Department of Veterans Affairs

Journal of Rehabilitation Research and Development
Vol. 40, No. 6, November/December 2003
Pages 487–500

Reliability and responsiveness of elbow trajectory tracking in chronic poststroke hemiparesis

Carolynn Patten, PhD, PT; Dhara Kothari, PT, MS; Jennifer Whitney, MPT; Jan Lexell, MD, PhD; Peter S. Lum, PhD

Rehabilitation Research and Development Center, Department of Veterans Affairs (VA) Palo Alto Health Care System, Palo Alto, CA; Department of Orthopaedic Surgery, Stanford University School of Medicine, Stanford, CA; Department of Rehabilitation, Lund University Hospital, Lund, Sweden; Department of Biomedical Engineering, Virginia Commonwealth University, Richmond, VA

Abstract—This study established the reliability of a novel upper-limb trajectory-tracking task for assessment of perceptual motor control in hemiparetic adults. Eleven persons with chronic poststroke hemiparesis (mean 58.6 months) and eleven nondisabled control subjects performed an elbow flexion-extension task against a low-resistance isotonic load at three speeds: 25°/s, 45°/s, and 65°/s. Both arms (paretic and nonparetic or dominant and nondominant) were tested during two identical sessions separated by 1 week. Relative reliability (intraclass correlation coefficient [ICC]) ranged from 0.5 to 0.8 and absolute reliability (standard error of measurement [SEM%]) ranged between 19% to 36% across both subject groups. No systematic errors between test sessions were revealed. Smallest real differences (SRDs) were determined to be ±2° to 3° in nondisabled, ±2° to 5° in nonparetic and ±9° in paretic arms. Responsiveness ratios derived with the use of the SRDs ranged between 1.91 to 2.45, indicating that this instrument is sensitive to clinically important change and suitable for demonstrating effects on upper-limb motor performance following clinical intervention.

Key words: motor control, outcome, rehabilitation, reliability, responsiveness, stroke.

INTRODUCTION

Following stroke recovery of functional use of the upper-limb is a primary goal of rehabilitation. Because it involves contact and manipulation of objects in the environment, the control of functional upper-limb motor tasks is complex. Perceptual motor control, also termed "perception-action coupling," emerges under the combined influences of vision, proprioception, central integration, temporal patterning, and corrective feedback mechanisms [1]. Evaluation of this genre of motor tasks is critical to the practice of neurorehabilitation; yet, appropriate instruments for their objective and quantitative measurement remain elusive.

Abbreviations: ANOVA = analysis of variance, CI = confidence interval, CNS = central nervous system, GRI = Guyatt's Responsiveness Index, ICC = intraclass correlation coefficient, MCID = minimal clinically important difference, RMS = root-mean-square, SD = standard deviation, SE = standard error, SEM = standard error of measurement, SRD = smallest real difference, VA = Department of Veterans Affairs.

This material was based on work supported by the Department of Veterans Affairs (VA), Rehabilitation Research and Development Service, merit review grant B2405R. Dr. Patten receives support through a VA Rehabilitation Research and Development Service Advanced Research Career Development Award.

Address all correspondence and requests for reprints to Rehabilitation Research and Development Center/153, VA Palo Alto Health Care System, 3801 Miranda Avenue, Palo Alto, CA 94304; 650-493-5000, ext. 63593; fax: 650-493-4919; email: patten@rrd.stanford.edu.

Clinical evaluation of upper-limb motor function typically involves ordinal instruments, such as the Functional Independence Measure (FIM) [2], the Wolf Motor Function Test [3], or the Chedoke-McMaster Stroke Assessment [4]. This type of clinical evaluation is accomplished through observation of commonplace movements and tasks, many of which require considerable interaction between the subject, object, and environment. Ordinal scales can be problematic because of poor consistency in differences between scale increments. Additionally, these measures lack sensitivity to document small, but potentially important, changes in motor function. During clinical or observational evaluation, performance of prescribed functional tasks is typically unconstrained with regard to positioning and specific details of task execution, most notably pace. Thus, qualitative scoring of task performance (i.e., success or failure of task completion, crude assessment of the amount of assistance required, use of assistive equipment, alteration in gross position, excessive time to completion) establishes the potential for each subject and/or patient to adopt a unique approach for successful task performance. Comparable scores between subjects may thus involve wholly different motor strategies.

Trajectory-tracking tasks involve components common to both perception-action coupling and functional motor tasks: perception of environmental and task constraints, motor planning, motor execution, and corrective monitoring of performance including explicit feedback. Biomechanical parameters such as joint excursion, movement speed, kinetic requirements, and task initiation can be reproduced in trajectory-tracking tasks. Therefore, quantitative comparisons between subjects and across subject groups can be made based on equivalent criteria. Moreover, because an instrumented task such as trajectory tracking yields continuous data, small changes in perceptual motor control (i.e., speed, accuracy, reciprocal activation, and kinetic or kinematic variability) can be readily detected where they might remain unapparent with the use of clinical assessments measuring at the ordinal level. A close relationship has been demonstrated between performance on functional and trajectory-tracking tasks [5]; thus a rationale is established that trajectory tracking is an appropriate surrogate measure for integrated upper-limb sensorimotor control.

Trajectory tracking clearly affords a superior approach and improved sensitivity for assessing upper-limb perceptual motor function. However, important clinimetric properties of this method, including reliability and responsiveness, have not been established. Few of the available data have reported reproducibility despite reports that various forms of trajectory-tracking tasks have been used to screen for impaired motor control and evaluate treatment outcome in persons with central nervous system (CNS) impairment [6,7]. Trajectory-tracking reliability has been reported in conjunction with experimental studies conducted in healthy individuals, but generalizing these findings to clinical populations is not possible [6]. Consequently, it remains unclear whether consistent, reproducible trajectory-tracking performance can be observed in persons with impaired motor control caused by a CNS injury. Persons with poststroke hemiparesis make up a heterogeneous population [8] (e.g., individual subject differences), characterized by high within-subject variability on repeated performance [6,9]. This characteristic of hemiparetic subjects renders measurement particularly challenging in this population and emphasizes the need to establish appropriate benchmark data for clinical or experimental inference.

The aim of this study was to determine whether an elbow trajectory-tracking task performed on a commercially available dynamometer could yield reliable assessment of impaired perceptual motor control in persons affected with poststroke hemiparesis. To address this problem, we used a simple test-retest design to evaluate the basic clinimetric properties of the trajectory-tracking task. We used an intraclass correlation analysis (ICC) to indicate relative reliability. By quantifying the measurement error and performing Bland and Altman analysis, we evaluated the absolute reliability. These parameters established a basis for deriving the smallest real difference (SRD), which was used as an objective indicator of Δ (effect size) for determining responsiveness of the elbow trajectory-tracking task. Our overriding motivation was to determine the utility of this measure for assessing clinically important change in upper-limb motor function following rehabilitation for poststroke hemiparesis.

METHODS

Subjects

The study population consisted of 11 adults with poststroke hemiparesis of greater than 12 months duration who were recruited from the local community,

including outpatient clinics at the Department of Veterans Affairs (VA) Palo Alto Health Care System. Criteria for participation included clinical presentation of a single, unilateral stroke; absence of pain or contracture in the upper limb; no more than minor impairment of upper-limb sensation or proprioception; ability to comprehend and follow three-step commands as evidenced on the Cognistat exam [10]; and demonstration of at least 70° of active elbow flexion with gravity eliminated. Eleven nondisabled adults of similar age ($t = 1.734$, $p > 0.05$) without evidence of neuromusculoskeletal or cardiac complaints served as control subjects. All subjects provided informed consent in accordance with the Declaration of Helsinki, and all aspects of this study were approved by the Stanford University panel on human subjects in medical research. Subject characteristics are enumerated in **Table 1**.

Instrumentation

We used a commercially available dynamometer (Biodex System 3.0 Pro, Shirley, New York, USA) to isolate and measure transverse plane elbow flexion and extension. The standard elbow attachment was modified with a prefabricated wrist splint and straps to accommodate persons with impaired grasp (**Figure 1**). Both hemiparetic and nondisabled subjects were tested with the modified attachment. The dynamometer was operated in isotonic mode with the sensitivity set at 5 and torque at 1 ft-lb. This combination of settings created a low-friction condition, which required minimal torque to move the apparatus. Thus, independent of strength deficits, both nondisabled and hemiparetic subjects were able to perform the task through the full range of motion. Prior to experimentation, the dynamometer was calibrated with the use of the standard procedure described in the Biodex operation manual. In addition, we manually calibrated both the position and torque channels by recording the voltages obtained over a range of positions and with a range of known weights.

Subjects were seated in the Biodex chair with the back angled at 85°, the trunk stabilized using waist and trunk straps, and the feet supported using the Biodex leg rest. A hand-held goniometer was used to position the shoulder in 5° of forward flexion and 70° of abduction, and the upper arm was stabilized with the use of an adjustable support, which balanced the weight of the limb and eliminated excess motion at the shoulder. The dynamometer axis of rotation was aligned with the medial epicondyle of the humerus, and the range of motion limit was anatomically referenced at 10° and 75° for extension and

Table 1.
Subject characteristics.

Characteristic	Subjects with Poststroke Hemiparesis (n = 11)	Nondisabled Subjects (n = 11)
Mean Age (yr) (±SD)	58.59 (±11.7)	49.19 (±13.1)
Gender		
Male	5	4
Female	6	7
Time Since Onset of Stroke (mo)	53.4	N/A
Side Affected/Dominance		
Right	9	10
Left	2	1
Mechanism of Stroke		
Ischemic	5	N/A
Hemorrhagic	6	N/A
Lesion Location		
Cortical	4	—
Subcortical	7	—
Upper-Limb Fugl Meyer Scores		
Median	42	—
Range	16–63	N/A

N/A = not applicable SD = standard deviation

Figure 1.
Tracking apparatus designed to accommodate persons with impaired grasp. Standard Biodex™ forearm attachment was fitted with a prefabricated forearm-wrist splint with a cone to fit cylindrical grasp. A range of sizes was available to fit each individual subject appropriately.

Figure 2.
Dynamometer configuration and positioning for transverse plane elbow flexion-extension task. An adjustable telescoping arm (foreground) attached to chair frame was fitted with a pad to support weight of upper arm, which provided shoulder stabilization and enabled independent elbow movement.

flexion, respectively, to provide a total excursion of 65° at the elbow. The dynamometer setup and subject positioning are illustrated in **Figure 2**.

We used custom written software to acquire data and present the test stimuli. Position and torque data were sampled from the dynamometer at 1 kHz with a 12-bit A/D converter (Keithly Instruments, Inc., Taunton, Massachusetts, USA) and written directly to disk for off-line analysis. We used a computer video display to present a position-time plot of the test criterion (**Figure 3**), which involved two cycles of reciprocal elbow flexion-extension. The trajectory-tracking task was performed at three speeds: 25°/s (slow), 45°/s (medium), and 65°/s (fast).

Protocol

Each subject participated in two identical test sessions scheduled 1 week apart at the same time of day. All subjects were tested by the same laboratory personnel (JAW and DK). Both the initial test and retest sessions for an individual subject were conducted by the same experimenter. Positioning adjustments for the dynamometer and chair were recorded on a laboratory form to aid in reproducing the subject setup for the retest session.

Following informed consent, subjects were positioned and then provided verbal instructions and opportunity for familiarization with the apparatus and trajectory-tracking task. Subjects were instructed to match a criterion trajectory as closely as possible, minimizing the difference between his or her performance and the criterion (**Figure 3**). Based on evidence that this type of task is learned rapidly [11], subjects were provided 30 practice trials structured in blocks of 10 trials at each of the three test speeds. Concurrent feedback was provided with the plot of the subject's performed trajectory displayed in a contrasting color. A numeric score reporting the root-mean-square (RMS) difference between criterion and performed trajectories was displayed on screen at the end of each practice trial for additional feedback. Following completion of the compulsory practice trials, five test trials were recorded at each speed. Test trials were ordered from slow to fast speeds and were performed without the numeric RMS error score feedback. All subjects performed the task with both arms. The nonparetic arm was tested first in hemiparetic subjects, and the nondominant arm was tested first in nondisabled

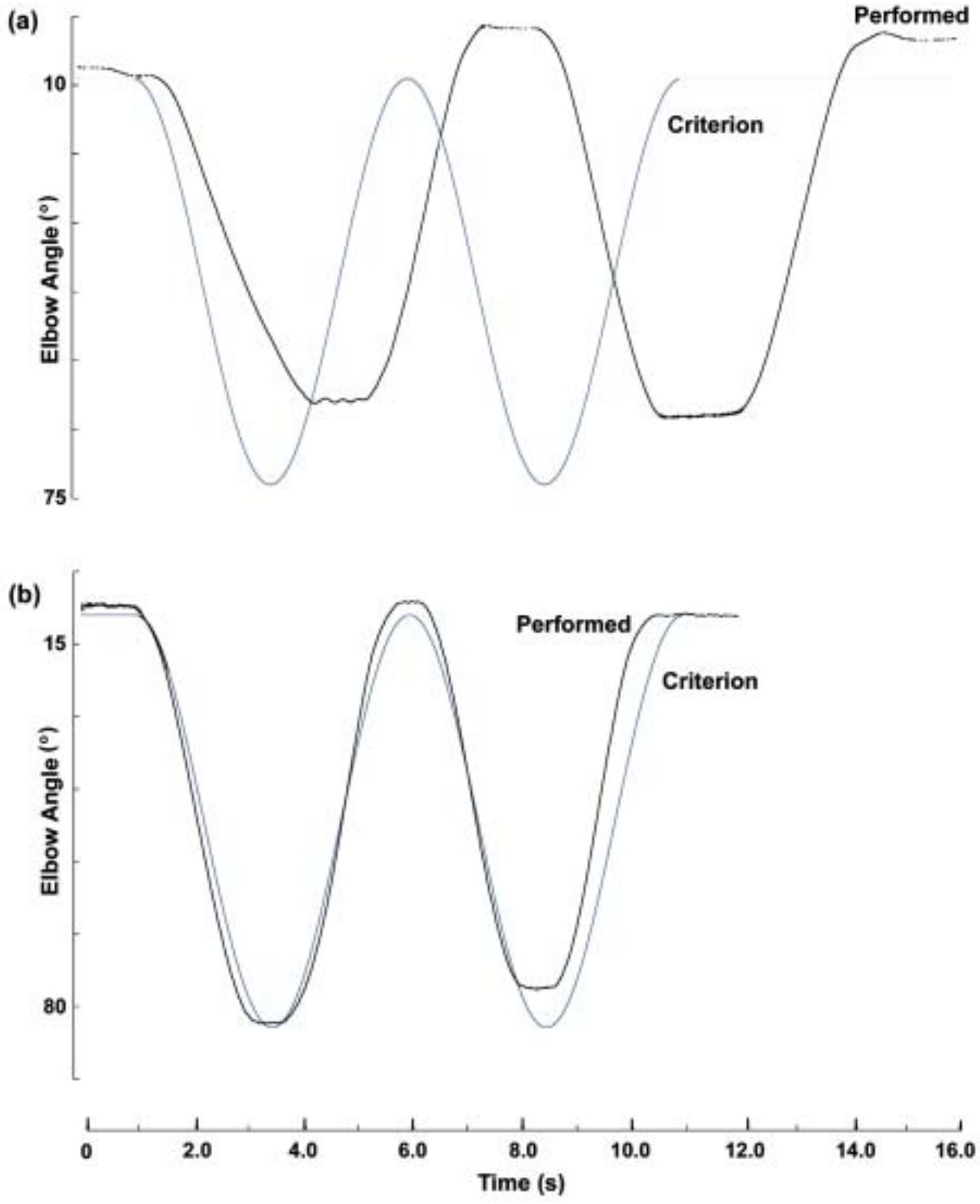

Figure 3.
Criterion trajectory and representative performance at medium speed (45° s^{-1}) for **(a)** paretic and **(b)** nonparetic arms. Movement was initiated from flexion and involved two cycles of reciprocal extension and flexion. Performance in nonparetic arm is relatively accurate and accompanying root-mean-square (RMS) error would be moderately low. Because both amplitude and temporal patterning of movement are significantly disrupted in paretic arm, RMS error would be quite large.

subjects. The entire procedure, including practice trials, was repeated in the retest session 1 week later at the same time of day.

Data and Statistical Analysis

We derived reliability and responsiveness statistics using the mean of the five individual test trial RMS error scores. To assure normal data distribution for analysis with parametric statistics, we log-transformed the raw RMS error scores. Relative reliability was determined through the calculation of the ICC [12]. Absolute reliability was established with the use of Bland and Altman analysis and the quantification of measurement error (within subject variance) through the assessment of both the standard error of measurement (SEM) and SEM% [13,14]. The SRD was derived on the basis of the measurement error, and responsiveness was assessed with the calculation of the Guyatt Responsiveness Index (GRI) [15,16].

Relative Reliability

The $ICC_{1,1}$ was calculated with the use of a one-factor (test session) repeated-measures analysis of variance (ANOVA) model. If BMS represents intersubject variability, WMS the variability within subjects, and k the number of iterations, then

$$ICC_{1,1} = \frac{BMS \angle WMS}{BMS + (k \angle 1)WMS} \quad .$$

Absolute Reliability

We constructed Bland-Altman plots by plotting the between session difference in mean RMS error score (test-retest) versus the mean of the test and retest RMS error scores [13]. From the Bland-Altman plots, the data were examined for their magnitude, range, and distribution around the zero line. We calculated and overlaid 95% confidence intervals (CIs) on the Bland-Altman plots to identify any systematic trends between repeated measurements indicative of task proficiency, session-specific effects, placebo effects, or task familiarization. The 95% CIs were derived as

$$CI = \bar{d} \pm 2.26\ (SE) \ ,$$

where $\bar{d}$ = the mean difference between the test and retest scores, and

$$SE\ (\text{standard error of } \bar{d}) = \frac{SD_{diff}}{\sqrt{n}} \ ,$$

with n = number of observations and SD_{diff} = standard deviation of the differences between test and retest scores.

We determined measurement error by calculating the SEM (SEM = WMS, WMS obtained from ANOVA as just described). Measurement error reports the variability in repeated trials within the same individual. Such variability may result from performance differences or nonspecific sources of error (i.e., the instrument, the experimental paradigm). For comparison across samples and experimental conditions, measurement error was also expressed as the SEM% (SEM% = (SEM/mean) × 100) to produce a unitless indicator of error magnitude.

Responsiveness

An instrument's responsiveness is directly related to the magnitude of change in subject scores, constituting a clinically important difference. As measurement error increases, larger treatment effects are necessary to demonstrate efficacy. Responsiveness was determined in two steps. First, we quantified the SRD using the within subject variability and method error:

$$SRD = 1.96\sqrt{2}(SEM) \ ,$$

where 1.96 is used to construct the 95% CI, $\sqrt{2}$ is used to account for the variance of two measurements, and SEM is the standard error of the measurement as described earlier.

The SRD is expressed in the same units as the measurement of interest and represents the smallest change that can be interpreted as a real difference, which exceeds measurement noise. In the second step, we calculated the GRI using the SRD as the numerator and the standard deviation (SD) of test-retest differences (SD_{diff}) as the denominator [15–17] as follows:

$$GRI = \frac{SRD}{SD_{diff}} \quad .$$

RESULTS

Figure 4 illustrates the mean RMS error by subject group, side, and movement speed. Across all task speeds, RMS error was generally low in the nondisabled adults, averaging 3.4°. No statistically or physiologically significant differences in trajectory-tracking performance were apparent between the dominant and nondominant sides of the nondisabled subjects ($F(1, 126) = 0.213$, $p = 0.65$). However, a nonmonotonic function was demonstrated in

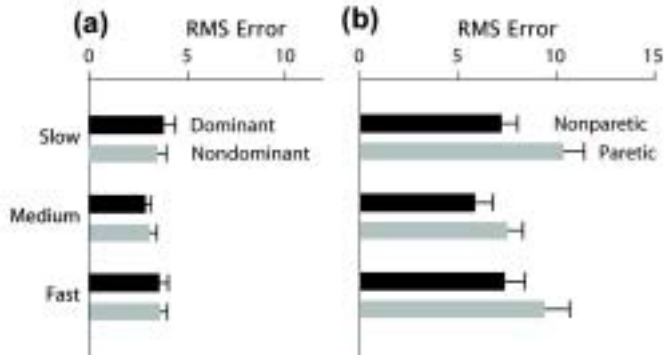

Figure 4.
Mean root-mean-square (RMS) error (±SD) by group, movement speed, and side. A nonmonotonic function, in which lowest RMS error was produced at medium speed (45°/s), was demonstrated in **(a)** nondisabled and **(b)** hemiparetic subject groups and both arms. Similarity of this pattern independent of performance group (i.e., nondisabled, paretic arm, nonparetic arm) suggests that while motor execution is clearly impaired in poststroke hemiparesis, general structure of perceptual-motor pattern is retained. Greater RMS error is characteristic of motor performance in hemiparesis. This type of impairment has led to question whether reliability can be demonstrated in this clinical population.

which error scores were significantly lower at the medium speed relative to the slow and fast speeds ($F(2, 126) = 2.97$, $p = 0.05$).

Across all conditions, the hemiparetic subjects demonstrated greater RMS error relative to nondisabled controls ($F(3, 120) = 39.88$, $p < 0.0001$) (**Figure 4(b)**). RMS error observed in the nonparetic arm was markedly greater (200%) than in nondisabled control subjects while the paretic arm RMS error exceeded the nonparetic arm by 134 percent ($F(1, 126) = 2.99$, $p < 0.0001$). Similar to observations in the nondisabled subjects, a nonmonotonic function, in which RMS error at the medium speed was significantly lower relative to both the slow and fast speeds, was demonstrated in both the paretic and nonparetic arms. Although the absolute magnitude of trajectory-tracking error varied across physiologic conditions (subject group and side), this nonmonotonic function is RMS error by target speed was observed consistently across all subject groups.

Relative Reliability

ICC values for RMS error are reported in **Table 2**. Across groups, sides, and movement speed, the ICC ranged between 0.51 to 0.80. ICC values were generally similar in the nonparetic, nondominant, and dominant arms, while in the paretic arm, ICC values were somewhat lower, ranging between 0.51 and 0.59.

Absolute Reliability

A Bland-Altman plot of test-retest differences in RMS error scores is represented in **Figure 5**. Test-retest differences demonstrated more variability in the hemiparetic as compared to nondisabled subjects. However, because zero fell within the 95% CI in all cases, no evidence of systematic bias between test sessions was revealed.

Measurement error reported as the SEM% ranged between 19 and 36 percent across both subject groups and all combinations of side and movement speed. These data are reported in **Table 3**. Across experimental groups and conditions, no consistent or systematic pattern was demonstrated in the magnitude of measurement error. Statistically, measurement error was similar between the hemiparetic and nondisabled subjects ($p > 0.05$).

Smallest Real Difference

SRD scores are reported in **Table 4**. The SRD represents a range, or error band, centered around the mean of the test-retest difference for any given experimental condition. The magnitude of the SRD is determined by the measurement error specific to each group and condition. In the nondisabled subjects, SRDs were similar between the nondominant (2.11 ± 0.7) and dominant (2.52 ± 0.7) sides. However, in persons with hemiparesis, SRDs differed considerably between the nonparetic (4.82 ± 0.8) and paretic (7.46 ± 2.2) sides and were markedly greater than in nondisabled persons. No systematic effects in the magnitude of the SRD were demonstrated on the basis of movement speed.

Table 2.
$ICC_{1,1}$ values.

Speed	Subjects with Poststroke Hemiparesis		Nondisabled Subjects	
	Paretic Arm	Nonparetic Arm	Nondominant Arm	Dominant Arm
Slow	0.51	0.77	0.56	0.79
Medium	0.59	0.68	0.80	0.71
Fast	0.52	0.75	0.71	0.74

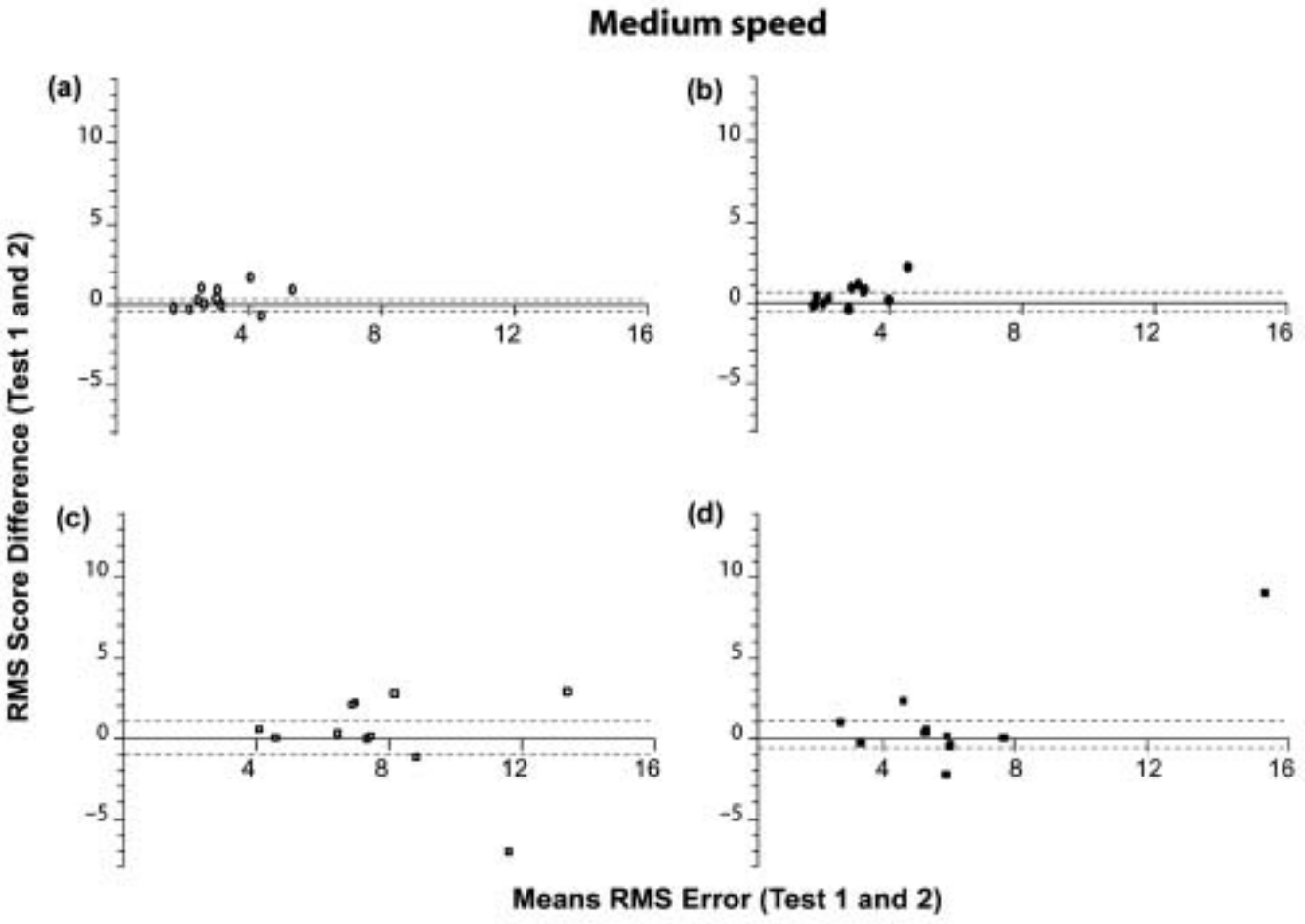

Figure 5.
Representative Bland-Altman plot of **(a)** nondominant, **(b)** dominant, **(c)** paretic, and **(d)** nonparetic arms. Test-retest difference plotted versus mean of test and retest score. Data from only medium speed (45°/s) was plotted for both subject groups and arms to illustrate differences in data distribution between subject groups. Dashed lines in all plots demarcate 95% confidence intervals (CIs). Positive values indicate lower root-mean-square (RMS) error on retest while negative values indicate lower RMS error on initial test session. Because zero is included in all 95% CIs, no systematic trends are present.

Responsiveness

Values for the GRI are presented in **Table 4**. Using the SRD derived from the measurement error to indicate clinically important change (Δ), the GRI ranged between 1.89 and 2.08 in the hemiparetic subjects and between 1.91 and 2.45 in the nondisabled subjects.

DISCUSSION

We present evidence of reliable performance on a novel task for assessing upper-limb perceptual motor control. We demonstrated reliability both in persons with poststroke hemiparesis and in nondisabled persons. In addition, we demonstrate the feasibility of modifying a commercially available dynamometer for this type of assessment and report benchmark performance values. We extended these fundamental observations to determine the magnitude of clinically important effects through quantification of the SRD. Estimated SRDs can be used in future studies for inference regarding whether genuine differences in upper-limb motor control occur in response to clinical rehabilitation. A final product of this sequence of statistical procedures was the calculation of the responsiveness ratio and estimation of sample size to be used in future clinical trials investigating the efficacy

Table 3.
Results—Measurement error.

Speed	Metric	Subjects with Poststroke Hemiparesis		Nondisabled Subjects	
		Paretic Arm	Nonparetic Arm	Nondominant Arm	Dominant Arm
Slow					
	$\bar{d}$	0.892	0.184	0.28	0.508
	95% CI	–0.42 to 2.20	–0.52 to 0.88	–0.69 to 1.25	–0.46 to 1.48
	SEM%	27.61	19.86	27.25	33.14
Medium					
	$\bar{d}$	0.152	0.49	0.21	0.30
	95% CI	–0.75 to 1.05	–0.48 to 1.46	–0.28 to 0.69	–0.21 to 0.81
	SEM%	24.17	34.83	19.37	20.86
Fast					
	$\bar{d}$	0.812	0.341	0.10	0.40
	95% CI	–0.79 to 2.41	–0.95 to 1.63	–0.68 to 0.79	–0.33 to 1.12
	SEM%	36.27	24.05	19.50	24.85

CI = confidence interval
SEM = standard error of measurement

Table 4.
Results—Smallest real differences (SRDs) and Guyatt Responsiveness Index (GRI).

Responsiveness Measures	Speed	Subjects with Poststroke Hemiparesis		Nondisabled Subjects	
		Paretic Arm	Nonparetic Arm	Nondominant Arm	Dominant Arm
SRD					
	Slow	±7.93	±3.94	±2.86	±3.24
	Medium	±5.02	±5.60	±1.54	±1.82
	Fast	±9.42	±4.91	±1.93	±2.51
GRI					
	Slow	1.98	2.08	2.07	2.28
	Medium	1.99	1.89	2.45	2.18
	Fast	1.96	1.99	2.37	1.91

of upper-limb rehabilitation for persons with poststroke hemiparesis.

Reliability of Motor Performance in Hemiparetic Adults

Few data are available in the literature reporting either motor performance or reliability of motor performance of any type in persons with poststroke hemiparesis [18–24]. We know of only one study reporting reliability of trajectory tracking in the upper limb, which involved finger flexion and extension [6]. Such observations are insufficient to support our current line of investigation because the tasks, joints, and subject populations differ substantially and thus provide no basis for generalization to our population of interest: persons with poststroke hemiparesis.

Because persons with poststroke hemiparesis are commonly described as a markedly heterogeneous population, obtaining consistent, reproducible measurements for assessing clinically important change has been problematic [25]. Much of the inconsistency in this group is ascribed to individual subject differences and these are typically attributed to variations in the severity, location, and mechanism of the cerebrovascular lesion [26]. The primary motivation for the present study was to establish reasonable expectations for reproducibility of upper-limb motor performance in a group of hemiparetic individuals. Our intention is to incorporate the protocol described

herein as one component of a suite of evaluation procedures for documenting recovery of motor function in hemiparetic adults. Accordingly, the current study was designed to emulate conditions under which the trajectory-tracking task would realistically be administered, and we chose to focus on assessment of its clinimetric properties in this context. This perspective differs somewhat from the development of a unique stand-alone test and exploration of optimal parameters for its administration. Indeed, our foci were first to determine whether reliability could be established using only five trials per condition and second to determine the requisite sample size for demonstrating clinically important effects with few data. Consistent with observations reported in the literature [9], regardless of side (paretic, nonparetic), the hemiparetic subjects were less accurate and more variable on repeated trials of the trajectory-tracking task as compared to nondisabled subjects. However, despite these marked impairments of perceptual motor control, both relative reliability, as reported by the ICC, and absolute reliability, as reported by the Bland and Altman analysis, were only slightly lower in the hemiparetic subjects as compared to nondisabled controls.

While reliability refers to the consistency of measurements, reliability analyses can focus on different aspects of the measurement problem: consistency across repeated trials within a given session (intrasession), variability between subjects, or repeated measurements across sessions (intersession). Especially in clinical populations, intrasession and intersession reliability are of great concern. Because our goal is to determine the suitability of this trajectory-tracking task for use as a clinical outcome measure, we focused the present study on intersession reliability and attendant issues of measurement stability across multiple test sessions. An issue related to the study of intersession reliability is the interval between test sessions, which varies considerably in published studies [14]. The length of the intersession interval was a less critical detail in this design because our subject sample was composed of chronic hemiparetic persons in whom clinically significant changes across test sessions were not expected. Thus, our decision to separate test and retest sessions by 1 week stemmed from several other methodological goals: minimizing systematic effects on performance such as fatigue or learning, avoiding physiological influences such as circadian rhythms, and establishing realistic measurement intervals such as those that may occur in a clinical setting.

Relative Reliability

Over the last several years, using ICC for reliability assessment in clinical research has become customary [16]. ICCs observed in all subject groups in the present study encompassed a range from 0.51 to 0.80. Shrout and Fleiss term reliability "fair" for ICCs in the range of 0.5 to 0.6, "good" in the range of 0.6 to 0.7, and "excellent" if ICCs exceed 0.75 [27]. Based on these criteria, reliability was fair to good in the hemiparetic group and good to excellent overall in the nondisabled adults. However, this statistic can be misleading if used solely to indicate reproducibility. As is the case with any correlation coefficient, the magnitude of the ICC is highly sensitive to both the heterogeneity of observations between subjects and the degree to which the rank order of subjects is replicated on retest. Moreover, correlation documents the strength of association rather than agreement between repeated observations. Because of these properties of correlation, if substantial between-subject differences are present, an instrument can produce a high ICC but reveal little information regarding agreement between repeated measurements. Accordingly, the ICC should be interpreted with caution.

One interpretation of the ICCs obtained in the present study might suggest that reproducibility of our trajectory-tracking measurements was imperfect. It is more likely, however, that our observations were narrowly distributed; thus individual subjects' data may have overlapped considerably, allowing observations to be reordered on retest. An additional source limiting the heterogeneity of the data might be that our sample was composed of hemiparetic persons with reasonably similar characteristics. Much of the existing clinical literature studying hemiparetic persons reports a vast range of deficits and functional abilities in their study samples. Accordingly, applying study findings to clinical practice or to any specific subsample of hemiparetic persons has often not been possible. In contrast to this customary practice of including subjects solely on the basis of having experienced a stroke, the present study included only hemiparetic persons meeting clearly predefined criteria. Moreover, because these subjects represent hemiparetic persons who qualify for and seek rehabilitative therapy, these current observations of reliable motor performance are generalizeable to a relevant clinical population.

Absolute Reliability

An important adjunct to analyzing the ICC is quantification of measurement error or reproducibility "noise." Measurement error reports the inherent methodological variability that is attributable to nonspecific sources (i.e., instrumentation, measurement, regression to the mean) rather than true biologic change (e.g., adaptation) and is obtained through the assessment of the variability of repeated measurements within the same subject. This error more directly indicates reproducibility than ICC and represents the aspect of reproducibility most relevant to longitudinal measurement in clinical settings. In contrast to ICC, absolute reliability is not influenced by the range of observations across subjects, and thus, irrespective of where a particular individual's performance ranks among the sample, it indicates the variability in repeated testing. Bland and Altman analysis has emerged as the statistical analysis of choice for assessing absolute reliability, especially in clinical research [12]. This contemporary approach complements ICC analysis and provides detail regarding the nature of the observed intrasubject variability. In the present study, Bland-Altman plots revealed somewhat larger absolute error scores and test-retest differences and a different distribution of the data in the hemiparetic versus control subjects. However, the 95% CIs for both groups and all conditions included zero. Thus, no systematic trends were demonstrated in the data leading to the conclusion that, beyond this quantifiable level of measurement error, performance on the trajectory-tracking task is reproducible. Importantly, measurement error was comparable between the subject groups. Assessment of absolute reliability demonstrates therefore that the performance differences observed between sessions result from nonspecific effects associated with experimental replication. Taken together, these findings clearly indicate reproducible data in a group of hemiparetic persons with moderately severe motor impairment.

Measuring Clinically Important Change

In clinical research, establishing reliability is important not only when instruments are used for repeated measurement to assure their stability but also especially when the investigative goal is to detect change over time. Central to this assessment are two closely related properties of assessment tools [15]: (1) responsiveness—the capacity and sensitivity of an instrument to detect a genuine change in subject performance and (2) the definition of minimally important, or clinically important, change. Responsiveness is the ratio of the treatment effect to the variability in individual subject response [16]. As measurement error increases, observing larger treatment effects becomes necessary. Because of this interrelationship of measurement parameters, an instrument may be reliable but unresponsive to change [15,16]. Responsiveness is thus an indicator of an instrument's usefulness, defined as its capacity to detect clinically important differences, and a more responsive instrument requires a smaller sample size to detect change [16].

Frequently encountered limitations in clinical measurement are lack of a single agreed-upon standard criterion of change and lack of information regarding the magnitude of change constituting a genuine performance difference [17,28]. For this issue to be addressed, the term "minimal clinically important difference" (MCID) has been found in literature with increasing frequency [15–17,28,29]. MCIDs evolve from clinical judgment regarding the measurement properties of an instrument, the patient population, and the magnitude of change deemed "minimally important" by the practitioner. The present study involved a wholly new task and paradigm and measurement in a population from whom few data are available. Thus, no basis exists from which to construct an MCID.

A more recent construct for addressing the problem of assessing change over time is the SRD [15]. To express the uncertainty caused by measurement error, one can construct an interval (or error band) surrounding the true score. The size of this error band is termed "SRD" and represents the minimal difference necessary to infer that genuine change, exceeding measurement error, has occurred. The MCID and SRD are not synonymous. Rather than relying on clinical judgment as when defining an MCID, the SRD is a clinimetric property of an instrument that incorporates attributes of the experimental paradigm, measurements obtained with the specific instrument, and any idiosyncratic limitations of the instrumentation. In contrast, the MCID relies on clinical judgment for definition, and importantly, if the established MCID is less than the SRD, the instrument is not valid for that particular assessment.

While measurement error in the present study appeared to be only slightly greater in the hemiparetic subjects and significant differences between the paretic and nonparetic arms were not apparent, multiplication of this parameter by $\sqrt{2}$ to account for the variances of two measurements amplifies any differences and clearly illustrates that even slightly greater variability between

measurements creates a markedly greater error band surrounding an observed score. Thus, one must observe larger effects to conclude that change has occurred. This point is well illustrated by observations from the present study. SRDs observed in the nondisabled subjects were small, on the order of ±2° to 3° RMS error, but were incrementally greater in both the nonparetic and paretic arms.

Once the responsiveness of an instrument is known, estimating the requisite sample size for experiments documenting change over time is possible [16]. As initially reported, the GRI proposed use of Δ, a difference based on clinical or experimental judgment, as the numerator. Beckerman et al. have proposed using the SRD as the numerator in this ratio, because it is a quantity based on the experimenter's paradigm and the specific subject population [15]. The SRD thus affords a critical link between reliability and responsiveness. With the numerator set equal to the SRD, responsiveness is by definition 1.96. An alternate approach to deriving the GRI uses the SD_{diff} of stable subjects as the denominator [17]. Because the subject population in the present study was composed of chronic hemiparetic adults in whom stable motor function can be expected, we used this alternate approach, which produced responsiveness ratios in the range of 1.91 to 2.4. Our findings are consistent with Beckerman et al.'s maxim that using the experimentally derived SRD to indicate clinically important change yields good responsiveness and requires modest sample sizes [16]. Based on Guyatt et al.'s estimates [16], clinically important change in elbow trajectory tracking apparently can be captured with the use of fewer than 10 subjects per group. Similar responsiveness ratios observed across all conditions suggest the motor task was appropriate and the instrument afforded consistent performance conditions for both nondisabled and neurologically impaired persons. Finally, use of the SRD to derive responsiveness illustrates the effects of both reproducibility and detecting clinical change. In many clinical populations, reduction in the SRD will be elusive; thus improvements in responsiveness can be attained only by decreasing measurement error. This objective emphasizes that the reliability of interest in longitudinal studies is the within subjects variability.

Trajectory Tracking

Increased interest in understanding the nature of motor dysfunction in hemiparetic adults has motivated the search for appropriate quantitative indicators of motor performance. Various trajectory-tracking tasks have been successfully used to differentiate impaired dexterity and strength and to demonstrate that hemiparesis primarily affects mechanisms of motor control and execution while mechanisms for motor learning remain relatively intact [30,9]. Findings from the present study add to this evolving body of evidence. While observation of increased RMS error suggests motor execution is significantly impaired in poststroke hemiparesis, the consistent observation of a nonmonotonic relationship in RMS error independent of subject group and experimental condition suggests the general structure of motor control involving perception and central integration remain functional in the nervous system damaged by cerebrovascular accident.

While it can be inferred that the elbow trajectory-tracking task investigated in this study requires pyramidal function, it does not directly address the critical contribution of hand dysfunction to overall upper-limb impairment in hemiparesis. Accordingly, a limitation of our work at present is the need to establish concurrent validity of the trajectory-tracking task through corroboration of performance on familiar clinical indicators of upper-limb motor impairment [2–4]. Efforts in future work can build on the basis of reliability reported herein as we continue to investigate the potential use of the trajectory-tracking task as an outcome measure.

CONCLUSION

Interpretation of results of clinical trials depends on the choice of the primary outcome measure, the use of which depends on its clinimetric properties, including the reproducibility of measurements and responsiveness to genuine clinical or physiological change. Various types of outcome tools have commonly been used in clinical research without examination of their clinimetric properties, and perhaps not surprisingly, results of clinical research studies frequently fail to demonstrate statistically significant differences or treatment effects of sufficient magnitude to draw positive conclusions regarding the efficacy of clinical intervention. Failure to understand and incorporate these critical aspects of measurement into study design can compromise the success of demonstrating positive effects of clinical intervention.

ACKNOWLEDGMENTS

At the time of this study, Ms. Whitney was a physical therapy intern from the California State University Long Beach. We thank Dr. Anna Maria Holmback of the Department of Physiotherapy, Lund University, Lund, Sweden, for her generous assistance with the reliability analysis. Eric Topp provided assistance with graphical illustrations.

REFERENCES

1. Turvey MT. Coordination. Am Psychol 1990;45(8):938–53.
2. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. Adv Clin Rehabil 1987;1:6–18.
3. Morris DM, Uswatte G, Crago JE, Cook EW, Taub E. The reliability of the Wolf Motor Function Test for assessing upper extremity function after stroke. Arch Phys Med Rehabil 2001;82(6):750–55.
4. Gowland C, Stratford P, Ward M, Moreland J, Torresin W, Van Hullenaar S, et al. Measuring physical impairment and disability with the Chedoke-McMaster Stroke Assessment. Stroke 1993;24(1):58–63.
5. Knight JL. Manual control and tracking. In: Salvendy G, editor. Handbook of human factors. New York: Wiley and Sons; 1987. p. 183–218.
6. Halaney ME, Carey JR. Tracking ability of hemiparetic and healthy subjects. Phys Ther 1990;69(5):342–48.
7. Carey JR, Bogard CL, Youdas JW, Suman VJ. Stimulus-response compatibility effects in a manual tracking task. Percept Mot Skills 1995;81(3 Pt 2):1155–70.
8. Gowland C, deBruin H, Basmajian JV, Plews N, Burcea I. Agonist and antagonist activity during voluntary upper-limb movement in patients with stroke. Phys Ther 1992; 72(9):9–14.
9. Winstein CJ, Merians AS, Sullivan KJ. Motor learning after unilateral brain damage. Neuropsychologia 1999;37(8): 975–87.
10. Mysiw WJ, Beegan JG, Gatens PF. Prospective cognitive assessment of stroke patients before inpatient rehabilitation. Am J Phys Med Rehabil 1989;68(4):168–71.
11. Vander Linden DW, Cauraugh JH, Greene TA. The effect of frequency of kinetic feedback on learning an isometric force production task in nondisabled subjects. Phys Ther 1993;73(2):79–87.
12. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil 1998;12(3):187–99.
13. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8(2): 135–60.
14. Holmback AM, Porter MM, Downham D, Lexell J. Reliability of isokinetic ankle dorsiflexor strength measurements in healthy young men and women. Scand J Rehabil Med 1999;31(4):229–39.
15. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001;10(7):571–78.
16. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987;40(2):171–78.
17. Stratford PW, Binkley FM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. Phys Ther 1996;76(10):1109–23.
18. Roberts L, Counsell C. Assessment of clinical outcomes in acute stroke trials. Stroke 1998;29(5):986–91.
19. Boissy P, Bourbonnais D, Carlotti MM, Gravel D, Arsenault BA. Maximal grip force in chronic stroke subjects and its relationship to global upper extremity function. Clin Rehabil 1999;13(4):354–62.
20. Goldie PA, Matyas TA, Evans OM, Galea M, Bach TM. Maximum voluntary weight-bearing by the affected and unaffected legs in standing following stroke. Clin Biomech (Bristol, Avon) 1996;11(6):333–42.
21. Hsu AL, Tang PF, Jan MH. Test-retest reliability of isokinetic muscle strength of the lower extremities in patients with stroke. Arch Phys Med Rehabil 2002;83(8):1130–37.
22. Pohl PS, Startzell JK, Duncan PW, Wallace D. Reliability of lower extremity isokinetic strength testing in adults with stroke. Clin Rehabil 2000;14(6):601–7.
23. Eng JJ, Chu KS. Reliability and comparison of weight-bearing ability during standing tasks for individuals with chronic stroke. Arch Phys Med Rehabil 2002;83(8):1138–44.
24. Eng JJ, Kim CM, Macintyre DL. Reliability of lower extremity strength measures in persons with chronic stroke. Arch Phys Med Rehabil 2002;83(3):322–28.
25. Gowland C, deBruin H, Basmajian JV, Plews N, Burcea I. Agonist and antagonist activity during voluntary upper-limb movement in patients with stroke [see comments]. Phys Ther 1992;72(9):624–33.
26. Sunderland A, Bowers MP, Sluman SM, Wilcock DJ, Ardron ME. Impaired dexterity of the ipsilateral hand after stroke and the relationship to cognitive deficit. Stroke 1999;30(5):949–55.
27. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psych Bull 1979;86:420–28.
28. Nitschke JE, McMeeken JM, Burry HC, Matyas TA. When is a change a genuine change? A clinically meaningful

interpretation of grip strength measurements in healthy and disabled women. J Hand Ther 1999;12(1):25–30.

29. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. Control Clin Trials 1991;12(4 Suppl): 142S–58S.

30. Ada L, O'Dwyer NJ, Green J, Yeo W, Neilson P. The nature of the loss of strength and dexterity in the upper limb following stroke. Hum Mov Sci 1996;15:671–87.

Submitted for publication September 16, 2002. Accepted in revised form April 1, 2003.