Performance Evaluation Methods for Assistive Robotic Technology

Katherine M. Tsui, David J. Feil-Seifer, Maja J. Matarić, and Holly A. Yanco

Abstract Robots have been developed for several assistive technology domains, including intervention for Autism Spectrum Disorders, eldercare, and post-stroke rehabilitation. Assistive robots have also been used to promote independent living through the use of devices such as intelligent wheelchairs, assistive robotic arms, and external limb prostheses. Work in the broad field of assistive robotic technology can be divided into two major research phases: *technology development*, in which new devices, software, and interfaces are created; and *clinical*, in which 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 poses a related set of challenges. In this paper, we survey several areas of assistive robotic technology in order to derive and demonstrate domain-specific means for evaluating the performance of such systems. We also present two case studies of applied performance measures and a discussion regarding the ubiquity of functional performance measures across the sampled domains. Finally, we present guidelines for incorporating human performance metrics into end-user evaluations of assistive robotic technologies.

1 Introduction

Assistive robots have the potential to provide therapeutic benefits in health care domains ranging from intervention for Autism Spectrum Disorders to post-stroke rehabilitation to eldercare. However, it is invariably challenging to transition an as-

Katherine Tsui and Holly Yanco
University of Massachusetts Lowell, Department of Computer Science, One University Avenue, Lowell, MA, USA, e-mail: {ktsui, holly}@cs.uml.edu

David Feil-Seifer and Maja Matarić
University of Southern California, Department of Computer Science, 3650 McClintock Ave, Los Angeles, CA, USA, e-mail: {dfseifer, mataric}@usc.edu

sistive 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 [80]. 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" [80]. Academic and industrial research labs are the piloting grounds for new concepts. Due to the institutional separation between the research environment and end-users, special care must be taken so that a technology, developed in the lab, properly addresses the needs of end-users in the real world. It is thus imperative for the development of assistive robotic technologies to involve the end-user in the design and evaluations [44]. 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 and time on task are common measures. Römer et al. propose an absolute measure for time to task completion, in which the time is compared to that of an able-bodied person's performance [80]. 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 context of 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" [80].

Aside from the very general metrics described above, should we even seek a ubiquitous set of performance metrics? The lack of ubiquitous performance metrics is a result of necessarily domain-specific performance needs. Most metrics do not translate well between domains or even sub-domains. Thus, the field of assistive robotic technology has used a wide variety of performance measures specific to the 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 intervention. Clinical evaluations use The Good Clinical Practice Protocol, which requires clearly stated objectives, checkpoints, and types and frequency of measurement [101]. Well-established domains have developed generally agreed-upon performance measures over time. For example, the Fugl-Meyer motor assessment, created in 1975, is commonly used in evaluating upper limb rehabilitation for patients post-stroke recovery [40]. On the other hand, FIM (formerly known as the Functional Independence Measure) is popular for measuring a person's functional independence with respect to activities of daily living (ADLs) [65]. The two evaluations have little, if any, relation to each other, because they emerged from different domains. However, they are both used broadly, albeit for

different user populations, and thusly can serve as a means for assessing performance relative to an established baseline.

In this paper, we explore contemporary end-user evaluations and the performance measures used in evaluating assistive robotic technologies. We present case studies from the University of Massachusetts Lowell and the University of Southern California. These studies illustrate the evolution of performance metrics in their respective domains: assistive robotic arms and Autism Spectrum Disorders. We also discuss the ubiquity of functional performance measures throughout all of the surveyed domains; we say that a performance measure is *functional* if it relates to an activity of daily living and is administered in a realistic setting. Finally, we present guidelines for incorporating human performance metrics into end-user evaluations of assistive robotic technologies.

2 Assistive Robotic Technologies

Assistive technology encompasses both "low-tech" and "high-tech" solutions. As a new technology is developed, new and/or improved assistive devices can be created. For example, the concept of a wheelchair was documented in China in the 6th century [108]. The manual self-propelled wheelchair was patented in 1894 [108]. The power wheelchair was invented during World War II [8].

As the field of robotics has matured, researchers began to apply this newest technology to surgery and rehabilitation. More recently, robots are being used to enhance the functional capabilities of people with physical and/or cognitive disabilities. For example, the first commercially available intelligent wheelchair entered the market in 2000 [14].

In 2002, Haigh and Yanco surveyed assistive robotics [47]. A historical survey of rehabilitation robotics through 2003 can be found in Hillman [49]. Simpson surveyed intelligent wheelchairs through 2004 [83]. 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 with secondary focus on the performance of the robot. This section highlights six areas of assistive robotic technology development. We discuss assistive robots used in intervention for Autism Spectrum Disorders, eldercare, post-stroke recovery, and independent living through intelligent wheelchairs, assistive robotic arms, and prosthetic limbs. For each area, we describe a few examples of performance metrics and how they have been employed or applied.

2.1 Autism Spectrum Disorders (ASD)

An increasing number of research institutions are investigating the use of robots as tools for intervention and therapy for children with Autism Spectrum Disorders (ASD), including the University of Hertfordshire [77, 78, 76], the Université de Sherbrooke [66, 81], the National Institute of Information and Communications Technology [56], the University of Southern California [34], and the University of Washington [88].The goal of these systems is typically to use robots as catalysts for social behavior in order to stimulate and train social and communicative behaviors of children with ASD for either assessment or therapeutic purposes.

2.1.1 End-User Evaluations

Researchers at the University of Hertfordshire have conducted several observational studies with children with ASD [77]. In one such 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 [78]. In addition, measures for observing social behavior were taken from the ASD research community; in particular, video coding for observing social behavior [94] was applied to determine if a robot was an isolator or mediator for children with ASD [76].

Researchers at the Université of Sherbrooke conducted an observational study of four children with ASD over seven weeks [66]. 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 either; 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 [81]. Performance measures included correlation of recognized play with observed behavior.

The National Institute of Information and Communications Technology (NICT) conducted a longitudinal observational study in a day-care setting [56]. 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 data of these interactions were analyzed in a qualitative fashion.

Researchers at the University of Southern California (USC) conducted a study with children with ASD interacting with a bubble-blowing robot [33]. This research used a repeated-measures study model to compare two types of robot behavior: contingent (the robot responds to the child's actions) and random (the robot executes an action at random times). The scenario involved the child, the robot, and a par-

ent, all of whom were 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, as well as the tagged by target (parent, robot, or none) of the interaction. These behaviors were taken from a standard ASD diagnostic exam, the Autism Diagnostic Observation Schedule (ADOS) [61], which uses a similar scenario to the one used in the experiment, providing a key for identifying relevant evaluative behavior. Performance measures included frequency and richness of the interaction observed between sessions.

Researchers at the University of Washington developed a study that compared a robot dog, AIBO, to a simple mechanical stuffed dog [88]. After a brief introductory period, the participants (i.e., parent and a child with ASD) interacted with the one of the artifacts for a period of thirty minutes. The sessions were video recorded and coded. The behavior coding included verbal engagement, affection, animating artifact, reciprocal interaction, and authentic interaction. The performance measure used was the amount of coded social behavior observed.

2.1.2 Discussion

Video coding is a commonly used technique for analyzing behavioral experiments [79]. Categories may be set prior to coding or may be the result of post hoc analysis, in which the categories are defined from keywords, phrases, or events. The data, such as open-ended responses to questions or comments, is then annotated with the categories. To ensure reliability, multiple coders (or raters) are trained on the units and definitions. When multiple coders are used, inter-coder reliability must be established, such as by using a kappa statistic.[1] However, in each experiment design, the basic unit of time for behavior data could be vastly different, ranging from tenths of a second (e.g., [77]) to twelve seconds (e.g., [66]) to assessments of the entire session (e.g., [56]). 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 ASD-specific settings, there is little evidence to date that the measures used that translate well to real-world improvements in learning, social skill development, and psychosocial behavior. ASD is considered a spectrum disorder with a great deal of symptom heterogeneity in the population [39], which creates a major challenge for diagnosis and treatment as well as research. Since assistive robotics studies to date have shown some effects for small groups of tested children, it is important to analyze how generalizable their results are. One strategy for ensuring that the observed data are somewhat grounded in the field of ASD research is to draw the analysis metrics from existing ASD diagnostics (e.g., [79] and [33]). This remains an open challenge for the growing field of socially assistive robotics for ASD.

[1] Cohen's kappa provides the level of agreement for nominal data between two raters [21]. For more than two raters Fleiss' kappa must be used [37].

2.2 Eldercare

Studies have shown that the elderly population is growing world-wide [11]. Roboticists from research institutions, including NICT [103], the University of Missouri [105], and USC [92], among others, are investigating robots for use as minders, guides, and companions for the elderly.

2.2.1 End-User Evaluations

Researchers at NICT conducted a five-week study of twenty-three elderly women in an eldercare facility. The participants interacted with Paro, the therapeutic care robot seal, one to three times per week [103]. Performance measures included self assessment of the participant's mood (pictorial semantic differential scale [71] of *1 = happy* to *20 = sad*) both before and after the interaction with Paro; questions from the Profile of Mood States questionnaire [64] to evaluate anxiety, depression, and vigor (semantic differential scale of *0 = none* to *4 = extremely*); and stress analysis of urinary specimens.

Researchers at the University of Missouri, together with TigerPlace, an independent living facility for the elderly, studied assistive technology for aging in place [105]. At TigerPlace, elderly people who would otherwise be required to have full-time nursing-home care are able to live in their individual residences and have health services brought to them. As part of this effort, the researchers developed a fuzzy logic-based augmentation of an existing day-to-day evaluation, the Short Physical Performance Battery (SPPB) [45]. SPPB measures the user's performance on balance, gait, strength, and endurance tasks. The fuzzy logic augmentation provided finer-grained performance measure for day-to-day monitoring. The team conducted observational studies of two elderly people recovering from surgery in their apartments at TigerPlace [98]. Sensors were placed in their apartments for a period of 14 and 16 months, respectively. Performance measures included a number of categorizations of restlessness (i.e., time vs. event frequency) and quality of life (i.e., ability to complete activities of daily living).

Researchers at the University of Southern California have developed a robot for exercise therapy for adults with dementia and Alzheimer's Disease [92, 93]. The experiment was designed based on the existing music therapy sessions conducted at a Silverado Senior Living community. In the experiment, the participant sat in front of a panel with large, bright, labeled buttons, and a mobile robot with an expressive humanoid torso and head. The robot played music and encouraged and coached the participant to "name that tune" by pushing the correct button. The complexity of the experiment was controlled by the amount of information provided by the robot (from no information, to the name of the song, to hints about the name of the song, to prompts for pushing a button). The performance measures included compliance with the game, enjoyment of the game (evaluated based on the type and amount of vocalizations and facial expressions of the participant), and response time in pushing the buttons, and correctness of responses. The experiment occurred twice per week

for eight months and the challenge level of the exercise was progressively adjusted in order to retain the participant's interest over multiple sessions.

2.2.2 Discussion

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 evaluations and behavioral studies of eldercare systems, such as the studies with Paro, describe the effects that such systems have on users and their environment. By emphasizing social interaction and fitness, these performance measures implicitly gauge the 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 will become increasingly important. Standardized questionnaires for observing QoL over time can be employed to observe any long-term effectiveness of such interventions in the eldercare environment [113]. For example, the SF-36 survey [1] is used to assess health-related QoL, while the 15-D [85] survey is used to measure QoL along several elements of a participant's lifestyle.

2.3 Stroke Rehabilitation

The use of robots is being investigated for gait training at Arizona State University [106], upper-limb recovery at the Rehabilitation Institute of Chicago and Northwestern University [51], and wrist rehabilitation at Hong Kong Polytechnic University [52]. It is well-documented that stroke patients regain most of their mobility through repetitions of task training [53]. The need for technology, such as robots, for supervising and guiding functional rehabilitation exercises is constantly increasing due to the growing elderly population and the large number of stroke victims. Matarić et al. [62] described the two sub-fields: hands-on rehabilitation systems which apply force to guide the affected limb in rehabilitation exercises and hands-off socially assistive systems that provide monitoring and coaching through verbal and gestural feedback without any physical contact. The two methods play complementary roles at different stages of the rehabilitation process.

2.3.1 End-User Evaluations

Pilot experiments are conducted with a small number of study participants in the study design to determine what needs to be altered. The results of a pilot experiment are used to justify a full-scale clinical trial. An example of a pilot experiment is Wada et al.'s case study ($n = 1$) of their Robotic Gait Trainer [106]. Twice per week for eight weeks, the participant walked on a treadmill with the Robotic Gait Trainer

assistance. The supination and pronation position of the participant's foot was measured to determine the quality of her gait. Other performance measure included the six-minute walk test (6MWT) [46] and the timed get-up-and-go test (TGUG) [104].

An example of a small-scale clinical study is Housman et al.'s evaluation of the Therapy Wilmington Robotics Exoskeleton (T-WREX) conducted at the Rehabilitation Institute of Chicago (RIC) and Northwestern University [51]. This clinical trial is an example of typical contact-based rehabilitation robot study with stroke patients. The team conducted a clinical evaluation of twenty-three stroke survivors over sixteen weeks comparing robot-assisted therapy to a traditional rehabilitation therapy regiment [51]. 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 [40] and Rancho Functional Test for Upper Extremity [109] 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 [102], 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 primary assessment of post-stroke rehabilitative robotics involves the use of clinical assessments of patient function. Discussed above were the Fugl-Meyer, Rancho Functional Test, 6MWT, and TGUG assessments. However, there are many others in clinical use today. At at Northwestern University and RIC, Ellis et al. supplemented the Fugl-Meyer with several other measures, including the Chedokee McMaster Stroke Assessment, the Reaching Performance Scale, and the Stroke Impact Scale [28]. At Hong Kong Polytechnic University, Hu et al. used four other measures [52]: the Motor Status Score (MSS, used to assess shoulder function) [35], the Modified Ashworth Scale (MAS, used to measure of increase of muscle tone) [7], the Action Research Arm Test (ARAT, used to assess grasp, grip, pinch, and gross movement) [26], and FIM (used to asses functionality in ADLs) [65]. These performance measures exemplify the clinical definition of effectiveness.

2.3.2 Discussion

Stroke rehabilitation is an established medical domain. The evaluations of assistive robot experiments in this domain must use relevant clinical evaluations to determine the effectiveness of the robot-augmented therapy. The scope of rehabilitative robotics for stroke-recovery patients 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, which is used to assess shoulder function, is applicable to the T-WREX [51] upper-arm rehabilitative aid but not to evaluating gait rehabilitation.

Functional evaluations, such as the Fugl-Meyer [43] and Wolf Motor Function [110], 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.

2.4 Intelligent Wheelchairs

Intelligent wheelchairs have the potential to 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.

2.4.1 End-User Evaluations

In 2005, MobileRobots (formerly ActivMedia) and researchers from the University of Massachusetts Lowell (UML) evaluated the Independence-Enhancing Wheelchair (IEW) [69, 68] with several end-users at a rehabilitation center. The original testing design planned to use a maze-like obstacle course constructed with cardboard boxes. However, this scenario did not work well for the participants. They were frustrated by a maze that was not like their regular driving environments and viewed boxes as movable 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 measures included the number of hits and near misses and time on task.These measures were compared to the same metrics gathered during a similar length observatin of the participant using his/her own wheelchair.

End-user trials have also been completed by intelligent wheelchair companies, such as DEKA [25] and CALL Centre [14] for government approval of the safety of those systems. Researchers at the University of Pittsburgh conducted an evaluation of DEKA's iBOT stair-climbing and self-balancing wheelchair with end-users [22].

2.4.2 Discussion

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 by Yanco [114], 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 participant's seating to be moved onto the prototype system. While seating can be moved from the participant's wheelchair to the prototype system (if compatible)

and back, such seating switches 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 [84]. 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 [100]. 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 [15, 91].

2.5 Assistive Robotic Arms

Robotic arms can improve a person's independence by aiding in activities of daily living (ADLs), such as self-care and pick-and-place tasks. Such arms can be used in fixed workstations, placed on mobile platforms, or mounted to wheelchairs. Ongoing research focuses on both the design of the arms and the human-robot interaction. The pick-and-place task, retrieving an object from a shelf or floor, is of particular interest as it is one of the most common ADLs [90]. Institutions where researchers are investigating assistive robotic arms include Georgia Institute of Technology [20], University of Pittsburgh [19], Clarkson University [41], University of Massachusetts Lowell [97], Delft University [95], and TNO Science & Industry [95].

2.5.1 End-User Evaluations

The Georgia Institute of Technology conducted an evaluation of laser pointers and a touch screen to control a mobile assistive robot arm, El-E [20]. Eight Amyotrophic Lateral Sclerosis (ALS or Lou Gehrig's Disease) end-users directed El-E to pick up objects from the floor in 134 trials. Performance measures included selection time for the participant to point to the object, movement time of the robot to the object, grasping time of the robot to pick up the object, and distance error. A post-experiment questionnaire with eight satisfaction questions yielded seven point Likert scale ratings. The participants' physical conditions were also assessed by a nurse using the Revised ALS Functional Rating Scale (ALSFRS-R) [17].

University of Pittsburgh researchers evaluated the effects of a Raptor arm, a commercially available wheelchair-mounted robotic arm, based on the independence of eleven users with spinal cord injury [19]. 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 researchers evaluated eight users with multiple sclerosis (MS) over five ADLs with and without the Raptor arm [41]. The participants in the study all required assistance with self-care ADLs. They 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) [24].

Researchers at the University of Massachusetts Lowell conducted an experiment of a new visual human-robot interface for the Manus Assistive Robotic Manipulator (ARM) [29]. Eight individuals who used wheelchairs and had cognitive impairments participated in an eight-week 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, and survey responses (i.e., preference of interface, improvements).

TNO Science & Industry and Delft University researchers conducted a four-person case study [95]. The end-users were people who used power wheelchairs and had weak upper limb strength and intact cognition. TNO Science & Industry evaluated their graphical user interface for the Manus ARM. The performance measures included number of mode switches, task time, Rating Scale of Mental Effort (RSME) [115], and survey responses including the participants' opinions about the tasks, the robot arm control methods, and impression of the user interface [95].

2.5.2 Discussion

As demonstrated by Tsui et al. [97], Tijsma et al. [95], and Fulk et al. [41], 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. [97] and Tijsma et al. [95], the participants were rated or rated themselves with respect to cognitive workload. In Tsui et al. [97], the level of prompting was a cognitive measure based on FIM, a measurement of functional independence [65], where in the user is rated on a semantic differential scale (*1 = needs total assistance* to *7 = has complete independence*) on a variety of ADLs. Choi et al. [20] indirectly investigated cognitive workload using an human-computer interaction inspired survey. The participants rated statements such as "It was easy to find an object with the interface" and "It was easy to learn to use the system" on a seven point Likert scale [60] (*-3 = strongly disagree* to *3 = strongly agree*).

FIM may also be applied as a cognitive measure to activities such as "comprehension, expression, social interaction, problem solving, and memory" [65]. In Tijsma et al. [95], 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 [70] is another cognitive performance measures used in older adults.

In Fulk et al. [41], participants 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 [23]. 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 [23]. An alternative emotional measure is the Profile of Mood States [64] used in Wada et al. [103].

2.6 External Limb Prostheses

Robotic prostheses can serve as limb replacements. Researchers have investigated creating novel robotic prostheses and control strategies. A number of prosthesis evaluations conducted have been feasibility studies on healthy subjects. As such, the focus of the experiments has largely been on the performance of the prostheses themselves. The performance measures include joint angle, joint torque, and power consumption. However, several research institutions have conducted end-user evaluations, including RIC [67, 57], Northwestern University [67, 57], Massachusetts Institute of Technology [4, 5], and Hong Kong Polytechnic University [58].

2.6.1 End-User Evaluations

RIC and Northwestern University conducted a clinical evaluation of six individuals who underwent targeted muscle reinnervation (TMR) surgery [67]. 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) [36], and the University of New Brunswick prosthetic function [82]. Performance measures included time to complete task, accuracy, and AMPS score.

Another RIC and Northwestern University study evaluated the effectiveness of the TMR procedure when controlling robotic prostheses with EMG signals [57]. Five participants with shoulder-disarticulation or transhumeral amputation who had the TMR procedure and five able-bodied participants controlled a virtual prosthetic arm to grip in three predetermined grasps. The performance measures included motion selection time (time from when motion began to correct classification), motion completion time, and motion completion rate.

Additionally, three of the participants who had undergone TMR also used physical robotic upper-limb prostheses (i.e., DEKA's ten degree-of-freedom "Luke arm" and a motorized seven degree-of-freedom prosthetic arm developed at John Hopkins University) using EMG signals [57]. The training and testing ran for two weeks with

one session in the morning and another in the afternoon; session lasted two to three hours. The participants were able to operate the prostheses in ADL-type tasks and controlling grasps. These results are largely anecdotal.

Researchers at the Massachusetts Institute of Technology (MIT) conducted a clinical evaluation with three unilateral, transtibial amputees [4]. 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, 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 researchers conducted a clinical evaluation with four transtibial amputees over the course of three consecutive days [58]. 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.

2.6.2 Discussion

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. [48] proposed the Orthotics and Prosthetics Users' Survey (OPUS). Burger et al. [12] 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 semantic differential scale fashion (*0 = unable to complete* to *3 = very easy to complete*. 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].

The clinical evaluations conducted with transtibial amputees discussed above used performance measures of the robotic system itself (i.e., gait symmetry and ground force reactions). Additionally, Hong Kong Polytechnic University administered a questionnaire asking about the participant's perception of the lower limb prosthesis, and, in an indirect manner, MIT measured the ease of use of the prosthesis by a biological means. In order for a prosthesis to gain clinical validity, performance of the device must also have a measure of use in daily life.

3 Case Studies

Next, we further explore two examples detailing the evolution of performance metrics on two different ongoing studies involving assistive robotics. At the University of Massachusetts Lowell (UML), we have conducted one able-bodied experiment and three end-user experiments with people who use wheelchairs involving an assistive robotic arm in the "pick-and-place" activity of daily living. At the University of Southern California (USC), we have conducted three preparatory experiments with end-users and are in the process of conducting an end-user experiment using a socially assistive robot designed to provoke or encourage exercise or social behavior in children with Autism Spectrum Disorders (ASD).

3.1 Designing Evaluations for an Assistive Robotic Arm

At UML, our research focuses on providing methods for independent manipulation of unstructured environments to wheelchair users using a wheelchair-mounted robot arm. Our target audience consists of people with physical disabilities who may additionally have cognitive impairments. We have investigated a visual interface compatible with single switch scanning [96], a touch screen interface [97], a mouse-emulating joystick [97], and a laser pointer joystick device [75]. By explicitly pointing to the desired object, it may be possible to expand the end-user population to include people with low cognition. We conducted a preliminary experiment with able-bodied participants as an evaluation baseline in August 2006 [96]. The first field trial was conducted with users who use wheelchairs and additionally had cognitive impairments in August and September 2007 [97]. The second field trial was conducted in August and September 2008. Our third field trial will begin in mid-July 2009 and run through the end of October 2009. In this section, we discuss our design of the end-user experiments.

In our first end-user evaluation, we compared the visual interface presentation (stationary camera vs. moving camera) and input device (touch screen vs. joystick) [97]. We collected data from video, manual logs, post-session questionnaires, and computer generated log files. We collected both qualitative and quantitative data. The qualitative data included the post-experiment questionnaire administered after each user session and the observer notes. The questionnaire posed open-ended questions about which interface the user liked most to date, which interface he/she liked least to date, and suggestions for improving the interface. The observer notes contained additional relevant data about the session, including length of reorientation.

The quantitative data included an attentiveness rating, prompting level, trial run time, close-up photos of the object selected, and computer-generated log files. The attentiveness rating and prompting level were developed by our clinicians. The experimenter, who was an assistive technology professional, rated the user's prompting level per trial based on the FIM scale, where *0 = no prompting needed* and *5 = heavy prompting needed* [65]. The experimenter also rated the user's attentiveness

to the task on a semantic differential scale, where *0 = no attention* and *10 = complete attention*. Two separate scales were used because it is not necessarily the case that a person who requires high levels of prompting is unmotivated to complete the task. Also, for each trial, the run time was recorded, specifically the time from object prompt to participant selection, the time from the Manus ARM movement to the object being visually confirmed, and the fold time. We focused on the object selection time, prompting level, and attention level as our primary performance metrics and computed paired t-tests to determine statistical significance.

In our second end-user evaluation, we compared a custom laser pointer device against the touch screen interface with stationary view[2] [75]. We were very satisfied with quality of the primary performance metrics from the first end-user evaluation. Thus, we based the next version of our data collection tools from the previous experiment and made several modifications. We recorded only the object selection time by the participant since the remainder of the process time is a robot system performance measurement. The post-session surveys were also based on the ones from the first end-user evaluation. Because the participants used two interfaces in the second end-user evaluation, we modified the post-session survey to investigate which aspects of each interface the participants liked and did not like, comments about the laser joystick and touch screen, and which interface they liked better.

At the suggestion of our clinicians, we updated the semantic differential scales to have the same range (i.e., [1, 5], where *1 = no prompting* needed and *5 = heavy prompting* for the prompting level, and *1 = not attentive* and *5 = very attentive* for attention level) which provided the same granularity across the performance measurements. We introduced a tally box for the number of prompts given by the experimenter which provided the ability to better understand what an experimenter considered a "high" level of prompting versus "low." In our observations of the first end-user evaluation, we noticed that the participants would seem quiet on one day and excited on another. To better understand how the performance of our robotic arm system was perceived by the participants, we added a mood and arousal level rating (i.e., *1 = very bad* and *5 = very good* for mood, and *1 = less than normal* and *5 = more than normal*) to be administered at the start of the session, before the condition change, and after the session. We added a health rating (i.e., *1 = poor* and *5 = good*) to be administered at the start of the session. These mood, arousal, and health ratings were items previously noted by the experimenter in the first end-user evaluation.

Our upcoming third end-user evaluation will investigate how different levels of cognition (i.e., *high*, *medium*, and *low* as classified by our clinician) impact a person's ability to use the robotic arm. We will continue to use the selection time and prompting, attention, mood, arousal, and health levels. We found that the second end-user experiment's ratings scales were not as effective as those in the first end-user evaluation. The second end-user experiment's rating scales were relative to each participant's typical performance, and we did not see much change. For the upcoming evaluation, we will instantiate the semantic differential scales in a concrete manner (i.e., for arousal, *1 = low* and *5 = high*). We will also incorporate

[2] The touch screen interface with the stationary camera view had the best overall performance from the first end-user evaluation [97].

aspects of the Psychosocial Impact of Assistive Devices Scale (PIADS), in which the participants will rate their perceived experiences with the Manus ARM [24].

3.2 Designing Evaluations for Socially Assistive Robots

At USC, our research focuses on the use of socially assistive robots that provide assistance through social interaction rather than physical interaction [31]. We are developing robot systems for encouraging and training social behavior for children with ASD. Our work focuses on the following goals: the automatic identification of social behavior; the creation of a toolkit of interactive robot behavior that can be used in order to provoke and encourage social interaction; and the determination of therapeutic effectiveness of socially assistive robots.

The robot we have developed is a humanoid torso (with movable arms, neck, head, and face) on a mobile base. The robot has microphones and speakers, so that it can make and recognize vocalizations, buttons that the user can press, and a bubble blower (typically used as part of standardized ASD evaluation and intervention [61]). The robot "rewards" a user's social behavior with its own social behavior, including gestures, vocalizations, and movements (approach, spinning in place, facing the user). Additionally, we use a camera and microphones in the experimental room at the clinic to collect high-fidelity multi-modal data of all aspects of the study.

We are developing a system that observes a child with ASD and automatically identifies social behaviors (e.g., approach, turn-taking, social referencing, appropriate affect, and/or vocalizations). Our long-range goal is to use the frequency and context of those behaviors in order for the robot to determine autonomously if the child is socially engaging with the robot or another person. Our studies use an overhead camera system to track and interpret the child's movement as he/she interacts with the robot [32]. The goal is for the overhead camera to autonomously identify the movement behavior of the robot and the child (such as A approaches B, B follows A, etc.). We conducted a preliminary experiment in which we collected supervised overhead camera data. We created a data set in which the person and the robot executed known actions.The system then performed an automated analysis of the data which was accompanied by blind human coding. Because the actions were known *a priori*, we were able to determine an absolute measure of the automatic coding mechanism's performance. We compared the accuracy of the automatic observations to human coding of the same actions which provided a relative measure of the systems performance. Our larger experiments employ a similar experimental data collection, coding, and evaluation model.

After validating that the socially assistive robot system is both effective at interacting with the child with ASD (i.e., it successfully elicits social behaviors) [34] and that our analysis of its effectiveness is valid (i.e., the coding and analysis algorithms) [32], the next step in validation must address any possible therapeutic benefits of the human-robot interaction. Our goal is not to presume or aim for a clinical benefit, but to validate reliably that such a socially assistive robot could have a potential thera-

peutic impact in order to plan follow-up studies. We are working with clinicians to determine if there are any therapeutic applications for such a robot system.

We are currently planing a validation experiment with end-users. The experiment will involve multiple sessions in which a child with ASD will participate in free-play with a trained therapist. Participants will be split into control and experimental groups. The control group will interact only with the therapist, while the experimental group will interact with the therapist and the robot for five sessions. The participants will be given a pre- and post-experiment evaluation consisting of the WISC-III Intelligence Test [107], the communication portion of the Vineland Adaptive Behavior Scales (VABS) [72], and an ADOS-inspired quantitative observation. These measures are used for repeated administration and comparison. Hypothesis testing will then involve comparing any change from the pre- and post-experiment evaluations between the control and experimental groups. Paired t-tests will be used to compare each test and any applicable sub-test. The variety of scales to be used in this upcoming study provides a rich range of measures pertaining to the social ability of the end-user.

4 Incorporating Functional Performance Measures

Evaluation of assistive robotic technology varies widely, as has been demonstrated by our exploration of several domains. It is clear that in order for assistive robotic technology to be accepted by clinicians at large, end-user evaluations must incorporate a functional performance measure based on the "gold standard" of the specific domain, if one has been established. We say that a performance measure is *functional* if it relates to an activity of daily living and is administered in a realistic setting. In this survey, we have found examples of functional performances used in the majority of the surveyed domains.[3]

Feil-Seifer et al. consulted the Autism Diagnostic Observation Schedule (ADOS) in their evaluation [33]. ADOS is one of the ASD "gold standard" assessment tools; it investigates "social interaction, communication, play, and imaginative use of materials" [61]. The Vineland Adaptive Behavior Scales (VABS) are a grouping of assessment tools for ASD and developmental delays [72, 16]. Unlike ADOS, VABS contains functional components, such as Daily Living Skill items and Motor Skill items. VABS has also been used in the domains of stroke (children) [63], wheelchairs (children) [27], prostheses (children) [73], and eldercare (developmentally disabled) [54].

[3] A functional performance measure was not surveyed for the domain of Autism Spectrum Disorders (ASD).

Table 1 Examples of Functional Performance Measurement Tools

Tool	Length	Categories of Functional Assessment	Rating	Applicable Domain(s)
15D [86]	15 items	mobility, vision, hearing, breathing, sleeping, eating, speech, elimination, activities of daily life, mental function, discomfort and symptoms, depression, distress, vitality, sexual activity	$n \in [1, 5]$, where 1 = cannot do, 5 = normal ability	Eldercare, post-stroke rehabilitation
FIM [9]	18 items	feeding, grooming, bathing, dressing (upper and lower body), toileting, bladder management, bowel management, bed transfer, toilet transfer, tub transfer, walking/wheelchair, climbing stairs, comprehension, expression, social interaction, problem solving, memory	$n \in [1, 7]$, where 1 = needs total assistance and 7 = complete independence	Autism Spectrum Disorders, assistive robotic arms, eldercare, post-stroke rehabilitation, wheelchair
Motor Assessment Scale [6]	8 items	supine to side lying, supine to sitting over the edge of a bed, balanced sitting, sitting to standing, walking, upper-arm function, hand movements, advanced hand activities	$n \in [0, 6]$, where 6 = optimal motor behavior	Post-stroke rehabilitation
SF-36 [74]	36 items	typical daily activities (strenuous activities, moderate activities, carrying groceries, climbing several flights of stairs, climbing one flight of stairs, bending/kneeling, walking more than a mile, walking several blocks, walking one block, bathing/dressing yourself)	$n \in [0, 3]$, where 0 = limited a lot, 3 = not limited	Eldercare, post-stroke rehabilitation, prostheses, wheelchair
Vineland Adaptive Behavior Scales (survey form) [72]	297 items	daily living skills, motor skills (gross motor, fine motor)	$n \in [0,2]$, where 0 = never performed without help or reminders and 2 = usually or habitually performed without help or reminders	Autism Spectrum Disorders, eldercare, post-stroke rehabilitation, prostheses, wheelchairs

We discussed the quality of life (QoL) measurements in the context of the eldercare domain. QoL measurement scales contain functional components to them, such as walking, climbing stairs, and self-care activities. The SF-36 [74] is a generic health measure which includes evaluations of physical function and limitations due to physical health. The SF-36 contains 36 items in total, including one multi-part question focused specifically on typical daily activities; see Table 1. The SF-36 has been used for assessment of eldercare [42], wheelchairs [13], stroke [3, 50], and a prosthesis [59]. The 15-D measures QoL as a profile with 15 dimensions; see Table 1. Mobility and eating are included as profile dimensions [86]. The 15-D has been used in the eldercare, orthopedic, and stroke rehabilitation domains [86].

The FIM scale [9] by definition is a functional performance measure; see Table 1. The FIM ADLs include "eating, grooming, bathing, dressing (upper body), dressing (lower body), toileting, bladder management, bowel management, transferring (to go from one place to another) in a bed, chair, and/or wheelchair, transferring on and off a toilet, transferring into and out of a shower, locomotion (moving) for walking or in a wheelchair, and locomotion going up and down stairs" [65]. FIM has been used largely in stroke rehabilitation [38] and to some extent in eldercare [55]. WeeFIM [99] is used for children between the ages of six months and seven years (present with functional abilities of or below age seven).WeeFIM has recently been used in clinical Autism studies in Hong Kong [111, 112]. FIM has been adapted for wheelchair users [87]. As described in Section 3.1, FIM inspired a scale for recording a user's prompting level while doing a task [97].

Currently, there is a large gap between robotic performance measures and functional performance measures. Robotic performance measures typically consider metrics such as time on task and number of collisions, while functional performance measures mentioned above do not employ such a fine level of granularity. The functional performance measures examine tasks and categorize a person's ability to complete those tasks in a n-nary manner (e.g., Motor Assessment Scale is ternary with "was able to complete easily," "was able to complete with some difficulty," "was not able to complete" [6]); see Table 1. To create finer granularity in functional performance measures, intermediate layers can be added. For example, Stineman et al. added intermediate layers to FIM in order to understand the causal relationship between impairments and disabilities [89].

Functional performance measures take a significant amount of time to administer, ranging from 20 to 60 minutes or more. The "gold standard" functional measure for a given domain thus should be administered once before the experiment and once after. However, a *subset* of the functional performance measures relevant to the specific area investigated can be used during each testing session, as has been done in some of the studies references above. Furthermore, robotic performance measures derived from this subset can provide continuous monitoring. These functional robotic performance measures may then help to bridge the gap between the strictly robotic performance measures and the functional performance measures commonly employed in clinical domains.

Table 2 Summary of Assistive Robotic Technology Performance Measures

Domain	Applicable Performance Measures
General AT	Activities of daily living, coding, instantiated Likert-type ratings, mood, quality of life, stress, time on task
Autism Spectrum Disorders	Behavior coding, correlate sensor modeling of behavior to human-rated behavior, standardized assessments (e.g., ADOS, Vineland Adaptive Behavior Scales)
Eldercare	Activities of daily living (e.g., FIM, SBBP), mood (e.g., Profile of Mood States), quality of life (e.g., 15-D, SF-36), response correctness, response time, stress (e.g., Standardized Mini-Mental State)
Post-Stroke Rehabilitation	Functional performance measures (e.g., FIM, Motor Activity Log, Motor Assessment Scale), quality of life (e.g., 15-D, SF-36), standardized assessments (e.g., ARAT, Chedoke-McMaster, Fugl-Meyer, Modified Ashworth Scale, MSS, Reaching Performance Scale, Wolf Motor)
Intelligent Wheelchairs	Accuracy, functional performance measures (e.g., FIM), gracefulness, number of hits/near misses, quality of life (e.g., SF-36), time on task
Assistive Robotic Arms	Activities of daily living (e.g., ALSFRS-R, FIM), attention, level of prompting, mental state (e.g., RSME, Profile of Mood States, PIADS), mood, quality of life, time to task completion
Prostheses	Accuracy, biological measures of effort (e.g., oxygen consumption), comfort, ease of use, functional performance measures (e.g., AMPS, OPUS, FIM), quality of life (e.g., SF-36), time to complete task

5 Conclusions

To be useful, performance measures should be specific to the domain and relevant to the task. Domains with clear, well-established medical or therapeutic analogs should leverage existing clinical performance measures. Domains without such strong therapeutic analogs can appropriately borrow and adapt clinical performance measures. Alternatively, they may draw inspiration from a clinical measure to create a new one or augment an existing one if none of the existing measures are appropriate [45].

Evaluations conducted with end-users should focus at least as much 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. examined trajectory error and smooth-

ness of motion with respect to Fugl-Meyer in the context of post-stroke rehabilitation [18]. Similarly, Brewer et al. have 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) [10, 30].

Existing performance measures for most of assistive robotic technologies do not provide sufficient detail for experimental and clinical evaluations. We have provided a summary of applicable performance measures (see Table 2) and offer the following guidelines for choosing appropriate and meaningful performance measures:

- Consult a clinician who specializes in the particular domain.
- Choose an appropriate clinical measure for the domain. A domain's "gold standard" will provide the best validity to clinicians, if one exists.
- Include a functional performance measure appropriate for the domain.
- 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 once before and after the experiment or study.
- Consider coding open-ended responses, comments, and/or video.
- Concretely define each enumeration on Likert and differential semantic scales.

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

In this paper, we have sought well-established performance measures to apply to assistive robotic technologies and encourage the practice of their use in our field. Common performance measurements will allow researchers to compare the state of the art approaches within specific robotics domains and to compare against the state of the practice within the relevant clinical field outside of the robotics community.

Acknowledgements This work is funded in part by the National Science Foundation (IIS-0534364, IIS-0546309, IIS-0713697, CNS-0709296), the National Academies Keck Futures Initiative (NAKFI), the USC NIH Aging and Disability Resource Center (ADRC) pilot program, and the Nancy Laurie Marks Family Foundation. The authors thank Kristen Stubbs of UMass Lowell.

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. AMPS Project International (Assessment of Motor and Process Skills), 2006. http://www.ampsintl.com/tasks.htm. Accessed Mar. 1, 2009.
3. C. Anderson, S. Laubscher, and R. Burns. Validation of the Short Form 36 (SF-36) Health Survey Questionnaire Among Stroke Patients. *Stroke*, 27(10):1812–1816, 1996.

4. S. Au. *Powered Ankle-Foot Prosthesis for the Improvement of Amputee Walking Economy*. PhD thesis, MIT, 2007.
5. S. Au, M. Berniker, and H. Herr. 2008 Special Issue: Powered Ankle-Foot Prosthesis to Assist Level-Ground and Stair-Descent Gaits. *Neural Networks*, 21(4):654–666, 2008.
6. L. Blum, N. Korner-Bitensky, and E. Sitcoff. StrokEngine (MAS), 2009. http://www.medicine.mcgill.ca/strokengine-assess/module_mas_indepth-en.html. Accessed Mar. 1, 2009.
7. R. Bohannon and M. Smith. Interrater Reliability of a Modified Ashworth Scale of Muscle Spasticity. *Physical Therapy*, 67(2):206–7, 1987.
8. R. Bourgeois-Doyle. *George J. Klein: The Great Inventor*. NRC Research Press, 2004.
9. Brain Injury Resource Foundation. Functional Independence Measure (FIM), 2009. http://www.birf.info/home/bi-tools/tests/fam.html. Accessed Mar. 1, 2009.
10. 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.
11. J. Brody. Prospects for an Ageing Population. *Nature*, 315(6019):463–466, 1985.
12. 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.
13. T. Bursick, E. Trefler, D. A. Hobson, and S. Fitzgerald. Functional Outcomes of Wheelchair Seating and Positioning in the Elderly Nursing Home Population. In *Rehab. Eng. Society of North America Conf.*, pages 316–318, 2000.
14. CALL Centre. Smart Wheelchair, 2006. http://callcentre.education.ed.ac.uk/Smart_WheelCh/smart_wheelch.html. Accessed Mar. 1, 2009.
15. T. Carlson and Y. Demiris. Human-Wheelchair Collaboration Through Prediction of Intention and Adaptive Assistance. In *IEEE Intl. Conf. on Robotics and Automation*, 2008.
16. A. Carter, F. Volkmar, S. Sparrow, J. Wang, C. Lord, G. Dawson, E. Fombonne, K. Loveland, G. Mesibov, and E. Schopler. The Vineland Adaptive Behavior Scales: Supplementary Norms for Individuals with Autism. *J. of Autism and Developmental Disorders*, 28(4):287–302, 1998.
17. J. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, and A. Nakanishi. The ALSFRS-R: A Revised ALS Functional Rating Scale that Incorporates Assessments of Respiratory Function. BDNF ALS Study Group (Phase III). *J. Neurol. Sci.*, 169(1-2):13–21, 1999.
18. 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.
19. 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.
20. Y. Choi, C. Anderson, J. Glass, and C. Kemp. Laser Pointers and a Touch Screen: Intuitive Interfaces for Autonomous Mobile Manipulation for the Motor Impaired. In *Intl. ACM SIGACCESS Conf. on Computers and Accessibility*, pages 225–232, 2008.
21. J. A. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46, 1960.
22. 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.
23. H. Day and J. Jutai. PIADS in the World, 2009. http://www.piads.ca/worldmapshtm/worldmap.asp. Accessed Mar. 1, 2009.
24. 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.
25. DEKA Research and Development Corporation. DEKA Evolved Thinking, 2007. http://www.dekaresearch.com. Accessed Mar. 1, 2009.

26. 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.
27. M. Donkervoort, M. Roebroeck, D. Wiegerink, H. van der Heijden-Maessen, and H. Stam. Determinants of Functioning of Adolescents and Young Adults with Cerebral Palsy. *Disabil. Rehab.*, 29:453–463, 2007.
28. 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.
29. Exact Dynamics. Assistive Robotic Manipulator, 2004. http://www.exactdynamics.nl/. Accessed Mar. 1, 2009.
30. S. Fahn, R. Elton, et al. Unified Parkinson's Disease Rating Scale. *Recent Developments in Parkinson's Disease*, 2:153–163, 1987.
31. D. Feil-Seifer and M. Matarić. Socially Assistive Robotics. In *Intl. Conf. on Rehab. Robotics*, pages 465–468, 2005.
32. D. J. Feil-Seifer and M. J. Matarić. B3IA: An Architecture for Autonomous Robot-Assisted Behavior Intervention for Children with Autism Spectrum Disorders. In *Intl. Workshop on Robot and Human Interactive Communication*, Munich, Germany, Aug 2008.
33. 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.
34. 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*, volume 54, pages 201–210, 2008.
35. 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. *Neurorehab. and Neural Repair*, 16(3):283, 2002.
36. A. Fisher. AMPS: Assessment of Motor and Process Skills Volume 1: Development, Standardisation, and Administration Manual. *Ft Collins, CO: Three Star Press Inc*, 2003.
37. J. Fleiss. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5):378–382, 1971.
38. W. Foczkowski and S. Barreca. The Functional Independence Measure: Its Use to Identify Rehabilitation Needs in Stroke Survivors. *Archives of Physical Medicine and Rehab.*, 74(12):1291–1294, 1993.
39. B. Freeman. Guidelines for Evaluating Intervention Programs for Children with Autism. *J. of Autism and Developmental Disorders*, 27(6):641–651, 1997.
40. 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.
41. G. Fulk, M. Frick, A. Behal, and M. Ludwig. A Wheelchair Mounted Robotic Arm for Individuals with Multiple Sclerosis. Technical report, Clarkson Univ., 2005.
42. B. Gale. Faculty Practice as Partnership with a Community Coalition. *J. Prof. Nurs.*, 14(5):267–271, 1998.
43. D. Gladstone, C. Danells, and S. Black. The Fugl-Meyer Assessment of Motor Recovery after Stroke: A Critical Review of Its Measurement Properties. *Neurorehab. and Neural Repair*, 16(3):232, 2002.
44. D. Greenwood, W. Whyte, and I. Harkavy. Participatory Action Research as a Process and as a Goal. *Human Relations*, 46(2):175–192, 1993.
45. 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.
46. G. Guyatt. The 6-minute walk: a new measure of exercise capacity in patients with chronic heart failure. *Canadian Medical Association Journal*, 132(8):919–923, 1985.

47. 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.
48. 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.
49. M. Hillman. Rehabilitation Robotics from Past to Present—A Historical Perspective. In *IEEE Intl. Conf. on Rehab. Robotics*, 2003.
50. J. Hobart, L. Williams, K. Moran, and A. Thompson. Quality of Life Measurement After Stroke Uses and Abuses of the SF-36. *Stroke*, 33(5):1348–1356, 2002.
51. 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.
52. 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.
53. 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.
54. D. Kerby, R. Wentworth, and P. Cotten. Measuring Adaptive Behavior in Elderly Developmentally Disabled Clients. *J. of Applied Gerontology*, 8(2):261, 1989.
55. R. Kleinpell, K. Fletcher, and B. Jennings. Reducing Functional Decline in Hospitalized Elderly. *Patient Safety and Quality: An Evidence-Based Handbook for Nurses. AHRQ Publication No. 08-0043*, 2007.
56. 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.
57. T. Kuiken, G. Li, B. Lock, R. Lipschutz, L. Miller, K. Stubblefield, and K. Englehart. Targeted Muscle Reinnervation for Real-time Myoelectric Control of Multifunction Artificial Arms. *J. of American Medical Assoc.*, 301(6):619–628, 2009.
58. 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.
59. M. Legro, G. Reiber, M. Del Aguila, M. Ajax, D. Boone, J. Larsen, D. Smith, and B. Sangeorzan. Issues of Importance Reported by Persons with Lower Limb Amputations and Prostheses. *J. of Rehab. Research and Development*, 36(3):155–163, 1999.
60. R. Likert. A Technique for the Measurement of Attitudes. *Archives of Psych.*, 140(5):1–55, 1932.
61. C. Lord, S. Risi, L. Lambrecht, E. H. Cook 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.
62. M. Matarić, J. Eriksson, D. Feil-Seifer, and C. Winstein. Socially Assistive Robotics for Post-Stroke Rehabilitation. *J. of NeuroEngineering and Rehab.*, 4(1):5, 2007.
63. J. Max, K. Mathews, A. Lansing, B. Robertson, P. Fox, J. Lancaster, F. Manes, and J. Smith. Psychiatric Disorders After Childhood Stroke. *J. of the American Academy of Child & Adolescent Psychiatry*, 41(5):555, 2002.
64. D. M. McNair, M. Lorr, and L. F. Droppleman. Profile of Mood States. In *Educational and Industrial Testing Service*, 1992.
65. MedFriendly.com. MedFriendly.com: Functional Independence Measure, 2007. http://www.medfriendly.com/functionalindependencemeasure.html. Accessed Mar. 1, 2009.
66. 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.

67. 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.
68. MobileRobots Inc. Robotic Chariot, 2006. http://activrobots.com/robots/robochariot.html. Accessed Mar. 1, 2009.
69. MobileRobots Inc. Independence-Enhancing Wheelchair, 2008. http://www.activrobots.com/RESEARCH/wheelchair.html. Accessed Mar. 1, 2009.
70. D. Molloy and T. Standish. A Guide to the Standardized Mini-Mental State Examination. *Intl. Psychogeriatrics*, 9(S1):87–94, 2005.
71. C. Osgood, G. J. Suci, and P. H. Tannenbaum. *The Measurement of Meaning*. Univ. of Illinois Press, 1957.
72. Pearson Education, Inc. Vineland Adaptive Behavior Scales, Second Edition (Vineland-II), 2009. http://www.pearsonassessments.com/vinelandadapt.aspx. Accessed Mar. 1, 2009.
73. S. Pruitt, J. Varni, and Y. Setoguchi. Functional Status in Children with Limb Deficiency: Development and Initial Validation of an Outcome Measure. *Archives of Physical Medicine and Rehab.*, 77(12):1233–1238, 1996.
74. QualityMetric Inc. The SF Community—Offering Information and Discussion on Health Outcomes, 2009. http://www.sf-36.org/. Accessed Mar. 1, 2009.
75. E. Rapacki, K. Tsui, D. Kontak, and H. Yanco. Design and Evaluation of a Laser Joystick in a Turret Assembly. In *Rehab. Eng. Society of North America Conf.*, 2008.
76. 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.
77. B. Robins, K. Dautenhahn, R. te Boekhorst, and A. Billard. Robots as Assistive Technology—Does Appearance Matter? In *IEEE Intl. Workshop on Robot and Human Interactive Communication*, 2004.
78. 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.
79. R. Robins, C. Fraley, and R. Krueger. *Handbook of Research Methods in Personality Psychology*. Guilford Press, 2007.
80. 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.
81. T. Salter, F. Michaud, D. Létourneau, D. Lee, and I. Werry. Using Proprioceptive Sensors for Categorizing Interactions. In *Human-Robot Interaction*, 2007.
82. 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.
83. R. Simpson. Smart Wheelchairs: A Literature Review. *J. of Rehab. Research Development*, 42(4):423–36, 2005.
84. R. C. Simpson. *Improved Automatic Adaption Through the Combination of Multiple Information Sources*. PhD thesis, Univ. of Michigan, Ann Arbor, 1997.
85. 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.
86. H. Sintonen. 15D Instruments, 2009. http://www.15d-instrument.net. Accessed Mar. 1, 2009.
87. R. Stanley, D. Stafford, E. Rasch, and M. Rodgers. Development of a Functional Assessment Measure for Manual Wheelchair Users. *J. of Rehab. Research and Development*, 40(4):301–307, 2003.
88. 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.
89. M. Stineman, A. Jette, R. Fiedler, and C. Granger. Impairment-Specific Dimensions Within the Functional Independence Measure. *Archives of Physical Medicine and Rehab.*, 78(6):636–643, 1997.

90. 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.
91. 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.
92. 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.
93. A. Tapus, C. Tapus, and M. J. Matarić. The Use of Socially Assistive Robots in the Design of Intelligent Cognitive Therapies for People with Dementia. In *IEEE Int. Conf. on Rehabilitation Robotics*, 2009.
94. 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.
95. 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.
96. K. Tsui and H. Yanco. Simplifying Wheelchair Mounted Robotic Arm Control with a Visual Interface. In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, pages 247–251, March 2007.
97. 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.
98. H. Tyrer, M. Aud, G. Alexander, M. Skubic, and M. Rantz. Early Detection of Health Changes In Older Adults. *Intl. Conf. of the IEEE Eng. in Medicine and Biology Society*, pages 4045–4048, 2007.
99. Uniform Data System for Medical Rehabilitation. UDSMR::WeeFIM II® System, 2009. http://www.weefim.org/WebModules/WeeFIM/Wee_About.aspx. Accessed Mar. 1, 2009.
100. Univ. of Illinois Chicago. Power Mobility Skills Checklist, 2006. http://internet.dscc.uic.edu/forms/0534.pdf. Accessed Mar. 1, 2009.
101. US Food and Drug Administration. Guidance for Industry, E6 Good Clinical Practice: Consolidated Guidance. *Federal Register*, 10:691–709, 1997.
102. 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.
103. 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.
104. J. Wall, C. Bell, S. Campbell, and J. Davis. The Timed Get-up-and-Go test revisited: measurement of the component tasks. *Journal of rehabilitation research and development*, 37(1):109, 2000.
105. 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.
106. 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.
107. D. Wechsler and Psychological Corporation and Australian Council for Educational Research. Wechsler Intelligence Scale for Children. 1949.
108. WheelchairNet. WheelchairNet: The history of wheelchairs, 2006. http://www. wheelchairnet.org/WCN_ProdServ/Docs/WCHistory.html. Accessed Mar. 1, 2009.
109. D. Wilson, L. Baker, and J. Craddock. Functional Test for the Hemiparetic Upper Extremity. *American J. of Occupational Therapy*, 38(3):159–64, 1984.
110. 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. *Neurorehab. and Neural Repair*, 19(3):194, 2005.
111. V. Wong. Use of Acupuncture in Children with Autism Spectrum Disorder - ClinicalTrials.gov, 2006. http://clinicaltrials.gov/ct2/show/locn/NCT00352352?term=Autism. Accessed Mar. 1, 2009.

112. V. Wong, Y. T. Au-Yeung, and P. Law. Correlation of Functional Independence Measure for Children (WeeFIM) with Developmental Language Tests in Children with Developmental Delay. *J. Child Neurol.*, 20(7):613–616, 2005.
113. 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.
114. H. Yanco. Evaluating the Performance of Assistive Robotic Systems. *Performance Metrics for Intelligent Systems (PerMIS) Workshop*, pages 21–25, 2002.
115. F. Zijlstra. *Efficiency in Work Behaviour: A Design Approach for Modern Tools*. PhD thesis, Delft Univ., 1993.