

Aside from the very general metrics described above, should there even be a ubiquitous set of performance metrics? This lack of a ubiquitous set has occurred in part because each domain has very specific needs in terms of performance. Most metrics do not translate well between domains or even subdomains. The field of assistive robotics technology has used a wide variety of performance measures specific to domains for end-user evaluations. However, there are observable similarities between various employed metrics and how they are devised. In order to evaluate an assistive robotic technology within a particular domain, clinical performance measures are needed to lend validity to the device.

Clinical evaluation is the mechanism used to determine the clinical, biological, or psychological effects of an evaluated intervention. Clinical evaluations use The Good Clin-


*PerMIS'08* August 19–21, 2008, Gaithersburg, MD, USA.

ical Practice Protocol, which requires clearly stated objectives, checkpoints, and types and frequency of measurement [68]. Well established domains can have well established performance measures. For example, the Fugl-Meyer motor assessment, created in 1975 [25], is commonly used when evaluating upper limb rehabilitation for patients post-stroke. FIM [42] is popular when measuring the function independence of a person with respect to activities of daily living (ADLs). The two evaluations have little correlation, if any, to each other because they are domain-specific. However, they are both used for studying potential end-users that do not use assistive technology, and can serve as an effective method for assessing performance relative to the established baseline.

In this paper, we explore contemporary end-user evaluations and the performance measures used in evaluating assistive robotic technology. We detail the performance measures and discuss for which evaluations and contexts they would be appropriate.

# 2. ASSISTIVE ROBOTIC TECHNOLOGIES

Haigh and Yanco surveyed assistive robotics in 2002 [30]. A historical survey of rehabilitation robotics through 2003 can be found in Hillman [32]. Simpson surveyed intelligent wheelchairs through 2004 [57]. We present a contemporary survey of assistive technologies that have been evaluated by end-users. We believe that the primary focus of end-user evaluations should be on the *human* performance measurements, and secondarily on the performance of the robot. This section highlights six areas of assistive technology development: autism; eldercare; intelligent wheelchairs, assistive robotic arms; prosthetic limbs; and post-stroke rehabilitation. For each area, we describe a few examples of performance metrics and how they are employed/applied.

## 2.1 Autism Spectrum Disorder

An increasing number of research institutions are investigating the use of robots as a means of interaction with children with autism spectrum disorder (ASD), including the National Institute of Information and Communications Technology [37], University of Hertfordshire [49, 50, 48], Université de Sherbrooke [43, 53], University of Southern California [21], University of Washington [60], and Yale University [55, 56]. The goal of these systems is to use robots as a means of affecting the social and communicative behavior of children with autism for either assessment or therapeutic purposes.

### *2.1.1 End-user Evaluations*

The University of Hertfordshire has conducted several observation studies with children with ASD [49]. In one study, four children interacted with Robota, a robot doll, over a period of several months. Post-hoc analysis of video footage of interaction sessions yielded eye gaze, touch, imitation, and proximity categories. Performance measures included frequency of the occurrence of the categories. Another study used the hesitation and duration of a drumming session as a task-specific measure of engagement with a drumming robot [50]. In addition, measures for observing social behavior were taken from existing work from the autism research community regarding methods for using video coding for observing social behavior [64] to determine if a robot was an isolator or mediator for children with autism [48].

The Université of Sherbrooke conducted an observation study of four children with autism spectrum disorder over seven weeks [43]. The children interacted with Tito, a human-character robot, three times per week for five minutes. Video was collected during the interactions. In post-hoc analysis, the interactions were categorized into shared attention, shared conventions, and absence of shared attention or conventions; all video data were coded using twelve-second windows. Performance measures included frequency of the occurrence of categories. Other work involved the use of automated interaction logs in order to model a user's play behavior with the robot [53]. Performance measures included correlation of recognized play with observed behavior.

The National Institute of Information and Communications Technology (NICT) conducted a longitudinal observation study in a day-care setting [37]. Groups of children interacted with a simple character robot, Keepon, in twenty five three-hour sessions over five months. Each session was a free-play scenario that was part of the regular day-care schedule. Children were given the opportunity to interact with the robot, or not, and children were allowed to interact with the robot in groups. Video of these interactions was recorded and analyzed in a qualitative fashion. In particular, they observed changes in dyadic interaction between the child, the robot, and peers.

The University of Southern California (USC) conducted a study with children with autism interacting with a bubble-blowing robot [20]. This research uses a repeated-measures study to compare two types of robot behavior, contingent (the robot responds to the child's actions) and random (the robot executes an action after a random amount of time has passed). The scenario involved the child, the robot, and a parent observed for forty-five minutes. Post-hoc analysis of video data was used to identify joint-attention, vocalizations, social orienting, and other forms of social interaction, identified by target (parent, robot, or none). These behaviors were taken from a diagnostic exam, the Autism Diagnostic Observation Schedule (ADOS) [40], which uses a similar scenario to the one used in the experiment, providing a key for identifying relevant evaluative behavior. Results from this work supported the hypothesis that a robot behaving contingently provoked more social behavior than a robot behaving randomly. Performance measures included frequency and richness of the interaction observed between sessions.

The University of Washington developed a study that compared a robot dog, AIBO, to a simple mechanical stuffed dog [60]. After a brief introductory period, the participants, parent and child, interacted with the one of the artifacts for a period of thirty minutes. The sessions were videotaped, and coded for behavior. The behavior coding included verbal engagement, affection, animating artifact, reciprocal interaction, and authentic interaction. The data were compared between sessions with each dog. The performance measure used was the amount of coded social behavior observed.

Yale University has been developing robots for diagnostic and therapeutic applications for children with autism. Specifically, they are developing passive sensing techniques along with robots designed to exhibit social "presses" in order to provoke and observe the behavior of children with autism [55]. One example of this approach was the use of observing gaze behavior as a means for providing diagnostic information [56]. In one study, children were outfitted with eye-tracking equipment and their gaze was tracked

with various visual and auditory stimuli. This experiment tested both children with autism and typically developing children. The performance measure for this study was to determine if the gaze tracker could identify significant differences between the gaze patterns of children with autism and typically developing children. Another study compared affective prosody given from either a human or robot speech therapist [36].

### 2.1.2 Analysis

One common technique for measuring performance in the ASD domain is coding, followed by a post-hoc analysis to create keywords, phrases, or categories from video data [51]. Categories and definitions are defined from these units. The data, such as open ended responses to questions or recorded, can be annotated with the categories. To ensure reliability, multiple coders are trained on the units and definitions. When multiple coders are used, inter-rater reliability needs to be established, usually assessed using Cohen's kappa [12]. However, in each case, the basic unit of time for behavior data could be vastly different, ranging from tenths of a second [49], to twelve seconds [43], to assessments of the entire session [37]. The resulting performance measures use the number of occurrences within the categories.

While these assessments are in most cases driven by existing tools used in developmental or autism-specific settings, there is little evidence shown so far that the measures used translate well to real-world improvements in learning, social skill development, and psychosocial behavior. It is important to note that autism is considered a spectrum disorder and that there is a great deal of heterogeneity to the population [24]. While studies can show effects for a small subgroup of children, it is important to analyze how generalizable the results are. One strategy for ensuring that the observed data are somewhat grounded in the field of autism research is to draw the analysis metrics from existing communities [51, 20].

## 2.2 Eldercare

Studies show that the elderly population is growing worldwide [6]. Roboticists from research institutions, such as NICT [70], USC [63], and University of Missouri [72] are investigating robots for use as minders, guides, and companions.

### 2.2.1 End-user Evaluations

The University of Missouri in conjunction with TigerPlace, an eldercare facility, studied assistive technology for aging in place [72], where residents who would otherwise be required to have full-time nursing-home care are able to live in their current residence and have health services brought to them instead. As part of this effort, they developed a fuzzy-logic augmentation of an existing day-to-day evaluation, the Short Physical Performance Battery (SPPB) [29]. This test measures the performance for balance, gait, strength, and endurance.

NICT conducted a five-week study of twenty three elderly women interacting with Paro, the therapeutic care robot seal, in an eldercare facility. Interaction occurred one to threes times per week [70]. Performance measures included self assessment of the participant's mood (pictorial Likert scale [39] of 1 (happy) to 20 (sad)) before and after the interaction with Paro; questions from the Profile of Mood States questionnaire [41] to evaluate anxiety, depression, and vigor (Likert scale of 0 (none) to 4 (extremely)); and urinary specimens to measure stress.

Researchers at the USC are currently developing a robot for exercise therapy in adults suffering from dementia [63]. Exercise therapy was part of the regular care regiment provided by the staff at the nursing home location of the experiment, but keeping the elders engaged in the task was a challenge for the staff. The experiment scenario involves using a robot to demonstrate, coach, and monitor exercises. The real-world performance measure for success is compliance to the exercise regimen, measured by time on task (from recorded video data post-hoc), or overall health of the residents. Initial studies involved using a focus group to assess resident's reactions to the robot. For the focus group interaction, performance was measured by the number of residents showing willingness to interact with the robot.

### 2.2.2 Analysis

Most of the above systems are currently at the feasibility stage of implementation, an important stage of evaluation for determining if the technology is ready for deployment in a real-world environment. User and behavior studies of eldercare systems, such as with Paro, serve to describe the effects that such systems have on users and their environment. By emphasizing social interaction and fitness, these performance measures implicitly measure changes in quality of life (QoL).

Current evaluations of eldercare systems occur over a period of days or weeks. As these systems become more permanent fixtures in eldercare environments, the assessment of QoL becomes more important. There exist standardized questionnaires for observing QoL at multiple points of time. Therefore, QoL can be a good method of observing the long-term effectiveness of a change in the eldercare environment [76]. For example, the SF-36 survey [1] is used to assess health-related QoL, while the 15-D [59] survey is used to measure QoL along several elements of a subject's lifestyle.

## 2.3 Intelligent Wheelchairs

Intelligent wheelchairs can potentially improve the quality of life for people with disabilities. Research has focused on autonomous and semi-autonomous collision-free navigation and human-robot interaction (i.e., novel input devices and intention recognition) and has been conducted by both research institutions and companies.

### 2.3.1 End-user Evaluations

In 2005, MobileRobots (formerly ActivMedia) and the University of Massachusetts Lowell evaluated the Independence – Enhancing Wheelchair (IEW) [45, 46] with several end-users at a rehabilitation center. The original testing design planned to use of a maze-like obstacle course made of cardboard boxes. However, this scenario did not work well with the participants. They were frustrated by a maze that was not like their regular driving environments and viewed boxes as moveable objects.

Instead, the participants operated the IEW as they would typically use a wheelchair in their everyday lives (e.g., going to class which entailed moving through corridors with other people and passing through doorways). The performance measure, number of hits/near misses and time on task, was not modified. The results have not yet been published.

End-user trials have also been completed by intelligent wheelchair companies, such as DEKA [16] and CALL Centre [8], seeking government approval to prove the safety of these systems. The University of Pittsburgh has conducted an evaluation of DEKA's iBOT with end-users [13].

### 2.3.2 Analysis

In the domain of intelligent wheelchairs, the majority of user testing has been in the form of feasibility studies with able-bodied participants. As noted in Yanco [77], able-bodied participants are more easily able to vocalize any discomforts and stop a trial quickly. These pilot experiments pave the way for end-user trials.

One barrier to end-user trials of robotic wheelchair systems is the need for the use of a participant's seating on the prototype system. While seating can be moved from the participant's wheelchair to the prototype system (if compatible) and back, this seating switch can take thirty to sixty minutes in each direction, making multiple testing sessions prohibitive.

We discuss performance measures commonly used thus far in feasibility studies. One of the most common tests of an autonomous intelligent wheelchair is passing through a doorway [58]. Passing through a doorway without collision is one of seven "environmental negotiations" that a person must perform in order to be prescribed a power wheelchair for mobility [67]. Other tasks include changing speed to accommodate the environment (e.g., cluttered = slow), stopping at closed doors and drop offs (e.g., stairs and curbs), and navigating a hallway with dynamic and stationary objects (e.g., people and furniture).

In the case of these power mobility skills, the user is rated based on his/her ability to *safely* complete the task. In contrast, robotic performance measures are not binary. Performance measures include time to completion (i.e., time to pass through the doorway), number of interactions, and number of collisions. Recent performance measures include accuracy, legibility, and gracefulness of the motion used to pass through the doorway [9, 62].

## 2.4 Assistive Robotic Arms

Robotic arms can improve the quality of life by aiding in activities of daily living (ADLs), such as self-care and pick-and-place tasks. Robotic arms can be used in fixed workstations, placed on mobile platforms, or mounted to wheelchairs. Research focuses both on building robot arms and the design of human-robot interaction. One topic of interest is retrieving an object from a shelf or floor (i.e., pick-and-place task), one of the most common ADLs [61]. Institutions investigating assistive robotic arms include Clarkson University [26], Delft University [65], Stanford University [71], University of Massachusetts Lowell [66], University of Pittsburgh [11], and TNO Science & Industry [65].

### 2.4.1 End-user Evaluations

Stanford University conducted an experiment with twelve spinal cord injury patients on two user interfaces for ProVAR, a vocational workstation [71]. After using each interface, each participant answered an evaluation questionnaire. Performance measures included open-ended responses to positive and negative questions on the robot's appearance, navigation, ease of use, error messages, complexity, usefulness, and functionality, and also on the participant's satisfaction.

The University of Pittsburgh evaluated the effects of a Raptor arm, a commercially available wheelchair-mounted robotic arm, on the independence of eleven spinal cord injury patients [11]. Participants first completed sixteen ADLs without the Raptor arm, then again after initial training, and once more after thirteen hours of use. At each session, the participants were timed to task completion and classified as *dependent*, *needs assistance*, or *independent*.

Clarkson University evaluated eight multiple sclerosis patients over five ADLs with and without the Raptor arm [26]. The participants in this study all required assistance with self-care ADLs. Participants were evaluated before and after training on the Raptor arm. At each session, the participants were timed to task completion and interviewed. They also rated the level of difficulty of task performance and the Psychosocial Impact of Assistive Devices Scale (PIADS) [15].

University of Massachusetts Lowell conducted an experiment of a new visual human-robot interface for the Manus Assistive Robotic Manipulator (ARM), a commercially available European robot arm. Eight individuals who used wheelchairs and had cognitive impairments participated in an eight week controlled experiment to control the robot arm in a pick-and-place task. Performance measures included time to task completion (i.e., object selection time), level of attention, level of prompting (based on measurement of functional independence [42]), and survey responses (i.e., preference of interface, improvements).

TNO Science & Industry and Delft University conducted a four person case study [65]. The end-users were people who use power wheelchairs and have weak upper limb strength and intact cognition. TNO Science & Industry evaluated their alternative graphical user interface for the Manus ARM. The performance measures included number of mode switches, task time, Rating Scale of Mental Effort (RSME) [78], and survey responses.

### 2.4.2 Analysis

As demonstrated by Tsui et al. [66], Tijsma et al. [65], and Fulk et al. [26], it is also important to account for the user's experience with respect to cognitive workload and mental and emotional state. The basis for the user's experience performance measure must be derived or adapted from an existing clinical measure.

In Tsui et al. [66] and Tijsma et al. [65], the participants were rated or rated themselves with respect to cognitive workload. In Tsui et al. [66], the level of prompting during a trial was a cognitive measure based on FIM, which is a scale that measures functional independence [42]. A person is rated on a Likert scale (1 = needs total assistance to 7 = has complete independence) on a variety of ADLs. FIM may also be applied as a cognitive measure to activities such as "comprehension, expression, social interaction, problem solving, and memory" [42]. In Tijsma et al. [65], RSME was used as a cognitive performance measure. RSME is a 150 point scale measuring the mental effort needed to complete a task, where 0 = no effort and 150 = extreme effort. The Standardized Mini-Mental State Examination [47] is another cognitive performance measures used in older adults.

In Fulk et al. [26], participants explicitly ranked the perceived difficulty of the task and their mental and emotional state were recorded using PIADS. PIADS is a twenty six item questionnaire in which a person rates their perceived

experience after completing a task with an assistive technology device [14]. It measures the person's feelings of competence, willingness to try new things, and emotional state. PIADS is well established and significantly used in the US and Canada [14]. An alternative emotional performance measure is the Profile of Mood States [41] used in Wada et al. [70].

## 2.5 External Limb Prostheses

Robotic prostheses can serve as limb replacements. Research institutions, such as Hong Kong Polytechnic University [38], Massachusetts Institute of Technology [3], Northwestern University [44], and the Rehabilitation Institute of Chicago [44], have investigated creating new robotic prosthetics and control strategies.

### 2.5.1 End-user Evaluations

The Rehabilitation Institute of Chicago (RIC) and Northwestern University conducted a clinical evaluation of six individuals who underwent targeted muscle reinnervation surgery [44]. After the upper limb prosthetic device was optimally configured for each patient's electromyography signals (EMG), functional testing occurred after the first month, third month, and sixth month. The functional testing was comprised of a series of standard tests: box and blocks, clothespin relocation, Assessment of Motor and Process Skills (AMPS) [23], and the University of New Brunswick prosthetic function [54]. Performance measures included time to complete task, accuracy, and AMPS score.

Researchers at the Massachusetts Institute of Technology conducted a clinical evaluation with three unilateral, transtibial amputees [3]. Data collection included oxygen consumption, carbon dioxide generation, joint torque, and joint angle. Kinematic and kinetic data were collected using a motion capture system for the ankle-foot prosthesis and unaffected leg. The resulting performance measures were metabolic cost of transport (using oxygen consumption as a parameter), gait symmetry between the legs, vertical ground reaction forces, and external work done at the center of mass of each leg.

Hong Kong Polytechnic University conducted a clinical evaluation with four transtibial amputees over the course of three consecutive days [38]. Data collected included motion capture and open-ended responses about the participant's comfort and the prosthesis' stability, ease of use, perceived flexibility, and weight. Stance time, swing time, step length, vertical trunk motion, and average velocity were derived from the motion capture data. Performance measures included ranking of the prostheses used (with respect to comfort, stability, ease of use, perceived flexibility, and weight), gait symmetry, and ground force reactions.

### 2.5.2 Analysis

Performance measures involving ADLs can be used in evaluating prostheses because ADLs include functions such a locomotion and self-care activities. Locomotion includes walking and climbing stairs, and self-care activities involve a high level of dexterity. Heinemann et al. [31] proposed the Orthotics and Prosthetics Users' Survey (OPUS). Burger et al. [7] in turn evaluated the Upper Extremity Functional Status of OPUS with sixty one users with unilateral, upper limb amputations and found that the scale was suitable for the measuring functionality of the population. The Upper Extremity Function Status is comprised of twenty three ADLs, rated in a Likert scale fashion (0 = unable to complete, 3 = very easy to complete. Similarly, AMPS is also comprised of ADLs but in a more flexible fashion; there are eighteen categories of ADLs with up to eleven choices within a category [2]. Another measure of quality of life is FIM, which is comprised of eighteen ADLs.

## 2.6 Stroke Rehabilitation

Robots are being investigated for gait training at Arizona State University [73], upper-limb recovery at RIC and Northwestern University [33], and wrist rehabilitation at Hong Kong Polytechnic University [34]. It is well documented that stroke patients regain most of their mobility through repetitions of task training [35]. Many researchers are investigating the use of robots as a way to augment current rehabilitation strategies for post-stroke patients.

### 2.6.1 End-user Evaluations

An example of a typical rehabilitation robot study using stroke patients was conducted by RIC and Northwestern University of the Therapy Wilmington Robotic Exoskeleton (T-WREX). The team conducted a clinical evaluation of twenty three stroke survivors over sixteen weeks comparing robot-assisted therapy to a traditional rehabilitation therapy regimen [33]. The researchers observed functional arm movement, quality of affected arm use, range of motion, grip strength, a survey of patient satisfaction of therapy, and the use of the affected arm in the home when not undergoing therapy. Performance assessments with or without the robot included Fugl-Meyer [25] and Rancho Functional Test for Upper Extremity [74] to measure ability to use the arm. In addition, they measured use of the arm outside of the experimental setting by using the Motor Activity Log [69], a self-report, to determine how the arm was used in the home. Finally, to assess the costs of using the robot, they measured the amount of time that the user needed assistance in order to use the T-WREX.

The early stages of rehabilitation robot development involves evaluations of the performance of the robot in a pilot setting. Some evaluations are users studies, where the robot is used with small number of users to determine what needs to be altered [73]. Performance measures used involve satisfaction surveys, measures of robustness, and analyses of the quantifiability of sensor data for clinical purposes. These measures are specific to the robot being evaluated, and in general cannot be used in the field in general.

The primary assessment of post-stroke rehabilitative robotics involves the use of clinical assessments of patient function. Discussed above was the Fugl-Meyer and Rancho Functional Test. However, there are many others used. At at Northwestern University and RIC, Ellis et al. [18] supplemented the Fugl-Meyer with several other measures, including the Chedokee McMaster Stroke Assessment, the Reaching Performance Scale, and the Stroke Impact Scale. At Hong Kong Polytechnic University, Hu et al. [34] used four other measures: the Motor Status Score (MSS, used to assess shoulder function) [22], the Modified Ashworth Scale (MAS, used to measure of increase of muscle tone) [4], the Action Research Arm Test (ARAT, used to assess grasp, grip, pinch, and gross movement) [17], and FIM (used to asses functionality in ADLs) [42]. These performance measures provide the picture of the clinical definition of effectiveness.

### 2.6.2 Analysis

Stroke rehabilitation is an established medical domain. Thus, the evaluations of these experiments use relevant clinical evaluations to determine the effectiveness of the robot-augmented therapy. The scope of rehabilitative robotics for patients post-stroke is quite large, ranging from upper-limb recovery to gait training and wrist rehabilitation. Even within a domain, the specific performance measures differ depending on the therapy and may not translate well to another sub-domain. For example, the MSS is applicable to the T-WREX [33] upper-arm rehabilitative aid, but not evaluating gait rehabilitation.

Functional evaluations, such as the Fugl-Meyer [27] and Wolf Motor Function [75], are crucial to comparing the effectiveness of robot-augmented therapies to one another in addition to comparing them with non-robot augmentations for current therapies. It is through these comparisons that robots can truly be evaluated as a rehabilitative device.

## 3. CONCLUSIONS

We believe that performance measures should be specific to the domain and relevant to the task. Domains with clear, well-established medical or therapeutic analogs can leverage existing clinical performance measures. For example, the Fugl-Meyer motor assessment, founded in 1975 [25], is popular when evaluating upper limb rehabilitation of post-stroke patients. Domains without strong therapeutic analogs can appropriately borrow clinical performance measures. Alternatively, they may draw inspiration from a clinical performance measure to create a new one or augment an existing one if none of the existing measures are appropriate [29].

Further, we believe that evaluations conducted with end-users should focus at least as highly on *human* performance measures as they do on system performance measures. By placing the emphasis on human performance, it becomes possible to correlate system performance with human performance. Celik et al. has taken the important first steps for stroke-rehabilitation by examining trajectory error and smoothness of motion with respect to Fugl-Meyer [10]. Similarly, Brewer et al. has used machine learning techniques on sensor data to predict the score of a person with Parkinson's disease on the Unified Parkinson Disease Rating Scale (UPDRS) [5, 19].

Existing performance measures for most of assistive robotic technologies do not provide sufficient detail for experimental and clinical evaluations. We provide a summary of performance measures used (see Table 1) and offer guidelines as to choosing appropriate and meaningful performance measures:

- Consult a clinician who specializes in the particular domain, if possible.
- Choose an appropriate clinical measure for the domain. A domain's "gold standard" will provide the best validity to clinicians.
- Choose an appropriate method to capture a participant's emotional and mental state.
- Consider an appropriate quality of life measurement.
- Administer the human performance measures at least before and after the experiment.
- Consider coding open ended responses, comments, and/or video.
- Concretely define each enumeration in a Likert scale.

By choosing meaningful performance measures, robotics researchers provide a common ground for interpretation and acceptance by the clinical community. In addition, the researchers of a given system are also given clear guidelines for how to observe and define performance of a given system.

Through this survey, we seek other well-established performance measures to apply to assistive robotic technologies. Common performance measurements will allow researchers to both compare the state of the art approaches within specific domains and also to compare against the state of the practice within the field outside of the robotics community.

## 4. ACKNOWLEDGMENTS

This work is funded in part by the National Science Foundation (IIS-0534364, CNS-0709296) and the Nancy Laurie Marks Family Foundation.

## 5. REFERENCES

[1] N. K. Aaronson, C. Acquadro, J. Alonso, G. Apolone, D. Bucquet, M. Bullinger, K. Bungay, S. Fukuhara, B. Gandek, S. Keller, D. Razavi, R. Sanson-Fisher, M. Sullivan, S. Wood-Dauphinee, A. Wagner, and J. E. Ware Jr. Intl. Quality of Life Assessment (IQOLA) Project. *Quality of Life Research*, 1(5):349–351, 2004.

[2] AMPS.com. Assess of Motor and Process Skills. In *http://www.ampsintl.com/tasks.htm*, 2008.

[3] S. Au. *Powered Ankle-Foot Prosthesis for the Improvement of Amputee Walking Economy*. PhD thesis, MIT, 2007.

[4] R. Bohannon and M. Smith. Interrater reliability of a modified Ashworth scale of muscle spasticity. *Physical Therapy*, 67(2):206–7, 1987.

[5] B. R. Brewer, S. Pradhan, G. Carvell, P. Sparto, D. Josbeno, and A. Delitto. Application of Machine Learning to the Development of a Quantitative Clinical Biomarker for the Progression of Parkinson's Disease. In *Rehab. Eng. Society of North America Conf.*, 2008.

[6] J. Brody. Prospects for an ageing population. *Nature*, 315(6019):463–466, 1985.

[7] H. Burger, F. Franchignoni, A. Heinemann, S. Kotnik, and A. Giordano. Validation of the orthotics and prosthetics user survey upper extremity functional status module in people with unilateral upper limb amputation. *J. of Rehab. Medicine*, 40(5):393–399, 2008.

[8] CALL Centre. Smart wheelchair. In *http://callcentre.education.ed.ac.uk/Smart_WheelCh/smart_wheelch.html*, 2008.

[9] T. Carlson and Y. Demiris. Human-wheelchair collaboration through prediction of intention and adaptive assistance. In *IEEE Intl. Conf. on Robotics and Automation*, 2008.

[10] O. Celik, M. K. O'Malley, C. Boake, H. Levin, S. Fischer, and T. Reistetter. Comparison of robotic and clinical motor function improvement measures for sub-acute stroke patients. In *IEEE Intl. Conf. on Robotics and Automation*, 2008.

[11] E. Chaves, A. Koontz, S. Garber, R. Cooper, and A. Williams. Clinical evaluation of a wheelchair mounted robotic arm. Technical report, Univ. of Pittsburgh, 2003.

[12] J. A. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[13] R. Cooper, M. Boninger, R. Cooper, A. Dobson, J. Kessler, M. Schmeler, and S. Fitzgerald. Use of the Independence 3000 IBOT Transporter at home and in the community. *J. of Spinal Cord Medicine*, 26(1):79–85, 2003.

**Table 1: Summary of Performance Measures Used in Assistive Robotic Technology Domains**

| Domain | Performance Measures Used |
|---|---|
| Autism | Behavior coding, correlate sensor modeling of behavior to human-rated behavior |
| Eldercare | Activities of daily living (SBBP, etc.), mood, stress (Standardized Mini-Mental State), quality of life (SF-36, 15-D, etc.) |
| Intelligent Wheelchairs | Number of hits/near misses, time on task, accuracy, gracefulness |
| Assistive Robotic Arms | Activities of daily living, time to task completion, mental state (RSME, Profile of Mood States), attention, level of prompting |
| Prostheses | Functional tests (AMPS, OPUS, FIM, etc.), measures of effort (oxygen consumption, etc.), accuracy, time to complete task, comfort, ease of use |
| Post-Stroke Rehabilitation | Functional measures (Fugl-Meyer, MSS, ARAT, FIM, Chedokee McMaster, Reaching Performance Scale, MAS, Wolf Motor, etc.), use of affected limb in home (Motor Activity Log, etc.) |


[14] H. Day and J. Jutai. Piads in the world. In *http://www.piads.ca/worldmapshtm/worldmap.asp*, 2008.

[15] H. Day, J. Jutai, and K. Campbell. Development of a scale to measure the psychosocial impact of assistive devices: lessons learned and the road ahead. *Disability and Rehab.*, 24(1–3):31–37, 2002.

[16] DEKA Research and Development Corporation. DEKA Evolved Thinking. In *http://www.dekaresearch.com*, 2008.

[17] W. DeWeerdt and M. Harrison. Measuring recovery of arm-hand function in stroke patients: a comparison of the Brunnstrom-Fugl-Meyer test and the Action Research Arm Test. *Physiother Can*, 37(2):65–70, 1985.

[18] M. D. Ellis, T. Sukal, T. DeMott, and J. P. A. Dewald. ACT$^{3D}$ exercise targets gravity-induced discoordination and improves reaching work area in individuals with stroke. In *IEEE Intl. Conf. on Rehab. Robotics*, 2007.

[19] S. Fahn, R. Elton, et al. Unified Parkinson's Disease Rating Scale. *Recent developments in Parkinson's disease*, 2:153–163, 1987.

[20] D. J. Feil-Seifer and M. J. Matarić. Robot-assisted therapy for children with autism spectrum disorders. In *Conf. on Interaction Design for Children: Children with Special Needs*, 2008.

[21] D. J. Feil-Seifer and M. J. Matarić. Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders. In *Intl. Symposium on Experimental Robotics*, 2008.

[22] M. Ferraro, J. Demaio, J. Krol, C. Trudell, K. Rannekleiv, L. Edelstein, P. Christos, M. Aisen, J. England, and S. Fasoli. Assessing the Motor Status Score: A Scale for the Evaluation of Upper Limb Motor Outcomes in Patients after Stroke. *Neurorehabilitation and Neural Repair*, 16(3):283, 2002.

[23] A. Fisher. AMPS: Assessment of Motor and Process Skills Volume 1: Development, Standardisation, and Administration Manual. *Ft Collins, CO: Three Star Press Inc*, 2003.

[24] B. Freeman. Guidelines for Evaluating Intervention Programs for Children with Autism. *J. of Autism and Developmental Disorders*, 27(6):641–651, 1997.

[25] A. Fugl-Meyer, L. Jaasko, I. Leyman, S. Olsson, and S. Steglind. The post-stroke hemiplegic patient. 1. a method for evaluation of physical performance. *Scandinavian J. of Rehab. Medicine*, 7(1):13–31, 1975.

[26] G. Fulk, M. Frick, A. Behal, and M. Ludwig. A wheelchair mounted robotic arm for individuals with multiple sclerosis. Technical report, Clarkson Univ., 2005.

[27] D. Gladstone, C. Danells, and S. Black. The Fugl-Meyer Assessment of Motor Recovery after Stroke: A Critical Review of Its Measurement Properties. *Neurorehabilitation and Neural Repair*, 16(3):232, 2002.

[28] D. Greenwood, W. Whyte, and I. Harkavy. Participatory Action Research as a Process and as a Goal. *Human Relations*, 46(2):175–192, 1993.

[29] J. Guralnik, E. Simonsick, L. Ferrucci, R. Glynn, L. Berkman, D. Blazer, P. Scherr, and R. Wallace. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. *J. of Gerontology*, 49(2):M85–94, 1994.

[30] K. Haigh and H. A. Yanco. Automation as caregiver: A survey of issues and technologies. In *AAAI-2002 Workshop on Automation as Caregiver: The Role of Intelligent Technology in Elder Care*, 2002.

[31] A. W. Heinemann, R. K. Bode, and C. O'Reilly. Development and measurement properties of the orthotics and prosthetics users' survey (opus): a comprehensive set of clinical outcome instruments. *Prosthetics and Orthotics Intl.*, 27(3):191–206, 2003.

[32] M. Hillman. Rehabilitation robotics from past to present - a historical perspective. In *IEEE 8th Intl. Conf. on Rehab. Robotics*, 2003.

[33] S. J. Housman, V. Le, T. Rahman, R. J. Sanchez, and D. J. Reinkensmeyer. Arm-training with t-wrex after chronic stroke: Preliminary results of a randomized controlled trial. In *IEEE Intl. Conf. on Rehab. Robotics*, 2007.

[34] X. L. Hu, K. Y. Tong, R. Song, X. j. Zheng, I. F. Lo, and K. H. Lui. Myoelectrically controlled robotic systems that provide voluntary mechanical help for persons after stroke. In *IEEE Intl. Conf. on Rehab. Robotics*, 2007.

[35] W. Jenkins and M. Merzenich. Reorganization of neocortical representations after brain injury: a neurophysiological model of the bases of recovery from stroke. *Progress in Brain Research*, 71:249–66, 1987.

[36] E. S. Kim, E. Newland, R. Paul, and B. Scassellati. A Robotic Therapist For Positive, Affective Prosody in High-Functioning Autistic Children. In *Poster Pres. at the Intl. Meeting for Autism Research*, 2008.

[37] H. Kozima and C. Nakagawa. Longitudinal child-robot interaction at preschool. In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, pages 27–32, 2007.

[38] W. Lee, M. Zhang, P. Chan, and D. Boone. Gait Analysis of Low-Cost Flexible-Shank Trans-Tibial Prostheses. *IEEE*

*Trans. on Neural Systems and Rehab. Eng.*, 14(3):370–377, 2006.

[39] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140(5):1–55, 1932.

[40] C. Lord, S. Risi, L. Lambrecht, E. H. C. Jr., B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter. The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *J. of Autism and Developmental Disorders*, 30(3):205–223, 2000.

[41] D. M. McNair, M. Lorr, and L. F. Droppleman. Profile of mood states. In *Educational and Industrial Testing Service*, 1992.

[42] MedFriendly.com. Functional independence measure. In *http://www.medfriendly.com/functionalindependencemeasure.html*, 2007.

[43] F. Michaud, T. Salter, A. Duquette, H. Mercier, M. Lauria, H. Larouche, and F. Larose. Assistive technologies and child-robot interaction. In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, 2007.

[44] L. Miller, K. Stubblefield, R. Lipschutz, B. Lock, and T. Kuiken. Improved Myoelectric Prosthesis Control Using Targeted Reinnervation Surgery: A Case Series. *IEEE Trans. on Neural Systems and Rehab. Eng.*, 16(1):46–50, 2008.

[45] MobileRobots Inc. Independence-enhaning wheelchair. In *http://www.activrobots.com/RESEARCH/wheelchair.html*, 2008.

[46] MobileRobots Inc. Robotic chariot. In *http://activrobots.com/robots/robochariot.html*, 2008.

[47] D. Molloy and T. Standish. A Guide to the Standardized Mini-Mental State Examination. *Intl. Psychogeriatrics*, 9(S1):87–94, 2005.

[48] B. Robins, K. Dautenhahn, and J. Dubowsky. Robots as Isolators or Mediators for Children with Autism? A Cautionary Tale. In *AISB05: Social Intelligence and Interaction in Animals, Robots and Agents*, 2005.

[49] B. Robins, K. Dautenhahn, R. te Boekhorst, and A. Billard. Robots as assistive technology – does appearance matter? In *IEEE Int. Workshop on Robot and Human Interactive Communication*, 2004.

[50] B. Robins, K. Dautenhahn, R. te Boekhorst, and C. Nehaniv. Behaviour Delay and Robot Expressiveness in Child-Robot Interactions: A User Study on Interaction Kinesics. In *Intl. Conf. on Human-Robot Interaction*, 2008.

[51] R. Robins, C. Fraley, and R. Krueger. *Handbook of Research Methods in Personality Psychology*. Guilford Press, 2007.

[52] G. Römer and H. Stuyt. Compiling a Medical Device File and a Proposal for an Intl. Standard for Rehabilitation Robots. *IEEE Intl. Conf. on Rehab. Robotics*, pages 489–496, 2007.

[53] T. Salter, F. Michaud, D. Létourneau, D. Lee, and I. Werry. Using proprioceptive sensors for categorizing interactions. In *Human-Robot Interaction*, 2007.

[54] E. Sanderson and R. Scott. UNB test of prosthetic function: a test for unilateral amputees [test manual]. *Fredericton, New Brunswick, Canada, Univ. of New Brunswick*, 1985.

[55] B. Scassellati. How Social Robots Will Help Us to Diagnose, Treat, and Understand Autism. *Robotics Research: Results of the 12th Intl. Symposium ISRR*, 28:552–563, 2007.

[56] F. Shic, B. Scassellati, D. Lin, and K. Chawarska. Measuring context: The gaze patterns of children with autism evaluated from the bottom-up. *Intl. Conf. on Development and Learning*, pages 70–75, 2007.

[57] R. Simpson. Smart wheelchairs: A literature review. *J. of Rehab. Research Development*, 42(4):423–36, 2005.

[58] R. C. Simpson. *Improved automatic adaption through the combination of multiple information sources*. PhD thesis, Univ. of Michigan, Ann Arbor, 1997.

[59] H. Sintonen. The 15-d measure of health related quality of life: Reliability, validity and sensitivity of its health state descriptive system. Working Paper 41, Center for Health Program Evaluation, 1994.

[60] C. Stanton, P. Kahn, R. Severson, J. Ruckert, and B. Gill. Robotic animals might aid in the social development of children with autism. In *Intl. Conf. on Human Robot Interaction*, pages 271–278, 2008.

[61] C. Stranger, C. Anglin, W. S. Harwin, and D. Romilly. Devices for assisting manipulation: A summary of user task priorities. *IEEE Trans. on Rehab. Eng.*, 4(2):256–265, 1994.

[62] T. Taha, J. V. Miró, and G. Dissanayake. Pompdp-based long-term user intention prediction for wheelchair navigation. In *IEEE Intl. Conf. on Robotics and Automation*, 2008.

[63] A. Tapus, J. Fasola, and M. J. Matarić. Socially assistive robots for individuals suffering from dementia. In *Human-Robot Interaction Intl. Conf., Workshop on Robotic Helpers: User Interaction, Interfaces and Companions in Assistive and Therapy Robotics*, 2008.

[64] C. Tardif, M. Plumet, J. Beaudichon, D. Waller, M. Bouvard, and M. Leboyer. Micro-analysis of social interactions between autistic children and normal adults in semi-structured play situations. *Intl. J. of Behavioral Development*, 18(4):727–747, 1995.

[65] H. Tijsma, F. Liefhebber, and J. Herder. Evaluation of new user interface features for the manus robot arm. In *IEEE Intl. Conf. on Rehab. Robotics*, pages 258–263, 2005.

[66] K. Tsui, H. Yanco, D. Kontak, and L. Beliveau. Development and evaluation of a flexible interface for a wheelchair mounted robotic arm. In *Intl. Conf. on Human Robot Interaction*, 2008.

[67] Univ. of Illinois Chicago. Power mobility skills checklist. In *http://internet.dscc.uic.edu/forms/0534.pdf*, 2008.

[68] US Food and Drug Administration. Guidance for industry, E6 good clinical practice: consolidated guidance. *Federal Register*, 10:691–709, 1997.

[69] G. Uswatte, E. Taub, D. Morris, K. Light, and P. Thompson. The Motor Activity Log-28: Assessing daily use of the hemiparetic arm after stroke. *Neurology*, 67(7):1189, 2006.

[70] K. Wada, T. Shibata, T. Saito, and K. Tanie. Effects of robot-assisted activity for elderly people and nurses at a day service center. *IEEE*, 92(11):1780–1788, 2004.

[71] J. Wagner and H. Van der Loos. Training strategies for the user interface of vocational assistive robots. *Intl. Conf. on Eng. in Medicine and Biology Society, EMBC*, 2, 2004.

[72] S. Wang, J. Keller, K. Burks, M. Skubic, and H. Tyrer. Assessing Physical Performance of Elders Using Fuzzy Logic. *IEEE Intl. Conf. on Fuzzy Systems*, pages 2998–3003, 2006.

[73] J. A. Ward, S. Balasubramanian, T. Sugar, and J. He. Robotic gait trainer reliability and stroke patient case study. In *IEEE Intl. Conf. on Rehab. Robotics*, 2007.

[74] D. Wilson, L. Baker, and J. Craddock. Functional test for the hemiparetic upper extremity. *American J. of Occupational Therapy*, 38(3):159–64, 1984.

[75] S. Wolf, P. Thompson, D. Morris, D. Rose, C. Winstein, E. Taub, C. Giuliani, and S. Pearson. The EXCITE Trial: Attributes of the Wolf Motor Function Test in Patients with Subacute Stroke. *Neurorehabilitation and Neural Repair*, 19(3):194, 2005.

[76] S. Wood-Dauphinee. Assessing quality of life in clinical research: From where have we come and where are we going? *J. of Clinical Epidemiology*, 52(4):355–363, 1999.

[77] H. Yanco. Evaluating the performance of assistive robotic systems. *Performance Metrics for Intelligent Systems (PerMIS) Workshop*, pages 21–25, 2002.

[78] F. Zijlstra. *Efficiency in work behaviour: A design approach for modern tools*. PhD thesis, Delft Univ., 1993.