HUMAN-ROBOT INTERACTION

A metric for characterizing the arm nonuse workspace in poststroke individuals using a robot arm

Nathaniel Dennler[1]*, Amelia Cain[2], Erica De Guzman[1], Claudia Chiu[1], Carolee J. Winstein[2], Stefanos Nikolaidis[1], Maja J. Matarić[1]

An overreliance on the less-affected limb for functional tasks at the expense of the paretic limb and in spite of recovered capacity is an often-observed phenomenon in survivors of hemispheric stroke. The difference between capacity for use and actual spontaneous use is referred to as arm nonuse. Obtaining an ecologically valid evaluation of arm nonuse is challenging because it requires the observation of spontaneous arm choice for different tasks, which can easily be influenced by instructions, presumed expectations, and awareness that one is being tested. To better quantify arm nonuse, we developed the bimanual arm reaching test with a robot (BARTR) for quantitatively assessing arm nonuse in chronic stroke survivors. The BARTR is an instrument that uses a robot arm as a means of remote and unbiased data collection of nuanced spatial data for clinical evaluations of arm nonuse. This approach shows promise for determining the efficacy of interventions designed to reduce paretic arm nonuse and enhance functional recovery after stroke. We show that the BARTR satisfies the criteria of an appropriate metric for neurorehabilitative contexts: It is valid, reliable, and simple to use.

INTRODUCTION

Stroke is a leading cause of serious long-term disability in the United States (*1*). Without sufficient rehabilitation efforts, functional decline will ensue, leading to increased difficulty in completing activities of daily living (ADLs), which contributes to decreased quality of life (*2*, *3*). The goal of poststroke neurorehabilitation is to restore functionality to the affected limb and enable stroke survivors to improve their quality of life. Several poststroke rehabilitative interventions, such as task-oriented training (*4*), biofeedback (*5*), and constraint-induced movement therapy (*6*), have demonstrated substantial improvements along levels of the International Classification of Functioning, Disability, and Health (*7*), including domains of body structure/function, activity limitations, and participation.

Despite these functional improvements, a subset of stroke survivors may still experience a discrepancy between what they are able to do in tests where they are constrained to using their stroke-affected arm and what they spontaneously do in real-world ADLs. This is of particular concern for individuals with hemiparetic stroke and other unilateral motor disorders, because the less-affected side can be used to compensate for movements of the impaired side, and such compensation interferes with the "use it or lose it" foundational principle of neurorehabilitation. The nonuse phenomenon, the discrepancy between capacity and actual use (*8*), was first characterized in an article titled "Stroke recovery: He can but does he?" (*9*). Nonuse has been shown to have a learned component (*10*) and can thus be reduced through practice. This makes nonuse a key behavioral phenomenon to assess when evaluating patient recovery, one with high clinical and scientific importance.

In neurological rehabilitation contexts, outcome metrics must meet three criteria to be considered useful for evaluation: validity, reliability, and ease of use (*11*). However, the two currently widely accepted instruments that provide metrics for nonuse, the motor activity log (MAL) (*12*) and the actual amount of use test (AAUT) (*13*), do not satisfy all three of those criteria. Although both tests have been found to be valid (*14*), they lack the other two desired qualities of neurorehabilitation assessment metrics: reliability and ease of use. The MAL relies on a structured interview for user-reported arm use over the course of a specified duration, for example, 1 week or 3 days. Because of the difficulty associated with remembering and accurately describing one's arm use over the period of a week, this test is not simple for the participants to complete. The AAUT is a covert assessment that is valid only if the participant is unaware that the test is being conducted. Once the test is revealed, it becomes invalid for repeated use, making the scale unreliable. Inspired by the current state of the field, this work introduces a metric for nonuse that meets all three criteria.

Previous work demonstrated that the bilateral arm reaching test (BART) can be used to reliably quantify nonuse (*15*). BART randomly lights up one of 100 equally spaced points between 10 and 30 cm in front of the user, and the user is required to reach to the lit-up point within a time limit. In the first condition, the user is instructed to choose either hand to reach the point as quickly and as accurately as possible. Because of the imposed time limit, the user must make a fast and spontaneous hand choice, even if they know they are being tested. In the second condition, the user is constrained to use only their stroke-affected arm to reach for the point. The automatic performance in the first condition is compared with the functional performance in the second condition to assess the level of nonuse. This approach has been shown to be both reliable and valid; however, it only assesses patients on a single plane of motion. Reaching tasks required to accomplish ADLs involve three-dimensional (3D) movements. In this study, we introduce a robot arm that enables a reaching task to quantify arm nonuse in three dimensions, allowing clinicians to tailor the rehabilitation process to specific patterns of nonuse as they occur in the user's real-world environment.

[1]Department of Computer Science, University of Southern California, Los Angeles, CA, USA. [2]Department of Biokinesiology and Physical Therapy, University of Southern California, Los Angeles, CA, USA.
*Corresponding author. Email: dennler@usc.edu

We describe the modified bimanual arm reaching test with a robot (BARTR), depicted in Fig. 1. The testing apparatus consisted of a general-purpose robotic arm that queried points in front of the user and a socially assistive robot (SAR) that supported the testing procedure by providing instruction and motivation. In a session of BARTR, the user completed two phases: a spontaneous phase and a constrained phase. Each phase can be completed in about 20 min. We used identical instructions to the original validated BART (*15*). In the spontaneous phase, the user was instructed to use the hand that can reach the button as quickly and accurately as possible. These instructions ensured that participants acted spontaneously while being aware that they were being tested. In the constrained phase, the user reached for the button with their stroke-affected hand. The nonuse metric, nuBARTR, was quantified from the reaching data collected from each session and repeated sessions that occurred at least 4 days apart, as in previous work (*15*).

To validate nuBARTR as a useful clinical metric, we developed the three following hypotheses based on the criteria for useful metrics in neurorehabilitation: First, nuBARTR is a valid metric, showing high correlation with the established metric for assessing nonuse, the amount of use (AOU) subscale of the AAUT. Second, nuBARTR is a reliable metric, having high test-retest reliability, as evidenced by high absolute agreement across repeated sessions taken at least 4 days apart. Third, nuBARTR is a simple-to-use metric, achieving a score of 72.6 of 100 or greater on the system usability scale (SUS), indicating above-average user experience as established in usability literature (*16*).

We found that nuBARTR satisfies these three criteria for a useful neurorehabiliation metric: It had high validity and high test-retest reliability, and study participants found it easy to use. The system can be used to aid clinicians in the quantification and tracking of stroke survivor arm nonuse.

RESULTS

We performed a user study with neurotypical and poststroke participants to evaluate the BARTR interaction. The nuBARTR was calculated from the BARTR interaction and assessed for the properties of useful neurorehabilitation metrics.

Participant demographics and stroke characteristics

Participants with chronic stroke were recruited from the Los Angeles, California area to take part in this study. Participants were recruited through the Institutional Review Board (IRB)–approved Registry for Aging and Rehabilitation Evaluation database of the Motor Behavior and Neurorehabilitation Laboratory at the University of Southern California (USC). All participants were right-hand dominant before their stroke. In total, 17 poststroke participants were recruited. Two participants did not meet the study criteria after screening, and one participant was excluded from analysis because of difficulties in completing the task. Of the 14 eligible participants, 12 completed all three sessions of the BARTR, and two were only able to complete two sessions because of scheduling constraints. One eligible participant did not receive AAUT scores because of technical problems in recording the exam. The average age of poststroke participants was 57 ± 11 years. Age and other participant demographic information is summarized in Table 1.

We also recruited 10 neurotypical adults to establish a normative value for performance. All neurotypical participants were right-hand dominant, and their demographic information is summarized in Table 2. The average age of neurotypical participants was 67 ± 10 years.

BARTR use characteristics

We performed multiple analyses to assess the validity of the BARTR interaction as a nonuse metric.

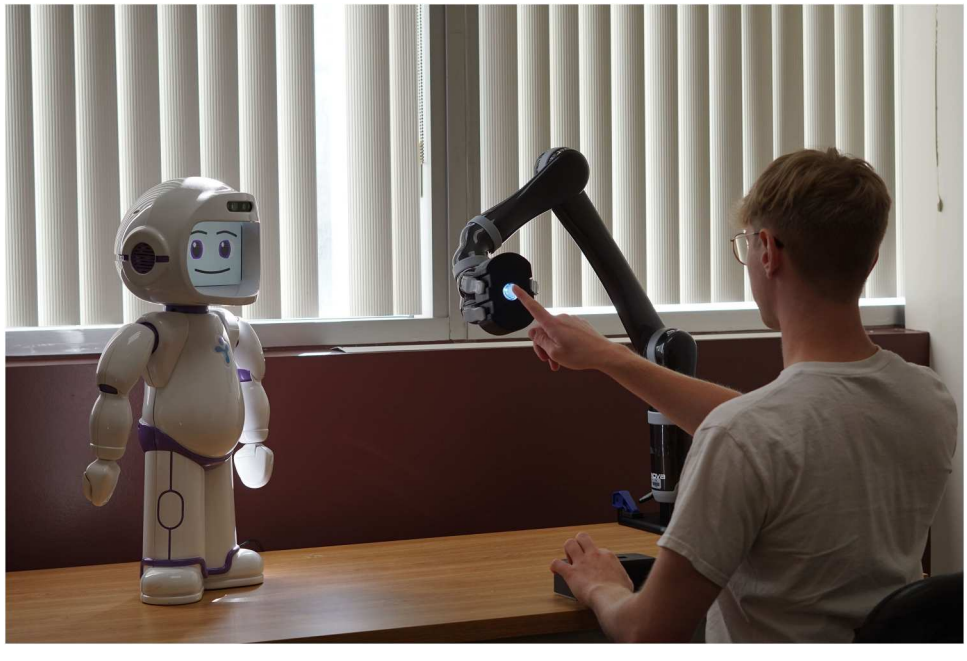

Fig. 1. Example reaching trial with the BARTR apparatus. The participant places their hands on the home position device. The SAR (left) describes the mechanics of the BARTR, and the robot arm (right) moves the button to different target locations in front of the participant. A reaching trial begins when the button lights up, and the SAR cues the participant to move.

Table 1. Demographic information for the poststroke group.

	Median	Minimum	Maximum
FM-UE motor score (66 maximum)	59.5	42	64
AAUT AOU score (1 maximum)	0.29	0.00	0.85
Age (years)	55	32	85
Time between sessions (days)	6.5	4	19
Gender	8 men, 6 women		
Affected side	5 left, 9 right		
Ethnicity	4 Asian, 2 Black, 4 Hispanic, 3 White, 1 mixed race		

Table 2. Demographic information for the neurotypical group.

	Median	Minimum	Maximum
Age (years)	69.5	45	82
Gender	5 men, 5 women		
Ethnicity	2 Asian, 2 Black, 6 White		

Arm use characteristics

To establish a normative baseline for comparisons, we examined the reaching data (time to reach and hand choice) from the 10 neurotypical participants. We found that, for the neurotypical group, there were no significant differences in average time from leaving the starting position to pressing the button across participant age (r^2 = 0.06 and P = 0.498) or gender (r^2 = 0.03 and P = 0.511), as evidenced by linear regressions. Similarly, for the neurotypical group, hand choice in the spontaneous condition had no significant differences due to participant age (r^2 = 0.004 and P = 0.861) or gender (r^2 = 0.015 and P = 0.739). Given the similarities across this group in performance on the BARTR task, we developed a single model of normative use based on an aggregate of the neurotypical participants' data.

A visualization of the neurotypical group's interaction metrics is shown in Fig. 2. On average, there was a handedness bias, where the right side was used to press the button in 60% of the workspace across participants, whereas the left side was used to press the button in 40% of the workspace, identically to the 60 of 40 handedness bias reported in the planar BART (*15*). In general, the time to reach the targets was relatively consistent for both hands, with farther points taking slightly more time, as expected.

In chronic stroke survivors, we observed high variability in hand choice and in the time to reach targets in the workspace. For illustrative purposes, in Fig. 3, we show these two interaction metrics modeled by Gaussian processes for two participants: one who was right-dominant affected and had high nonuse (P23) and one who was left–non-dominant affected and had low nonuse (P31).

These data plots highlight the importance of including 3D movement in the evaluation of nonuse. For example, P23 exhibited lower use of the right hand (63% left handed and 37% right handed), specifically in areas that appeared higher on the right side but maintained a high probability of using the affected arm for lower areas on the same side. P31 exhibited more symmetric use (37% left handed and 63% right handed) but also used the less-affected side slightly more often for points that were higher up and closer to the midline.

Selecting kernels for Gaussian process modeling

Three quantities were modeled through Gaussian processes to calculate arm nonuse: reaching success in the constrained phase of the BARTR, arm choice in the spontaneous phase of the BARTR, and reaching time for the affected arm across both phases of the BARTR. Success and arm choice are classification problems that leverage the Laplace approximation to model a non-Gaussian posterior with a Gaussian process, as is standard for classification (*17*). Reaching time is modeled directly as a regression problem. The results across several kernel choices for the Gaussian processes are shown in Table 3.

In the context of the reaching task, both the distance the hand travels and the spatial location of the target are important for predicting time to reach and the selected reaching arm (*15*, *18*). The kernels we tested were composed of two key components: the linear kernel and the radial basis function kernel. The linear kernel indicates that points of similar distances from the origin will have similar values and is defined as

$$k_{\mathrm{l}}(x, x') = \sigma_0^2 + x \cdot x' \tag{1}$$

where σ_0 is a hyperparameter learned from the data. The radial basis function kernel indicates that points near each other in space will have similar values and is defined as

$$k_{\mathrm{rbf}}(x, x') = \exp\left(-\frac{d(x, x')^2}{2\ell^2}\right) \tag{2}$$

where ℓ is the length-scale hyperparameter learned from the data.

The kernels were combined with addition and multiplication to represent different relationships between distance and locality, in accordance with recommendations from Duvenaud *et al.* (*19*). In total, 15 kernels were tested that combined five kernels that encapsulated the signal (k_{l}, k_{rbf}, $k_{\mathrm{l}} + k_{\mathrm{rbf}}$, $k_{\mathrm{l}} \times k_{\mathrm{rbf}}$, and $k_{\mathrm{l}} + k_{\mathrm{rbf}} + k_{\mathrm{l}} \times k_{\mathrm{rbf}}$) and five kernels that encapsulated the noise in responses (k_{n}, $k_{\mathrm{n}} + k_{\mathrm{l}} \times k_{\mathrm{n}}$, and $k_{\mathrm{n}} + k_{\mathrm{rbf}} \times k_{\mathrm{n}}$).

Kernels were evaluated through fivefold cross validation within each participant visit. The performance was averaged over all participant visits to evaluate each kernel. We selected kernels on the basis of their negative log marginal likelihood for the observed data. This value reflects the goodness of fit and additionally accounts for model complexity but is only applicable to Bayesian models. We additionally examined other non-Bayesian models to compare accuracies and negative log likelihoods.

We found that modeling the interaction data with Gaussian processes reached similar levels of accuracy and negative log likelihoods as other machine learning models. For completeness, we also evaluated the use of other classifiers and regressors for calculating our metric for nonuse and found them to perform similarly to Gaussian processes, as reported in the Supplementary Materials.

BARTR as a metric of nonuse

To evaluate BARTR as a metric for neurorehabilitation, we evaluated the three criteria of effective metrics: validity, test-retest reliability, and ease of use.

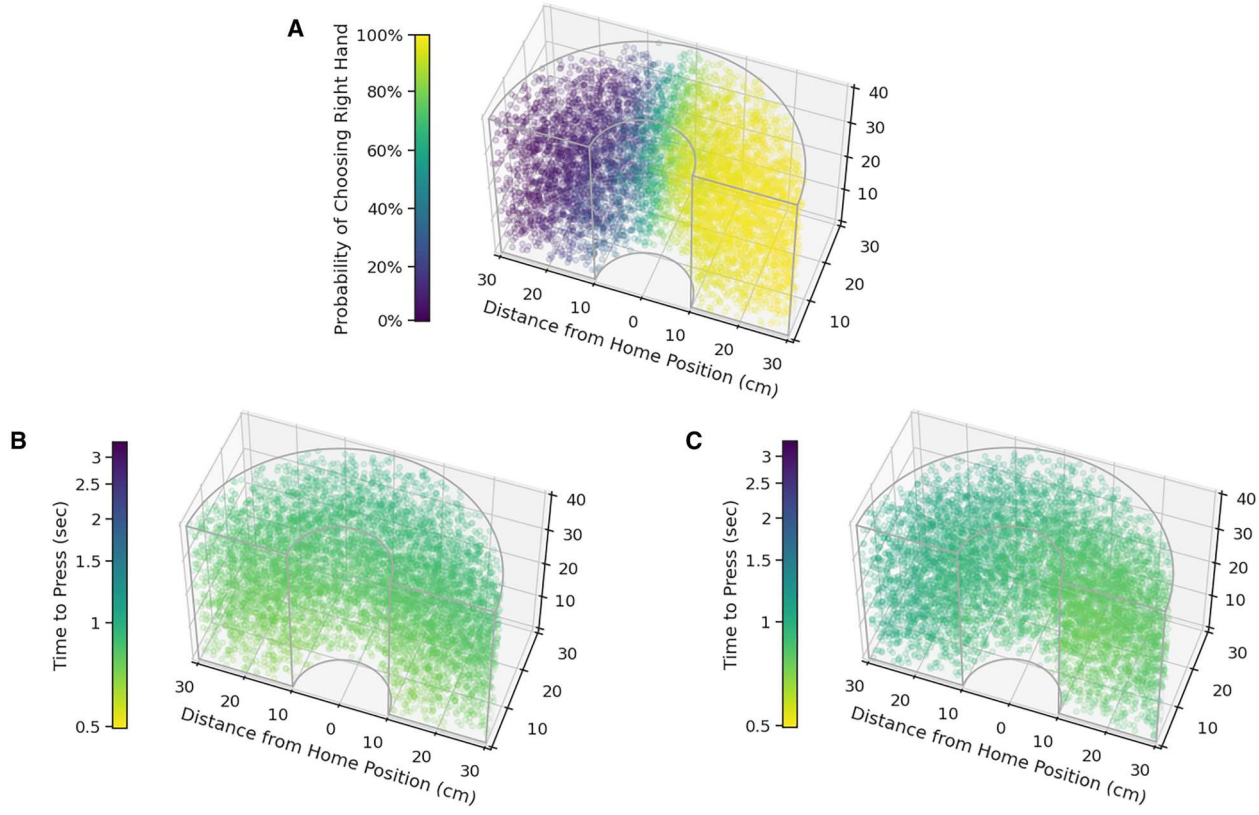

Fig. 2. Normative data collected from neurotypical participants. The normative data consists of hand choice in the spontaneous condition (**A**) and average time for participants to reach with their left hand (**B**) and their right hand (**C**). Lighter colors indicate high probability of participants choosing their right hand (A) or faster times to reach (B and C).

Validity

We evaluated the validity of BARTR by comparing the quantification of nonuse produced by the system with the values of nonuse collected from poststroke participants using the AAUT, the clinical standard for assessing nonuse. Participants had a wide range of nonuse, with AAUT AOU values ranging from 0.00 to 0.85 and nuBARTR scores ranging from 0.849 to 1.71. We determined the validity of nuBARTR with the nonparametric Spearman correlation between AAUT AOU and the averaged value of nuBARTR across the three sessions. Figure 4 shows that the calculated nonuse from BARTR is correlated with the clinical AAUT AOU metric of nonuse [$r(13) = 0.693$ and $P = 0.016$].

We also examined the correlation with the individual subscales of the AAUT with the nonparametric Spearman correlation. The cBARTR shows a high correlation with the cAAUT [$r(13) = 0.773$ and $P = 0.002$], and the sBARTR shows a correlation with sAAUT [$r(13) = 0.769$ and $P = 0.002$].

Test-retest reliability

We examined the absolute agreement [intraclass correlation coefficient (ICC)] of the three BARTR sessions to assess test-retest reliability. Absolute agreement of the BARTR metric is the recommended test of reliability in the medical field (*20*). We found that between sessions there was very high reliability of nuBARTR scores, ICC(1, k) = 0.908 and $P < 0.001$. A visualization of nuBARTR scores by participant is shown in Fig. 4.

We noted correlations between all pairs of sessions via a Pearson correlation. The first and second session are significantly correlated [$r(14) = 0.662$ and $P = 0.010$], the second and third sessions are significantly correlated [$r(12) = 0.948$ and $P < 0.001$], and the first and third sessions are correlated [$r(12) = 0.686$ and $P = 0.012$]. We examined scores across all three sessions and note that the BARTR interaction showed increased reliability after the first session, supporting repeated evaluations using this method to evaluate participants' nonuse over time.

Ease of use

To evaluate ease of use, we applied the standard, commonly used SUS (*16*, *21*, *22*). The SUS is scored out of 100 and calculated from 10 items. SUS meta-analyses provide full distributions of SUS scores across 446 extant systems and recommend evaluating systems on the basis of percentiles of systems examined in the meta-analysis (*16*). For example, a mean SUS score of 72.6 represents a system that is in the top 65% of all systems evaluated in the meta-analysis, and the meta-analysis provides a rating system for understanding these percentiles. A score of 78.9 or higher is in the "A" range, a score of 72.6 to 78.8 is in the "B" range, a score of 62.7 to 72.5 is in the "C" range, and a score of 51.7 to 62.6 is in the "D" range. The middle values of these ranges are denoted by the dashed lines in Fig. 4.

We administered the SUS to all participants enrolled in the study. To determine usability, we examined the SUS scores of only the poststroke group. The average rating of scores was 8.93 ±

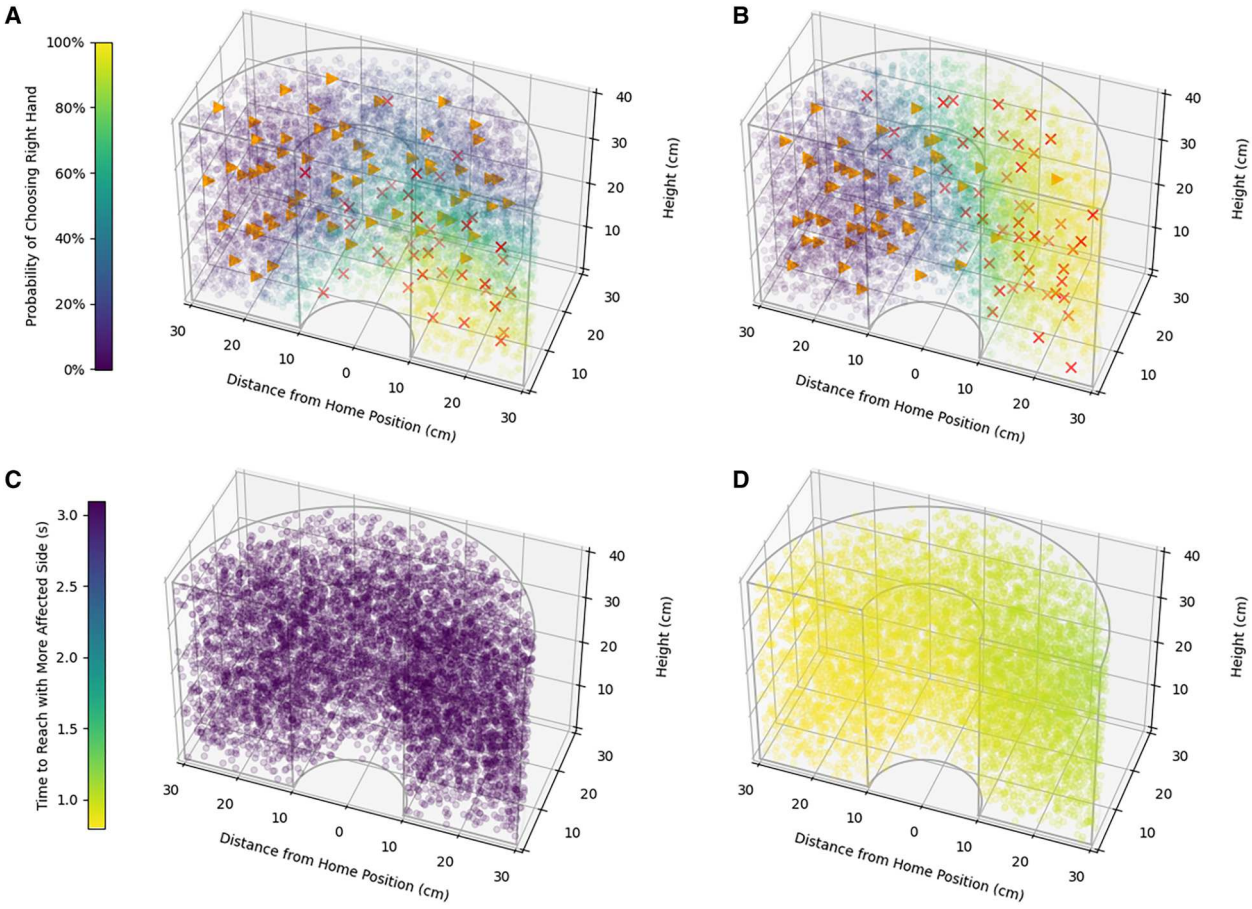

Fig. 3. Comparisons of data collected from two participants. P23 was right-dominant affected and showed lower right arm use (**A**) as well as longer reaching times (**C**). P31 was left–non-dominant affected and showed more balanced right arm use (**B**) and faster reaching times (**D**). Raw data are shown for arm choice data, with a red "x" denoting right hand reaches and an orange triangle denoting left hand reaches (A and B).

11.67, placing the mean usability of the BARTR apparatus in the 80th percentile of systems included in the SUS meta-analysis. Because of the high variance in participants' scores, we determined that the score is significantly greater than 72.6, which corresponds to an above-average user experience (*16*). We found from a nonparametric Wilcoxon signed-rank test that participants rated our system significantly above the 72.6 threshold (Z = 16.0 and P = 0.040). On the basis of this result, the system is easy to use and readily satisfies the ease of use criterion. The distribution of SUS scores across all participants is shown in Fig. 4.

Qualitative results

The qualitative analysis we performed considered the semi-structured interviews from 12 poststroke participants who completed all three study sessions. The full set of interview questions is provided in the Supplementary Materials. The interviews were conducted after the third session of the BARTR and lasted for an average of 14 min (minimum, 4 min and maximum, 44 min). The questions were structured around the four themes that prior work identified as important for interaction with rehabilitation systems (*23*): safety throughout interaction, ease of interpretation, predictability of actions, and adaptation of behaviors to task. We show an overview of the participants' responses to these four themes in Fig. 5. Positive responses described the system as being unequivocally helpful within the theme, mixed-positive responses described the system as helpful but provided room for improvement, and mixed responses were unsure whether the system was helpful with respect to the theme. No participants found the system unhelpful. We also report the participants' suggestions for improvement and future tasks.

Safety throughout interaction

All participants (n = 12) found the interaction to be safe. In addition to the safety precaution we took of moving the arm slowly, participants also reported feeling safe because they "figured [the experimenter] knew what [they] were doing" (P29) and that "it felt pretty safe because I had this shoulder harness on" (P27).

Some participants (n = 3) identified that they worried about the robot arm when it came close to the home position but reported that this did not influence how safe they felt throughout the interaction. One participant viewed the perceived risk as beneficial to them because "it was good to have my brain react to having it come close" (P36).

Ease of interpretation

All participants (n = 12) also found the robot easy to use. Most participants (n = 9) specified that they felt this way because the interaction itself was easy to learn. Participants found that they "got used to the robot after the first command it gave" (P37) and that the

Table 3. Modeling results for the three interaction metrics. Arrows indicate direction of better fits. Values with asterisks represent best values for each column. ACC, accuracy; NLL, negative log likelihood; NLML, negative log marginal likelihood; MSE, mean squared error; ME, maximum error; k_l, linear kernel; k_{rbf}, radial basis function kernel; $k_{comb.}$, $k_l * k_{rbf}$; N_1, constant noise; N_2 constant noise + linear noise; N_3, constant noise + radial basis function noise; k-NN, k-nearest neighbors; MLP, multilayer perceptron; SVM, support vector machine. Dash entries indicate not applicable.

Kernel	ACC↑	NLL↓	NLML↓	ACC↑	NLL↓	NLML↓	MSE↓	ME↓	NLML↓
$k_l + N_1$	0.838	0.330*	31.100	0.909	0.193	29.829	0.545	1.306	72.005
$k_l + N_2$	0.838	0.332	31.018	0.909	0.195	29.539	0.545	1.307	69.675
$k_l + N_3$	0.837	0.330*	31.100	0.909	0.193	29.829	0.545	1.306	72.005
$k_{rbf} + N_1$	0.838	0.342	31.664	0.912*	0.193	29.039	0.542	1.314	66.644
$k_{rbf} + N_2$	0.836	0.342	31.637	0.912*	0.196	29.003	0.546	1.313	64.941
$k_{rbf} + N_3$	0.836	0.342	31.660	0.912*	0.193	29.039	0.541	1.317	66.417
$k_l + k_{rbf} + N_1$	0.837	0.331	30.814*	0.910	0.193	28.933	0.540*	1.308	66.762
$k_l + k_{rbf} + N_2$	0.838	0.334	30.859	0.910	0.196	28.899	0.540*	1.306	65.210
$k_l + k_{rbf} + N_3$	0.839*	0.330*	30.865	0.910	0.193	28.933	0.541	1.307	66.762
$k_{comb.} + N_1$	0.834	0.345	31.209	0.910	0.193	29.071	0.542	1.305	66.462
$k_{comb.} + N_2$	0.836	0.346	31.274	0.911	0.195	29.017	0.541	1.304	64.863
$k_{comb.} + N_3$	0.836	0.345	31.268	0.911	0.192*	29.046	0.542	1.303*	66.462
$k_l + k_{rbf} + k_{comb.} + N_1$	0.837	0.342	30.870	0.910	0.192*	28.942	0.542	1.310	65.978
$k_l + k_{rbf} + k_{comb.} + N_2$	0.836	0.341	30.983	0.910	0.194	28.858*	0.541	1.307	64.718*
$k_l + k_{rbf} + k_{comb.} + N_3$	0.836	0.343	30.983	0.910	0.193	28.927	0.541	1.312	66.110
Non-GP method									
AdaBoost	0.809	0.570	-	0.889	0.264	-	0.586	1.399	-
k-NN	0.832	1.131	-	0.899	0.731	-	0.614	1.457	-
MLP	0.842*	0.331*	-	0.908	0.195	-	0.568*	1.368*	-
Random forest	0.834	0.437	-	0.899	0.327	-	0.617	1.476	-
Linear SVM	0.834	0.332	-	0.905	0.192*	-	0.606	1.453	-
RBF SVM	0.842*	0.339	-	0.914*	0.193	-	0.623	1.407	-

interaction was "a normal everyday task, so it wasn't hard to learn" (P27). Participants also found the task easy to learn because "there wasn't anything...that you have to put on" (P23). Two participants (P29 and P36) mentioned that they had done several other studies using other devices and that this interaction was easy because "it was all right in front of me and the instructions were clear" (P29).

Eight participants also directly described the SAR's voice as easily understandable. One participant (P29) noted that they "liked the mouth moving, it helped to understand the speech" of the SAR providing instructions. In addition to understanding the words, another participant (P27) also found that "the voice was comforting and the instructions were very clear." Participants found the instruction from the SAR valuable toward understanding the task as well as socially motivating.

Predictability of actions

Seven participants directly commented on the predictability of the interaction. The comments addressed both the physical predictability of the task and the social predictability of the SAR. Participants found the task predictable because it was repetitive and simple. A participant (P37) found this to be particularly important because "with stroke you're also going through a psychological situation, and with [this task], you don't have to grapple with anything. This way is straightforward."

Participants had a variety of interpretations of the social component of the interaction. Because we used randomness in the SAR's movements and feedback to make it appear more natural and lifelike [as is standard in human-robot interaction work (*24*–*26*)], some participants viewed the unpredictability as a benefit. One participant (P27) referred to the unpredictable social behaviors as "natural" and thought the SAR "doesn't feel like technology"; another participant (P25) became engaged in "trying to find a pattern in the robot's eyes." Another participant (P21) had a more neutral reaction to the randomness and said, "The fluctuations in cuing, I don't know if that was a hindrance or a help." One participant (P37) greatly appreciated that the exercise was led by a robot, because the overall social interaction was predictable and the robot was not getting tired, and stated, "With the SAR it is like no judgement...there is no feeling of changing in the delivery...if a person had to repeat 'go, go, go', sometimes they might get tired, and when you're doing the exercise you can see that."

Adaptation of behaviors to task

Eight participants commented on how the system could adapt to them specifically throughout the task. The participants were also concerned with either the task or the social component of the task. For the task, six participants identified that the robot could adapt more to different levels of task difficulty. With the goal of developing a standardized test, the robot sampled points randomly in

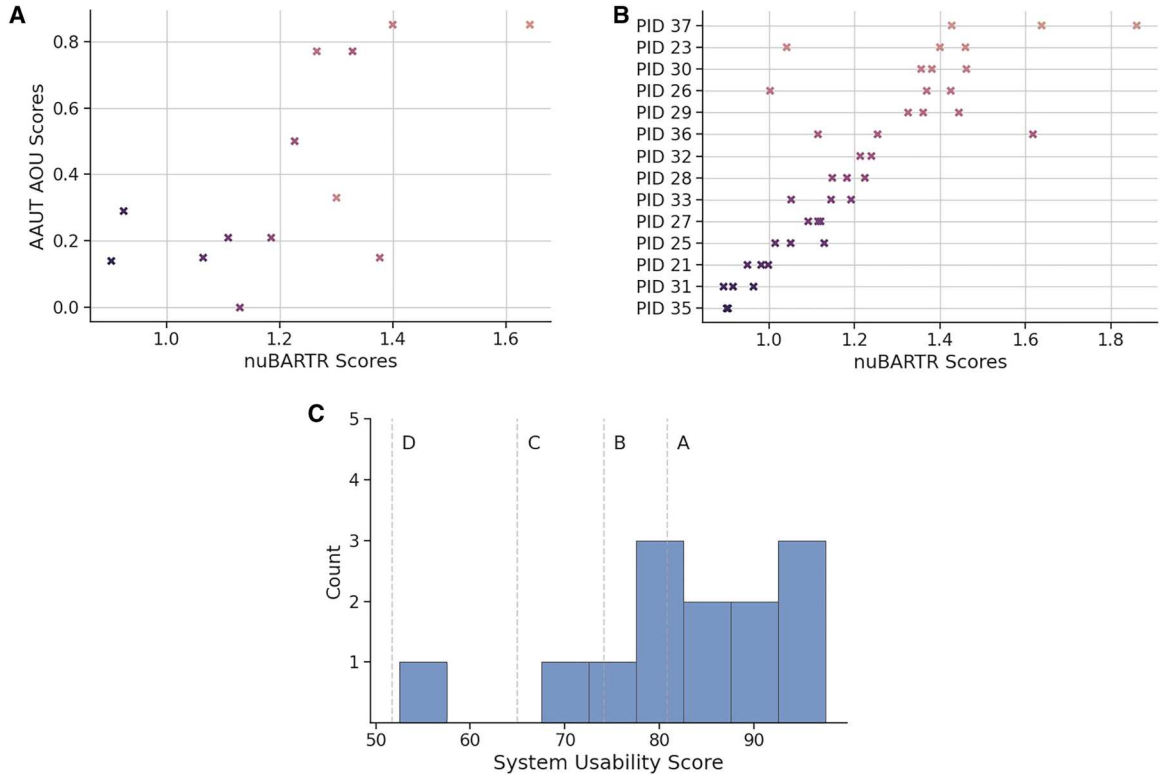

Fig. 4. Evaluations of the proposed metric. We demonstrate the BARTR metrics' validity through its correlation with clinical measurements of nonuse through a nonparametric Spearman correlation, $r(13) = 0.693$ and $P = 0.016$ (**A**). We demonstrate reliability with the absolute agreement of BARTR scores across three sessions through the ICC(1,k) = 0.908 and $P < 0.001$ (**B**). We demonstrate its ease of use through usability ratings of the system, showing that the average rating is above 72.6 through a nonparametric Wilcoxon signed-rank test, $Z = 16.0$ and $P = 0.040$ (**C**).

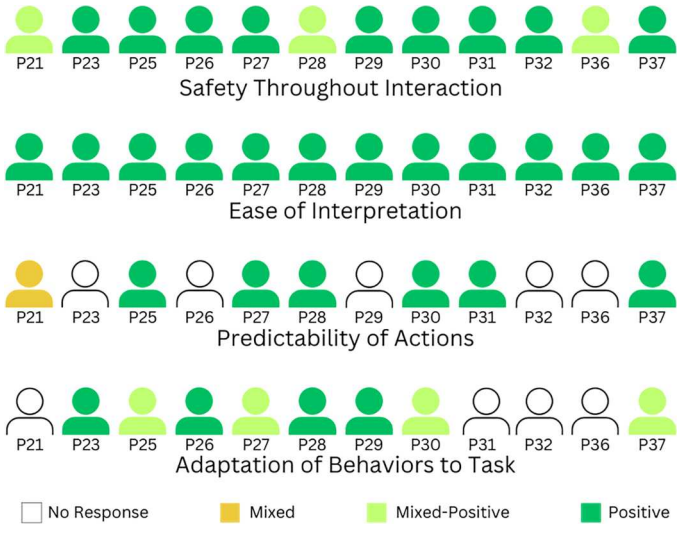

Fig. 5. Qualitative responses from participants. We show overall perceptions of each of the four factors of trust (*23*) that each participant mentioned.

the interaction, but participants asked whether the arm "could go all the way up or all the way back...it would be nice if I could extend my whole arm" (P25), while recognizing that "if you had more damage in your arm it would be harder to do" (P28). Four participants also described the timing of the robot placing points. Three of them found the speed appropriate; one of these participants (P37) described it as "when the arm was moving it was moving at the right speed." One participant (P21) thought that the arm "could move faster or something...it was very methodical where it went."

With regard to the SAR's verbal communication, four participants described the feedback that the SAR gave as evidence of it adapting to their good performance. P26 also specified that the SAR's progress updates were helpful because they "gave you an idea of where you stand at the time." However, one participant (P30) wished that the SAR would be "more responsive" to the specifics of their performance, for example, by commenting on how fast their reach was.

Suggestions for improvements and future tasks

Participants also provided feedback on how the system could be improved or adapted to other forms of exercises for evaluating arm nonuse. The suggestions for improvement largely addressed how the system could be more personalized to individual tastes. Participants discussed how visual components could be adapted, for example, how the SAR's exterior could "change to USC colors, which would work better...I have some stickers I could put on the robot" (P37) or how the button could "turn green when you press it" (P23). Other participants described how the SAR's audio could be personalized by "choosing music to play, just to make it more pleasant" (P25). Participants also suggested gestures for the robot to perform, such as "when you make a mistake, you could have the robot hold its arms up and point to the button" (P37).

Despite these suggestions, participants (n = 10) described the system as being effective and helpful. Several (n = 4) explicitly stated that they thought about the interaction outside of the experiment. One participant reported that when they were "trying to open a cabinet, I had a flashback to this button pressing when I was thinking about how to orient my hand to open the cabinet" (P37). Participants found the "fact that it is 3D is effective" (P37) and suggested several other 3D interactions that would be useful.

The most popular tasks that the participants described as being useful were "3D tasks that involved more finger dexterity" (n = 7). Participants described how the interaction could "integrate a little ball...because once you put it in your hand your fingers start working" (P36) or how the robot arm could "hold a pocket or something and have people put pennies over here or over there" (P37). One participant also described how they would like to control the robot to practice finger dexterity by using "a glove or something to control the robots, so you simulate grabbing and the robot moves with the glove" (P23).

The second type of task that multiple participants suggested was gross motor tasks (n = 4). For example, two participants suggested using the robot arm to passively move their more stroke-affected side by "grabbing what the robot is holding and have it drag my arm around" (P25). Two other participants suggested actively pushing against the arm as a form of strength training. One participant suggested "you could add on pressure sensing...I am interested in seeing the pressure and strength of both sides" (P27).

DISCUSSION

This work introduced robotics as an enabling methodology for the evaluation of difficult-to-evaluate yet clinically substantial constructs such as arm nonuse poststroke. We demonstrated that robots can provide a way to objectively and reliably assess motor behaviors that are meaningful for neurorehabilitation. The following discussion highlights the efficacy of using a SAR and a robot arm together for rehabilitation and evaluation to complement the work of neurorehabilitation clinicians. The quantitative and qualitative results and insights contributed by this work, including the interviews with poststroke participants after their BARTR sessions, inform the design of future systems for effective rehabilitation and assessment of patients' rehabilitation progress.

Robots as tools for evaluation

A key benefit of using a robotic system for administering rehabilitation assessments is in enabling highly controlled, repeatable, and precise measurements. Robot arms/end effectors can administer tests with exact instruction, intonation, pacing, and placement of reaching targets as well as randomize variables that may affect outcomes, allowing these variables to be assessed covertly, an important aspects of valid assessment methods. In addition, SARs can provide instruction and motivation for sustained effort of long-term rehabilitation exercises. Users do not need to wear any sensors, allowing for unencumbered, natural behavior more representative of ADLs, a key consideration that allows the BARTR to achieve high reliability and usability compared with other interactions that rely on wearable sensors to collect data (*27*), because worn sensors can affect behavior by encumbering or being uncomfortable (*28*).

In the context of other metrics, using robotics provides the benefit of precise spatial information, enabling quantification of areas of difficulty for arm use. Precise quantification of those regions can be tedious or difficult for clinicians to obtain, yet the data can support the development of personalized therapy regimens. Once the regions are quantified, they can be used to adapt the system's behavior to select targets for the user that are at the appropriate level of challenge, with sufficient variety as well. Another alternative is to use environmental or ambient sensors to track human movement, but such sensors can present privacy concerns.

With the increased richness of the resulting data from robotic systems administering physical assessments, more nuanced information can be communicated to rehabilitation therapists and clinicians. For example, machine learning techniques can be applied to summarize a patient's data for easy visualization by clinicians, who can then indicate where in the workspace the patient may need to focus most. These techniques can be connected with data visualization techniques to effectively communicate particular measurements of interest and enable more personalized care.

Past research has also shown that end-effector robots can be functionally effective in rehabilitation (*29*). This work further demonstrates that such robots have the potential to be used for assessments that are otherwise difficult to obtain because of lack of reliability or ease of use. Our metric of nonuse is not specific to a particular robot embodiment and can be applied to any robot that

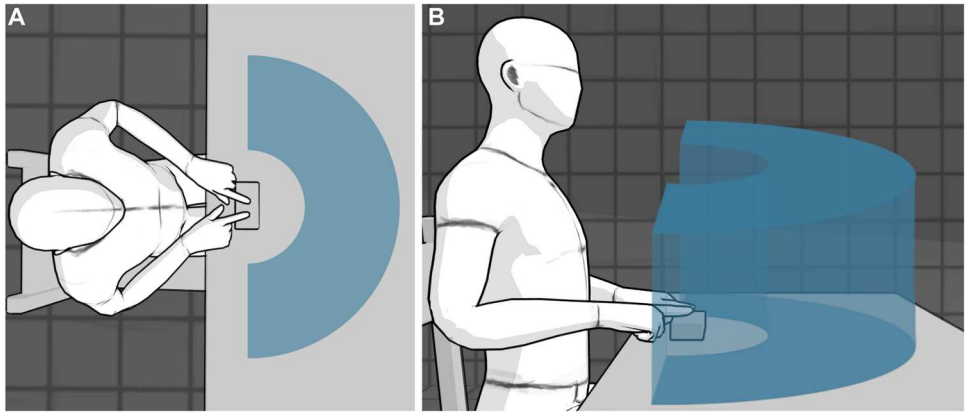

Fig. 6. Visualization of the participant's workspace. Viewed above, the workspace tested extends radially from the home position from a distance of 10 to 30 cm (**A**). Viewed from the side, the workspace extends upward 40 cm (**B**).

can reach points within the 3D space around the patient in a cylindrical region 30 cm in radius and 40 cm in height (as illustrated in Fig. 6). Such robots can be used for other forms of assessment (*30*), adaptive exercise practice that requires a model of the difficulty of reaching different points in the participant's workspace (*31*, *32*), and other gamified tasks to practice reaching (*33*). In addition to rehabilitation exercise, other assistive tasks throughout the rehabilitation process can be completed by end-effector robots, such as dressing (*34*), hair combing (*35*), shaving (*36*), etc. Because of their multiple possible uses and their portability, end-effector robots have the potential to facilitate the in-home rehabilitation process. Analogously, SAR systems have been shown to be effective in increasing user motivation for a wide variety of tasks, from physical exercise (*37*) to cognitive and social skill learning (*38*), in early work on SAR for supporting stroke rehabilitation exercises (*39*). This paper demonstrates a synergistic combination of both a robot arm and a SAR in a rehabilitation context.

Considerations for using robots for in-home assessments

Incorporating robots into home environments is a broad and highly active area of robotics research. In-home rehabilitation, also called in-place rehabilitation, is a frontier that presents a unique set of challenges. Naturally, safety is a key concern. In this work, a robotics specialist was present for all sessions to monitor any potential system failures. More than 8000 trials were conducted as part of this work, and there was only one case of the robot moving outside of the workspace. The error was quickly corrected, but such expert intervention would not be readily available for in-home systems. Although all of our study participants reported not feeling scared or anxious of the robots' movements, future unsupervised systems must be developed with considerations for all failure cases, no matter how improbable.

Another concern is the privacy of the collected user data. Some data used in this work were personally identifiable; for example, the robot used the participants' names when addressing them, which was appreciated by participants, who noted that it added to the interaction (P25). Although identifiable data are valuable for engaging and personalizing interactions, they also represent a data security risk.

Another barrier to bringing robots to users' homes is cost. Design guidelines for robotic systems for rehabilitation therapy indicate that, to be useful in home settings, such systems must cost less than US $10,000 (*40*). Hospital loans and insurance coverage are necessary to subsidize patient costs. Low-cost (under US $1000) SAR systems have already been developed (*41*), and more cost-effective end-effector robots are also being developed (*42*). In our work, poststroke participants reported that they would miss the robot after their third session. If the system is used for a longer period of time in the home, a stronger bond may develop, and thus future work should evaluate how removal of the system may affect users (*43*).

Limitations and future work

Although the AAUT is still used for clinical evaluation of nonuse, it may have become outdated. It includes tasks such as inserting a Polaroid into a photo album and opening a physical newspaper; such tasks are commonly accomplished digitally and therefore may compromise the covert nature of the test. Although no participants reported being aware that they were being observed before the nature of the test was revealed, outdated/atypical tasks may also influence arm use because of task unfamiliarity within the context of ADLs. The question of aptness of the AAUT provides a further impetus to develop additional assessments of nonuse, such as the BARTR.

Although we show that the BARTR meets the criteria for a good neurorehabilitation metric, we note that there are two important considerations to make for its use in clinical contexts. First, we observed much higher reliability in the second and third BARTR sessions and variability in some of the participants between sessions. We recommend one abbreviated session to habituate the participants to the BARTR assessment for increased reliability. Second, the BARTR is focused on reaching motions and does not include finger/hand manipulation tasks. Therefore, a limitation of the presented study is that the button task was relatively simple compared with the fine manipulation tasks used to evaluate nonuse in the AAUT, such as removing business cards from a box and placing a picture in a scrapbook. Such tasks may not be correctly evaluated through our current system; however, future extensions of the BARTR test can include fine manipulation tasks. Our qualitative results show participants suggesting the inclusion of additional held objects as part of the BARTR trials to enable the assessment of grasping and stabilization motions in addition to reaching motions (*44*). The participants suggested tasks that involved the manipulation of pennies and golf balls; the BARTR could include a screwdriver-like instrument to press the button, analogous to the cylindrical grasp item of the Fugl-Meyer upper extremity (FM-UE). Such additional tools can also be equipped with sensors to evaluate grasp force, which cannot be evaluated visually (*45*).

Our results are limited to only right-hand–dominant participants; we did not recruit any left-handed poststroke participants to reduce variance in our tested population, because handedness affects baseline hand choice. Previous work has shown that the side of the stroke lesion is more important in determining limb selection than pre-morbid handedness (*46*). As more data are collected from BARTR sessions, stronger results can be drawn about premorbid handedness in this specific task. However, because of well-known difficulties in recruiting premorbidly left-hand–dominant stroke survivors, data scarcity may be a barrier for evaluating all participants' nonuse. A potential direction may be to evaluate whether techniques from domain adaptation (*47*) may be used to leverage data collected from right-handed participants to perform more accurate assessments of future left-handed users.

Our qualitative results are influenced by the particular identity of the analyst. One author (N.D.) conducted the interviews and analyzed the interview data. In the spirit of reflexivity (*48*), it is important to understand that the analysis represents the views of a nondisabled computer scientist who designed the interaction and values the co-construction of knowledge with participants. N.D. has experience both as a facilitator and a participant of participatory design in several contexts and views the feedback of the users of systems as a requirement in the design process.

The most important direction for future work is to deploy and evaluate this methodology for longer periods of time as a tool for clinicians to assess and develop progress throughout the patients' rehabilitation process. Several opportunities for further research would arise from such deployments. Researchers could examine how information is communicated to neurorehabilitationists, how information from the BARTR test can improve functional outcomes or other behavioral metrics in users, how SAR personality and

communication styles may be adapted to participants over time, how BARTR tests could use data from previous sessions with the same participant to be more efficient and engaging, and how assessment can be combined with regular rehabilitation exercises. Overall, the combination of social and functional components offers a unique opportunity for more personalized, engaging, and effective human-robot interaction for neurorehabilitation.

MATERIALS AND METHODS

Participants

Participants who had experienced a hemispheric stroke at least 6 months before enrollment were recruited for this study. Participants were screened for eligibility before the interaction and were deemed eligible if they satisfied the following criteria: 18 years of age or older, able to reach at least 30 cm anterior to the midline of the trunk and at least 30 cm high without pain or assistance, normal or corrected-to-normal hearing and vision, proficiency in understanding English, right-hand–dominant prestroke, and a score greater than 25 of 66 on the Fugl-Meyer upper extremity motor assessment (*49*).

The Fugl-Meyer assessment was administered to all participants by a board-certified physical therapist specializing in neurorehabilitation who had more than 2 years of experience. We also administered the mini-mental state exam (MMSE) (*50*) to ensure that the participants could give consent. If a participant scored lower than 25 of 30 on the MMSE, a caregiver was required to be present as a witness for consent.

We additionally recruited neurotypical participants of similar ages to establish normative use of the robot system. These participants provided a baseline for zero nonuse, because neurotypical participants favor their dominant hand in bimanual tasks (*15*). The purpose of the neurotypical adult group was to establish a normative value for handedness bias and for the time it takes to reach different points in the task workspace.

All participants reviewed and signed a consent form before the experiment. Participants with chronic stroke performed up to three sessions using the BARTR testing apparatus, scheduled at least 4 days apart. Neurotypical adults performed one session with the BARTR testing apparatus. Participants were paid US $50 for each hour-long session they completed. All study protocols and consent forms were approved by the USC IRB under #UP-22-00461.

Actual amount of use test

Before the first session of BARTR, the experimenter (N.D.) administered the AAUT to the chronic stroke survivors, and the research physical therapist (A.C.) rated performance from the offline video data. The AAUT is a covert assessment of spontaneous arm use for 14 tasks that regularly occur in daily life, such as pulling out a chair from a table before sitting in it and flipping through the pages of a book. First, the tasks were completed covertly (spontaneous $AAUT_s$), so the participant did not know that they were being video-recorded and tested. Then, the experimenter revealed that arm use was being observed, and participants completed the 14 tasks again while being encouraged to use their stroke-affected arm as much as possible (constrained $AAUT_c$).

The research physical therapist rated both the AAUT amount (binary yes or no) if the participant attempted to use their stroke-affected arm (AAUT AOU score) for that task and the AAUT quality of movement (QOM; on an ordinal scale of 0 to 5) if they used the paretic arm in the task. In the context of this study, we considered the AAUT AOU score as the metric of interest for three reasons. First, the AOU and QOM scales have been found to be redundant (*51*). Second, in our sample, the AOU achieved higher values of internal consistency for each participant ($\alpha = 0.87$) than the QOM scale ($\alpha = 0.72$). Last, the poststroke participants exhibited a high coverage of the values of the AOU scale. The final nonuse score for each participant was calculated as the average of the differences between the constrained AAUT AOU score and the spontaneous AAUT AOU score over all tasks, resulting in a scalar value between 0 and 1.

Testing apparatus

The BARTR apparatus, designed to test arm nonuse, consists of a robot arm and a SAR. The robot arm was the Kinova JACO2 assistive arm (*52*), selected because it has already been used in assistive domains, is lightweight, and safely interacts with and around people. The arm has the same affordances as end-effector robots typically used for other rehabilitative interactions that have been shown to be effective in the rehabilitation context (*29*). The SAR was the Lux AI QTRobot (*53*) that consists of a screen face on a 2–degree-of-freedom head and two 3–degree-of-freedom arms that can gesture. This SAR platform has already been validated in our past work with children with arm weakness due to cerebral palsy (*24*) and in other human-robot interaction contexts (*54*). The SAR provided the participant with verbal instructions at the start of the BARTR session and with positive feedback on a random schedule, similarly to previous SAR use in other rehabilitation contexts (*24*, *55*, *56*).

In addition to the two robots, we developed two low-cost devices for the BARTR apparatus: the target object and the home position. Both devices were 3D printed, had self-contained power supplies and processors, and communicated wirelessly with the BARTR apparatus using low-level UDP protocols.

The target device, held by the robot arm, consisted of a 3D-printed housing with a single button. It received commands to turn on a light and start a timer to begin each reaching trial and logged the time taken by the participant to reach for and press the button to turn off the light.

The home position was the location that participants returned to between reaching trials, implemented as a 3D-printed block with two shallow holes 2 cm in diameter and 5 cm apart, with capacitive touch sensors inside. Participants placed their left pointer finger in the left hole and their right pointer finger in the right hole. The device communicated at 20 Hz, reporting the locations that were being actively touched by the participant.

Bilateral arm reaching test with a robot

Participants were seated at a table with the home position aligned with the center of their chest, as shown in Fig. 6. They were instructed to maintain approximately 90° angles of their elbows when their index fingers were resting at the home position. To limit upper-body compensatory movement, participants wore a shoulder harness attached to the chair (*57*). Participants were instructed to verbally cue the experimenter when they were ready to begin each section of BARTR. Following previous work, the two experiment phases were the spontaneous BARTR phase (sBARTR), where the participants were instructed to use either arm to reach the target,

and the constrained BARTR phase (cBARTR), where the participants were instructed to use their more-affected arm to reach the target (*15*).

For both phases of BARTR, the robot arm placed the reaching target at a different location in 3D space in front of the participant. The participant was instructed to reach the target as quickly and accurately as possible when prompted by the SAR. Each reaching trial began with the robot arm moving to one of the randomly sampled locations. When the robot arm arrived at the location and the participant was in the home position, the light on the target device turned on, and the SAR cued the participant to reach to the target after a random interval between 0 and 2 s to prevent the participant from anticipating movement to the target. After the audiovisual cue, the participant was given 3.1 s to reach to the target. When the participant pressed the button, the light turned off. If the participant did not reach the target in 3.1 s, the light turned off after the 3.1 s had elapsed. This time period was selected to make the maximum time of each experiment phase approximately 20 min in duration, given the variability in travel time between points for the robot arm. This period was sufficient for all neurotypical participants to reach all of the target placements.

In total, 100 locations were tested each for sBARTR and cBARTR. The locations were evenly spaced in the 3D workspace volume in front of the participant, defined by the region that was 10 to 30 cm from the center of the home position, forming a semicircle that extended in front of the participant in their transverse plane, and heights that ranged from 0 to 40 cm above the table, as shown in Fig. 6. These points were selected randomly without replacement, namely, each point was selected exactly one time; participants reached for all 100 targets one time per session. Participants attempted up to 100 reaching trials for each section of BARTR, for a total of 200 reaches.

Calculation of the BARTR metric

We used the data collected through the BARTR interaction to estimate a user′s workspace. Following previous work, nonuse was modeled as the subtraction of two components: the constrained component and the spontaneous component (*15*). The constrained component of the workspace W is defined for every point $x \in W$ for a particular participant p as

$$\mathrm{cBARTR_p}(x) = p_\mathrm{p}(\mathrm{success} \mid X = x, S = s_\mathrm{p}) \tag{3}$$

where $p_\mathrm{p}(\cdot)$ denotes the function that returns the probability of the poststroke participants selecting each side in the spontaneous condition. The side of the participant that was affected by stroke is denoted as s_p and is in the set of values {′left′, ′right′}. This quantity represents the total area that the participant is expected to be able to reach within the time limit, 3.1 s, based on the times from the neurotypical group.

The spontaneous component of the workspace is defined over all points $x \in W$ as

$$\begin{aligned}\mathrm{sBARTR_p}(x) = {} & p_\mathrm{p}(S = S_\mathrm{p} \mid X = x) * p_\mathrm{n}(S = s_\mathrm{p} \mid X \\ & = x) * E[t_\mathrm{n}^{s_\mathrm{p}}(x) - t_n^{s_\mathrm{p}}(x) \mid X = x]\end{aligned} \tag{4}$$

–where $p_\mathrm{p}(\cdot)$ denotes the probability of the poststroke participants selecting either side in the spontaneous condition and $p_\mathrm{n}(\cdot)$ denotes the probability of the neurotypical group selecting either side in the spontaneous condition. $t_\mathrm{n}^{s_\mathrm{p}}(x)$ and $t_\mathrm{n}^{s_\mathrm{p}}(x)$ represent the movement time for the poststroke and neurotypical participants, respectively, to reach the point x in the workspace with the arm on the participant′s more affected side, s_p. This quantity represents how close the participants′ spontaneous arm use is to spontaneous neurotypical use. Higher usage of the participant′s more-affected arm and faster movements result in higher spontaneous scores.

The final value for nonuse is calculated as the difference of these functions summed over all of the points in the workspace:

$$\mathrm{nuBARTR} = \sum_{x \in W} \mathrm{cBARTR}(x) - \mathrm{sBARTR}(x) \tag{5}$$

To obtain these values, we modeled the interaction metrics, time to reach points and arm choice, as Gaussian processes for the normative participants and for each poststroke participant. We summed more than 10,000 samples from a uniform distribution over the workspace to accurately estimate the difference of these two functions.

User-reported data

In addition to the data-driven evaluation of nonuse, we asked participants for their perceptions about using the system with two self-reported surveys: the SUS (*16*, *21*, *22*) and a semi-structured interview (*58*).

SUS

The SUS (*21*) is a 10-item scale that assesses the ease of use of a technological system. Each item is rated on a five-point Likert scale that ranges from "strongly agree" to "strongly disagree." Five of the items are positively worded, where higher ratings indicate a highly usable system, and five items are negatively worded, where lower ratings indicate a more usable system. This scale was selected for its high reliability, validity, and broad applicability to technological systems. Participants completed the SUS after their second session with the system.

Semi-structured interviews

The combination of a SAR and a robot arm has the potential to support participants throughout their rehabilitation process at home. Because of the socially interactive nature of the BARTR session, rehabilitation effectiveness depends on how much users can trust the robots to help them. Trust in rehabilitation robotics has four key facets: safety throughout the interaction, predictability of actions, ease of interpretation, and adaptation of behaviors to the task (*23*). We developed questions to identify ways that our system could support participants in other exercises and how the system they interacted with achieved or did not achieve the above four key aspects of trust. Participants qualitatively reflected on their experience of the system and discussed improvements for future systems after the third session.

One of the authors conducted and analyzed all of the interviews. We used an iterative four-phase deductive qualitative analysis approach to analyze the interview data (*59*). The first phase consisted of the transcription and open coding of interview data, assigning specific meaning to the phrases participants spoke. The second phase grouped similar codes into subthemes. The third phase categorized the subthemes according to the themes found in previous research in rehabilitation with SARs (*23*). Some codes did not fit into the theorized categories, and thus new emergent categories were developed. The fourth phase iterated on the second and

third phases, developing categories on a subset of six interviews that were expanded to include all the interviews until the final categorization reached theoretical saturation, similar to the method used by Ando *et al.* (*60*).

Supplementary Materials

This PDF file includes:

Text
Fig. S1
Tables S1 to S3

REFERENCES AND NOTES

1. C. W. Tsao, A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, Y. Commodore-Mensah, M. S. V. Elkind, K. R. Evenson, C. Eze-Nliam, J. F. Ferguson, G. Generoso, J. E. Ho, R. Kalani, S. S. Khan, B. M. Kissela, K. L. Knutson, D. A. Levine, T. T. Lewis, J. Liu, M. S. Loop, J. Ma, M. E. Mussolino, S. D. Navaneethan, A. M. Perak, R. Poudel, M. Rezk-Hanna, G. A. Roth, E. B. Schroeder, S. H. Shah, E. L. Thacker, L. B. VanWagner, S. S. Virani, J. H. Voecks, N.-Y. Wang, K. Yaffe, S. S. Martin, Heart disease and stroke statistics—2022 update: A report from the American Heart Association. *Circulation* **145**, e153–e639 (2022).
2. N. E. Mayo, S. Wood-Dauphinee, R. Côté, L. Durcan, J. Carlton, Activity, participation, and quality of life 6 months poststroke. *Arch. Phys. Med. Rehabil.* **83**, 1035–1042 (2002).
3. C. Winstein, B. Kim, S. Kim, C. Martinez, N. Schweighofer, Dosage matters: A phase Ib randomized controlled trial of motor therapy in the chronic phase after stroke. *Stroke* **50**, 1831–1837 (2019).
4. M. Rensink, M. Schuurmans, E. Lindeman, T. Hafsteinsdottir, Task-oriented training in rehabilitation after stroke: Systematic review. *J. Adv. Nurs.* **65**, 737–754 (2009).
5. R. Stanton, L. Ada, C. M. Dean, E. Preston, Biofeedback improves performance in lower limb activities more than usual therapy in people following stroke: A systematic review. *J. Physiother.* **63**, 11–16 (2017).
6. S. L. Wolf, C. J. Winstein, J. P. Miller, E. Taub, G. Uswatte, D. Morris, C. Giuliani, K. E. Light, D. Nichols-Larsen, Effect of constraint-induced movement therapy on upper extremity function 3 to 9 months after trial. *JAMA* **296**, 2095–2104 (2006).
7. World Health Organization, ICF: International Classification of Functioning, Disability and Health (2001); www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health.
8. E. Taub, J. E. Crago, G. Uswatte, Constraint-induced movement therapy: A new approach to treatment in physical rehabilitation. *Rehabil. Psychol.* **43**, 152–170 (1998).
9. K. Andrews, J. Stewart, SROKE recovery: He can but does he? *Rheumatology* **18**, 43–48 (1979).
10. L. J. Buxbaum, R. Varghese, H. Stoll, C. J. Winstein, Predictors of arm nonuse in chronic stroke: A preliminary investigation. *Neurorehabil. Neural Repair* **34**, 512–522 (2020).
11. D. T. Wade, Measurement in neurological rehabilitation. *Curr. Opin. Neurol. Neurosurg.* **5**, 682–686 (1992).
12. G. Uswatte, E. Taub, D. Morris, K. Light, P. Thompson, The motor activity log-28: Assessing daily use of the hemiparetic arm after stroke. *Neurology* **67**, 1189–1194 (2006).
13. A. Sterr, S. Freivogel, D. Schmalohr, Neurobehavioral aspects of recovery: Assessment of the learned nonuse phenomenon in hemiparetic adolescents. *Arch. Phys. Med. Rehabil.* **83**, 1726–1731 (2002).
14. S. Chen, S. L. Wolf, Q. Zhang, P. A. Thompson, C. J. Winstein, Minimal detectable change of the actual amount of use test and the motor activity log: The excite trial. *Neurorehabil. Neural Repair* **26**, 507–514 (2012).
15. C. E. Han, S. Kim, S. Chen, Y.-H. Lai, J.-Y. Lee, R. Osu, C. J. Winstein, N. Schweighofer, Quantifying arm nonuse in individuals poststroke. *Neurorehabil. Neural Repair* **27**, 439–447 (2013).
16. J. R. Lewis, The system usability scale: Past, present, and future. *Intl. J. Hum–Comp. Int.* **34**, 577–590 (2018).
17. C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, vol. 1 (MIT Press, 2005).
18. A. Roby-Brami, A. Feydy, M. Combeaud, E. V. Biryukova, B. Bussel, M. F. Levin, Motor compensation and recovery for reaching in stroke patients. *Acta Neurol. Scand.* **107**, 369–381 (2003).
19. D. Duvenaud, J. Lloyd, R. Grosse, J. Tenenbaum, G. Zoubin, Structure discovery in nonparametric regression through compositional kernel search. *International Conference on Machine Learning* (PMLR, 2013), pp. 1166–1174.
20. T. K. Koo, M. Y. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **15**, 155–163 (2016).
21. J. Brooke, SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* **189**, 4–7 (1996).
22. A. Bangor, P. T. Kortum, J. T. Miller, An empirical evaluation of the system usability scale. *Intl. J. Hum–Comput. Int.* **24**, 574–594 (2008).
23. P. Kellmeyer, O. Mueller, R. Feingold-Polak, S. Levy-Tzedek, Social robots in rehabilitation: A question of trust. *Sci. Robot.* **3**, eaat1587 (2018).
24. N. Dennler, C. Yunis, J. Realmuto, T. Sanger, S. Nikolaidis, M. Matarić, Personalizing user engagement dynamics in a non-verbal communication game for cerebral palsy, in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (IEEE, 2021), pp. 873–879.
25. A. Abubshait, A. Wykowska, Repetitive robot behavior impacts perception of intentionality and gaze-related attentional orienting. *Front. Robot. AI* **7**, 565825 (2020).
26. M. M. de Graaf, S. Ben Allouch, J. A. Van Dijk, What makes robots social?: A user's perspective on characteristics for social human-robot interaction, in *Social Robotics: 7th International Conference, ICSR 2015, Proceedings 7* (Springer, 2015), pp. 184–193.
27. I. Boukhennoufa, X. Zhai, V. Utti, J. Jackson, K. D. McDonald-Maier, Wearable sensors and machine learning in post-stroke rehabilitation assessment: A systematic review. *Biomed. Signal Process. Control* **71**, 103197 (2022).
28. R. Yin, D. Wang, S. Zhao, Z. Lou, G. Shen, Wearable sensors-enabled human–Machine interaction systems: From design to application. *Adv. Funct. Mater.* **31**, 2008936 (2021).
29. S. H. Lee, G. Park, D. Y. Cho, H. Y. Kim, J.-Y. Lee, S. Kim, S.-B. Park, J.-H. Shin, Comparisons between end-effector and exoskeleton rehabilitation robots regarding upper extremity function among chronic stroke patients with moderate-to-severe upper limb impairment. *Sci. Rep.* **10**, 1–8 (2020).
30. S. Balasubramanian, R. Colombo, I. Sterpi, V. Sanguineti, E. Burdet, Robotic assessment of upper limb motor function after stroke. *Am. J. Phys. Med. Rehabil.* **91**, S255–S269 (2012).
31. A. A. Blank, J. A. French, A. U. Pehlivan, M. K. O'Malley, Current trends in robot-assisted upper-limb stroke rehabilitation: Promoting patient engagement in therapy. *Curr. Phys. Med. Rehabil. Rep.* **2**, 184–195 (2014).
32. S. Y. A. Mounis, N. Z. Azlan, F. Sado, Assist-as-needed control strategy for upper-limb rehabilitation based on subject's functional ability. *Meas. Control* **52**, 1354–1361 (2019).
33. D. Eizicovits, Y. Edan, I. Tabak, S. Levy-Tzedek, Robotic gaming prototype for upper limb exercise: Effects of age and embodiment on user preferences and movement. *Restor. Neurol. Neurosci.* **36**, 261–274 (2018).
34. A. Jevtić, A. F. Valle, G. Alenyà, G. Chance, P. Caleb-Solly, S. Dogramadzi, C. Torras, Personalized robot assistant for support in dressing. *IEEE Transactions on Cognitive and Developmental Systems* **11**, 363–374 (2019).
35. N. Dennler, E. Shin, M. Matarić, S. Nikolaidis, Design and evaluation of a hair combing system using a general-purpose robotic arm, in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2021), pp. 3739–3746.
36. K. P. Hawkins, P. M. Grice, T. L. Chen, C.-H. King, C. C. Kemp, Assistive mobile manipulation for self-care tasks around the head, in *2014 IEEE Symposium on computational intelligence in robotic rehabilitation and assistive technologies (CIR2AT)* (IEEE, 2014), pp. 16–25.
37. J. Fasola, M. J. Matarić, A socially assistive robot exercise coach for the elderly. *Interaction* **2**, 3–32 (2013).
38. C. Clabaugh, K. Mahajan, S. Jain, R. Pakkar, D. Becerra, Z. Shi, E. Deng, R. Lee, G. Ragusa, M. Matarić, Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders. *Front. Robot. AI* **6**, 110 (2019).
39. M. J. Matarić, J. Eriksson, D. J. Feil-Seifer, C. J. Winstein, Socially assistive robotics for post-stroke rehabilitation. *J. Neuroeng. Rehabil.* **4**, 1–9 (2007).
40. A. L. van Ommeren, L. C. Smulders, G. B. Prange-Lasonder, J. H. Burke, P. H. Veltink, J. S. Rietman, Assistive technology for the upper extremities after stroke: Systematic review of users' needs. *JMIR Rehabil. Assist. Technol.* **5**, e10510 (2018).
41. M. Suguitan, G. Hoffman, Blossom: A handcrafted open-source robot. *ACM Trans. Hum-Robot Interact. (THRI)* **8**, 1–27 (2019).
42. Trossen Robotics, LoCoBot (2023); www.trossenrobotics.com/locobot-overview.aspx.
43. N. Taylor, K. Cheverst, P. Wright, P. Olivier, Leaving the wild: Lessons from community technology handovers, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), pp. 1549–1558.
44. R. Varghese, J. J. Kutch, N. Schweighofer, C. J. Winstein, The probability of choosing both hands depends on an interaction between motor capacity and limb-specific control in chronic stroke. *Exp. Brain Res.* **238**, 2569–2579 (2020).
45. Y. Choi, J. Gordon, H. Park, N. Schweighofer, Feasibility of the adaptive and automatic presentation of tasks (adapt) system for rehabilitation of upper extremity function post-stroke. *J. Neuroeng. Rehabil.* **8**, 1–12 (2011).

46. S. Kim, C. E. Han, B. Kim, C. J. Winstein, N. Schweighofer, Effort, success, and side of lesion determine arm choice in individuals with chronic stroke. *J. Neurophysiol.* **127**, 255–266 (2022).
47. G. Wilson, D. J. Cook, A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol. (TIST)* **11**, 1–46 (2020).
48. R. Berger, Now I see it, now I don't: Researcher's position and reflexivity in qualitative research. *Qual. Res.* **15**, 219–234 (2015).
49. A. R. Fugl-Meyer, L. Jaaskö, I. Leyman, S. Olsson, S. Steglind, The post-stroke hemiplegic patient. 1. A method for evaluation of physical performance. *Scand. J. Rehabil. Med.* **7**, 13–31 (2011).
50. M. F. Folstein, S. E. Folstein, P. R. McHugh, Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* **12**, 189–198 (1975).
51. G. Uswatte, E. Taub, Implications of the learned nonuse formulation for measuring rehabilitation outcomes: Lessons from constraint-induced movement therapy. *Rehabil. Psychol.* **50**, 34–42 (2005).
52. Kinova, Assistive solutions; www.kinovarobotics.com/sector/medical-robotics.
53. Qtrobot: Humanoid social robot for research and teaching (2020); http://luxai.com/qtrobot-for-research/.
54. M. Spitale, S. Okamoto, M. Gupta, H. Xi, M. J. Matarić, Socially assistive robots as storytellers that elicit empathy. *ACM Trans. Hum-Robot Interact.* **11**, 1–29 (2022).
55. K. Swift-Spong, E. Short, E. Wade, M. J. Matarić, Effects of comparative feedback from a socially assistive robot on self-efficacy in post-stroke rehabilitation, in *2015 IEEE International Conference on Rehabilitation Robotics (ICORR)* (IEEE, 2015), pp. 764–769.
56. R. Feingold-Polak, O. Barzel, S. Levy-Tzedek, A robot goes to rehab: A novel gamified system for long-term stroke rehabilitation using a socially assistive robot—Methodology and usability testing. *J. Neuroeng. Rehabil.* **18**, 1–18 (2021).
57. S. Cai, G. Li, X. Zhang, S. Huang, H. Zheng, K. Ma, L. Xie, Detecting compensatory movements of stroke survivors using pressure distribution data and machine learning algorithms. *J. Neuroeng. Rehabil.* **16**, 1–11 (2019).
58. H. Kallio, A.-M. Pietilä, M. Johnson, M. Kangasniemi, Systematic methodological review: Developing a framework for a qualitative semi-structured interview guide. *J. Adv. Nurs.* **72**, 2954–2965 (2016).
59. S. Elo, H. Kyngäs, The qualitative content analysis process. *J. Adv. Nurs.* **62**, 107–115 (2008).
60. H. Ando, R. Cousins, C. Young, Achieving saturation in thematic analysis: Development and refinement of a *Codebook*. *Psychology* **3**, 03–CP (2014).

Acknowledgments: We thank all participants and pilot testers who participated in our study for their time. **Funding:** Supported by a NSF Graduate Research Fellowship under the award number #DGE-1842487. This work was also partially supported by an Agilent Early Career Professor Award. **Author contributions:** N.D. designed and performed all experiments and wrote the manuscript; A.C. administered Fugl-Meyer assessment, scored the AAOUT, and provided feedback on experimental design; E.D.G. and C.C. developed code for analyzing results; C.J.W. provided the lab space, participants, and collaboratively developed the experimental protocol and paper editing; S.N. helped with analysis plans and paper editing; and M.J.M. helped design experiments and edited the paper. **Competing interests:** The authors declare that they have no competing interests to disclose. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials. The data for this study will be made available upon request in accordance with our IRB protocol.

Submitted 11 November 2022
Accepted 19 October 2023
Published 15 November 2023
10.1126/scirobotics.adf7723