Journal of Neurology, Neurosurgery, and Psychiatry 1994;57:326–332

Accuracy, reproducibility, and variability of hand-held dynamometry in motor neuron disease

Ajith Goonetilleke, Hamid Modarres-Sadeghi, Roberto J Guiloff

Abstract

A spring-loaded device that "breaks" at preset forces was used to assess readings obtained by hand-held dynamometry by three raters with varying experience in the method. Overall accuracy (3%), but not reproducibility or variability, was improved by greater experience. Readings obtained jointly by three raters had 53% greater variability than those obtained by a single rater. Nine muscle groups in 19 patients with motor neuron disease were assessed at 10 sessions (three replications per session) over six days by the experienced rater. Muscle force was expressed relative to that of 22 matched normal controls. The reproducibility was good with a mean % difference of 13·2 and repeatability coefficient of 2·17 kg-force for readings six days apart; the overall correlation coefficient was 0·98. The mean coefficient of variation (CV) of 10 readings was 9·9%. The poorer reproducibility and greater variability seen in clinically weaker muscles may account for differences in patients with bulbar palsy and classical amyotrophic lateral sclerosis; the degree of spasticity had no effect. The rater was estimated to contribute 37% of the total variability when testing patients. The use of a composite score by combining normalised dynamometry readings of eight limb muscles improved mean % difference to 6·7 and mean CV to 5·8%. The reproducibility and variability of hand-held dynamometry readings obtained by a single rater compare well with those of fixed devices. Readings from single raters, irrespective of experience, have similar reproducibility and variability. If, however, multiple raters are used in longitudinal assessments of individual patients, as occurs in clinical trials, the variability of their combined readings should be estimated when calculating the sample size required.

(*J Neurol Neurosurg Psychiatry* 1994;57:326–332)

Regional Neurosciences Centre, Charing Cross Hospital, London, UK
A Goonetilleke
H Modarres-Sadeghi
R J Guiloff

Correspondence to:
Roberto J Guiloff, Neuromuscular Unit, Charing Cross Hospital, Fulham Palace Road, London W6 8RF, UK.

Received 25 June 1992 and in revised form 21 July 1993.
Accepted 21 July 1993

Assessments of muscle strength are helpful in giving a topographical distribution of weakness and in monitoring progression of disease. The need for a more objective and accurate method of strength assessment than manual muscle testing has long been accepted. Children with polio were often classified as "normal" by manual muscle testing, when in fact dynamometry revealed that they were about 50% of normal.[1] Lovett and Martin[2] reported on a spring-balance mechanism designed to assess muscle force. Since then there has been a proliferation of fixed and portable devices.[1 3]

Fixed dynamometers produce highly reproducible readings but can be inconvenient to use in disabled patients. Hand-held devices overcome this problem and are widely used;[4] they have been shown to give reproducible results in normal adults and children,[5–8] and in patients with various disorders.[9–13] Sources of variability in such readings include: (1) transducer readings, (2) inter-rater, (3) intrarater between sessions and for different forces tested, (4) interpatient, (5) intrapatient between sessions and for different forces and muscle groups tested.

We are unaware of previous studies on the accuracy of hand-held dynamometry. Many studies have analysed the degree of inter- and intrarater variability, including between muscles.[8–14] The relative contributions of raters and patients, and the effects of weakness and tone, on reproducibility and variability have not been systematically studied.

An experiment in two stages was planned. Firstly, a spring-loaded mechanism that would "break" at preset forces was used. Three raters, with varying levels of experience in dynamometry, obtained readings from four different forces. The accuracy, reproducibility, and variability of these readings were studied. The effects on variability of obtaining readings by a single rater were compared with those obtained jointly by three raters.

The second experiment involved testing nine muscle groups in 19 patients with motor neuron disease. The effects of using maximum, median, or mean of three replications per session on reproducibility and variability were analysed, as were the effects of weakness and tone, and the use of composite scores.

The contributions of different factors to reproducibility and variability in patients were then estimated by combining the results of both experiments.

Methods

APPARATUS

Two hand-held electronic dynamometers (Penny and Giles Instrumentation Ltd, Christchurch, Dorset, England) were used. The devices had a range of 0–30 and 0–60 kg-force (1% error rates); the electronic

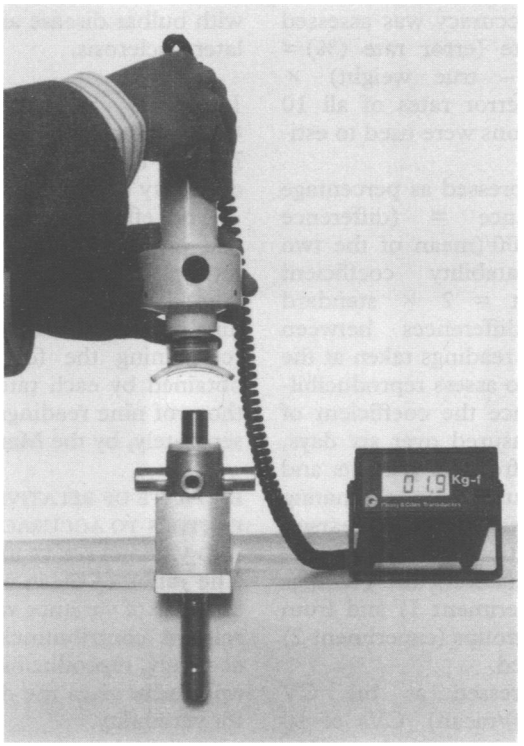

Figure Spring loaded device used in experiment 1.

transducers had a variability of 0·5% (manufacturer's figures).

For the first experiment, a spring-loaded device that would "break" (give way) at preset forces was constructed (figure). It was securely fixed to a rigid work surface at a height of 0·75 m. An engineer calibrated the device to break at 6·3, 9·3, 14·0, and 16·4 kg-force.

RATERS

Three male raters were blinded to the preset forces. Rater 1, aged 33, had one month's experience of dynamometry. Rater 2, aged 44, had used it for 2 years. Rater 3, aged 25, had only received instruction in the technique before the experiment. Ten replications were performed at each force level in each of two sessions (separated by three hours).

SUBJECTS

For the second experiment, rater 1 performed dynamometry on 19 patients with motor neuron disease (mean age 57 (range 30–80) years); the criteria for this diagnosis have been published before.[15] Six had progressive bulbar palsy (three men, three women; mean age 64), and 13 had classical amyotrophic lateral sclerosis (nine men, four women; mean age 53). Mean duration of illness was 17 (range 2–42) months.

Twenty two age, sex, height, and weight matched normal subjects (13 men, nine women; mean age 54 (range 26–75) years) were used to obtain control values for the muscles tested in patients.

MEASUREMENTS

For controls isometric dynamometry was performed with a "break" method (reading taken at the moment that the subject's force was overcome). The subject increased the force of contraction to a maximum over a period of about five seconds. Three such replications were performed at one assessment with a 10–15 s interval between each contraction. A total of nine muscle groups was tested, with standard positions (table 1); testing positions for some groups (elbow flexion/extension, hip flexion) were modified for ease and speed of assessments. Both sides of the body were tested in turn for each muscle group, the order of testing being kept the same.

For patients, the procedure used for controls was repeated on 10 separate occasions over a one week period. The first five assessments were performed on successive days, and the last five over a single 24 hour period.

The mean of the dynamometry readings obtained for each muscle from the 22 control subjects was then used to transform the readings from subjects with motor neuron disease into a % normal value. This value was used in the construction of composite scores as follows (see table 1 for abbreviations): upper limb (UL) = (SA + EF + EE + FE + IFA)/5, lower limb (LL) = (HF + KE + FD)/3, global limb = (SA + EF + EE + FE + IFA + HF + KE + FD)/8. Composite scores were thus obtained for 10 assessments.

Muscles were also assessed with the Medical Research Council scale; grade 4 was split into three subgrades (4–, 4, and 4+; overcome by mild, moderate, and strong forces respectively). Of the nine muscle groups tested in the 19 patients, 28 muscle groups were considered of normal strength (MRC grade 5); 37, 33, 34, and 39 were graded $\leqslant 3$, 4–, 4, and 4+ respectively. Wasting was present in the upper limbs in 15 patients (six mild, six moderate, three severe), and in the lower limbs in eight (four mild, four moderate). In the upper limbs 13 had normal tone, three were spastic (one mild, one moderate, one severe), and three were hypotonic; in the lower limbs seven had normal tone, 10 were spastic (eight mild, two moderate), and two were hypotonic.

DEFINITIONS

Accuracy (experiment 1) is the difference between dynamometry estimations of the forces tested from the true magnitude of the

*Table 1 Positioning for muscle group testing in experiment 2**

Muscle group	*Limb position*	*Dynamometer placement*
Shoulder abduction (SA)	Abducted to 90° (palm down, arm flexed to 30°)	Proximal to lateral epicondyle of humerus
Elbow flexion (EF)	Shoulder adducted, elbow flexed to 90°, forearm supine	Flexor surface of distal forearm, at wrist crease
Elbow extension (EE)	Shoulder adducted, elbow flexed to 90°, forearm mid-prone	Distal forearm, at ulnar styloid
Finger extension (FE)	Forearm prone, wrist fixed, fingers extended	Dorsum of fingers, at proximal phalanges
Index finger abduction (IFA)	Index finger abducted, other fingers fixed	Radial surface of index, at proximal interphalangeal joint
Hip flexion (HF)	Hip flexed to 45°, knee flexed to 90°	Extensor thigh, proximal to knee joint
Knee extension (KE)	Hip and knee flexed to 90°	Tibial surface, proximal to ankle joint
Foot dorsiflexion (FD)	Knee extended, ankle at 90°	Dorsum foot, over metatarsals
Neck flexion (NF)	Neck flexed to 45°	Centre of forehead

*All muscle groups were tested with the subject in a seated position;

forces. The degree of accuracy was assessed by calculating error rate (error rate (%) = (dynamometer reading − true weight) × 100/true weight). The error rates of all 10 replications at both sessions were used to estimate accuracy.

Reproducibility was expressed as percentage difference (% difference = (difference between readings) × 100/(mean of the two readings)) and repeatability coefficient (repeatability coefficient = 2 × standard deviation (SD) of differences between repeated readings).[16] All readings taken at the two sessions were used to assess reproducibility in experiment 1; since the coefficient of variation (CV) was measured over six days, readings six days apart from all patients and muscle groups were used for estimating reproducibility in experiment 2. Pearson product-moment correlation coefficients were also calculated (see discussion), the readings at all forces tested (experiment 1) and from all patients and muscle groups (experiment 2) being normally distributed.

Variability was expressed as the CV (CV(%) = (SD × 100)/mean). CVs of 10 replications for each force at each session were used in experiment 1; CVs in experiment 2 were calculated from 10 readings at separate sessions over six days for each muscle group in each patient.

DATA ANALYSIS

Accuracy

Multifactor analysis of variance was performed on all error rates, which were normally distributed, to assess the effects of rater, force tested, and sessions on accuracy in experiment 1.

Reproducibility

Log transformations of all % differences (and all correlation coefficients) were normally distributed; analysis of variance was then used to assess the effect of choice of test index (maximum, median or mean of three readings per session) on reproducibility in experiment 2.

Subsequent analyses of data in experiment 2 was performed on the *maximum* of three readings. Multifactor analyses of variance on log % difference were used to assess the effects of rater, force tested, and sessions on reproducibility in experiment 1; and of sex, motor neuron disease type, upper or lower limb, and right or left sides of body in experiment 2. With data obtained from one side of the body only, multifactor ANOVA showed the effects of different muscle groups and patients on reproducibility.

To study differences in reproducibility between motor neuron disease types, Spearman rank correlation coefficients were used to show the relation between MRC grade and tone on % differences and correlation coefficients. Sums of MRC grades for all muscles tested (muscle scores) were made for all patients; non-parametric methods (Mann-Whitney U test) were used to compare MRC grades (by individual muscles, and by patient's muscle scores) between patients with bulbar disease and classical amyotrophic lateral sclerosis.

Variability

CVs were normalised by log transformations. Further analysis was performed as for reproducibility (see earlier).

The effects on accuracy, reproducibility, and variability of using readings obtained by multiple raters was studied further. In experiment 1 error rates, % differences, and CV of nine readings obtained jointly by three raters (combining the first three of 10 readings obtained by each rater) were compared with those of nine readings obtained by each rater separately, by the Mann-Whitney U test.

ESTIMATE OF RELATIVE CONTRIBUTIONS OF FACTORS TO ACCURACY, REPRODUCIBILITY, AND VARIABILITY IN EXPERIMENTS 1 AND 2

The ratios of mean squares from multifactor analyses of variance were used to estimate the relative contributions of various factors to accuracy, reproducibility, and variability. The **appendix** gives the details of this calculation for variability.

Accuracy

A multifactor analysis of variance on error rates in experiment 1 was used to assess the relative contributions of rater, force tested, order of reading (within a session), and different sessions on accuracy.

A close comparison of reproducibility and variability between both experiments was achieved by: (1) using only the first three readings in experiment 1 for each force level at each session, (2) using readings from the first two sessions, which were three hours apart in experiment 2; (3) as there were four force levels tested in experiment 1, readings in experiment 2 were also divided into four MRC force grades (4−, 4, 4+, 5).

Reproducibility

The relative contribution of rater to reproducibility in testing patients was estimated by comparing the variance of % differences for rater 1 in experiment 1 with the overall variance of readings in experiment 2. Multifactor analyses of variance on log % differences were then used to estimate the effects of force tested and order of reading (within a session) on reproducibility in experiment 1, and of force, order, patient, and muscle group tested in experiment 2.

Variability

All readings were first standardised by expressing them as a percentage of the mean of three readings in each session for that force level (experiment 1) or muscle tested (experiment 2). This was necessary to make the variances of readings at forces of different magnitudes comparable. The variance of the standardised readings for rater 1 in experiment 1 was taken as an estimate of rater variability. Similarly, the variance of the standardised readings in experiment 2 was an estimate of overall variability. The ratio

Table 2 Summary of main results from both experiments

	Accuracy	*Reproducibility*			*Variability*
	% Error Mean (95% CI)	*% Difference Mean (95% CI)*	*Repeat coeff*	*Pearson coeff*	*CV Mean (95% CI)*
Experiment 1:			kg-force		
Rater 1	5·4 (3·7 to 7·2)	8·0 (2·0 to 13·9)	2·20	0·95	4·8 (3·5 to 6·2)
Rater 2	0·0 (−1·8 to 1·7)	1·8 (−4·1 to 7·7)	1·24	0·98	5·0 (3·7 to 6·4)
Rater 3	3·6 (1·8 to 5·3)	4·0 (−2·0 to 9·9)	1·81	0·95	6·3 (5·0 to 7·6)
Experiment 2 (muscle group):					
Shoulder abduction (n = 18)		12·7 (6·1 to 19·4)	1·98	0·97	10·1 (7·8 to 12·3)
Elbow extension (n = 19)		9·0 (2·5 to 15·4)	2·36	0·98	9·1 (6·9 to 11·3)
Elbow flexion (n = 16)		8·9 (2·0 to 15·7)	1·49	0·98	8·1 (5·7 to 10·5)
Finger extension (n = 16)		14·6 (7·5 to 21·7)	0·78	0·94	10·8 (8·4 to 13·2)
Index finger abduction (n = 13)		15·6 (7·6 to 23·6)	0·46	0·93	13·1 (10·5 to 15·7)
Hip flexion (n = 16)		18·4 (11·6 to 25·3)	2·93	0·96	9·3 (6·9 to 11·7)
Knee extension (n = 7)		14·0 (3·2 to 24·8)	2·14	0·96	8·8 (5·2 to 12·4)
Foot dorsiflexion (n = 11)		15·9 (7·6 to 24·3)	2·00	0·98	13·5 (10·6 to 16·4)
Neck flexion (n = 18)		12·5 (5·8 to 19·1)	2·85	0·90	8·2 (5·9 to 10·4)
Composite scores:			% Normal		
Upper limb (n = 19)		6·8 (3·9 to 9·7)	12·11	0·99	6·0 (3·0 to 8·9)
Lower limb (n = 19)		10·6 (7·6 to 13·7)	3·39	0·97	7·0 (3·9 to 10·1)
Global limb (n = 19)		5·8 (2·8 to 8·9)	7·98	0·98	6·7 (3·7 to 9·6)

Repeat coeff = repeatability coefficient (of readings 6 days apart); Pearson coeff = Pearson's correlation coefficient (of readings six days apart); CV = coefficient of variation (of 10 readings, taken over 6 days); data for maximum of three replications; only right side used for muscle groups tested, both sides for composite scores.

between these two quantities gave an estimate of rater contribution to overall variability in readings obtained from patients. Another estimate was obtained by comparing the overall mean CV (of three readings per force level per session) for rater 1 in experiment 1 with the overall mean CV in experiment 2. The ratios of mean squares were calculated on the standardised readings.

Results

Table 2 presents a summary of the main results.

EXPERIMENT 1

Accuracy

The average accuracy of readings from all raters and forces tested was 3%. There were significant differences in error rate of readings obtained by different raters ($p = 0·0001$); this was solely due to the most experienced rater, who was significantly more accurate (mean error rate of 0%). There were significant differences in accuracy between forces tested ($p = 0·03$) and between sessions ($p = 0·0003$).

Reproducibility

For readings obtained by all raters at all forces, the mean difference between readings was 4·6%, with a repeatability coefficient of 1·79 kg-force—that is, less than 1·79 kg-force difference between repeated readings on 95% of occasions. The Pearson correlation coefficient was 0·96. The % difference was not affected by either rater or force tested.

Variability

The CVs averaged 5·4%. There were no significant differences in variability between raters, or forces tested, or between sessions.

Groups of nine readings obtained jointly by three raters were 53% more variable than nine readings obtained by a single rater in experiment 1 (CV of 8·7 and 5·7% respectively; $p < 0·009$). Accuracy and reproducibility were similar.

EXPERIMENT 2

Reproducibility

The % differences (and correlation coefficients) between two assessements were similar with different test indices; *maximum* was used subsequently. Mean difference was 13·2% for readings taken six days apart (right side only and neck flexion), with a mean repeatability coefficient of 2·17 kg-force. The overall mean correlation coefficient was 0·98 (for all patients 0·98 and for all muscles 0·95).

Patients with bulbar palsy had smaller % differences than those with classical amyotrophic lateral sclerosis ($p = 0·001$), but there were no significant differences between sexes, upper and lower limbs, or right and left sides of the body. The % differences were highly correlated to MRC grade (stronger muscles producing smaller % differences, $p = 0·002$) but not to tone. The MRC grades were higher in bulbar palsy (by muscles $p < 0·0001$; by patients $p = 0·06$); this probably accounts for the better reproducibility in these patients. There were no significant effects of patients or muscle groups on % differences. Readings taken three hours apart and six days apart resulted in similar % differences. The same analysis for correlation coefficients showed no difference between types of motor neuron disease, otherwise results were similar to % differences.

Variability

Neither test index, nor sex, assessing upper or lower limbs, or right or left sides of the body had any significant effects on the CV; however, patients with bulbar palsy had less variability than those with amyotrophic lateral sclerosis ($p = 0·0003$). When only right sided muscle groups were used, there were significant differences in variability between patients ($p = 0·0001$) and muscle groups ($p = 0·0004$); the mean CV of 10 assessments from all patients and muscle groups was 9·9%. The CVs were similar between five assessments performed in a single 24 hour period and those on five separate days.

The CVs were highly correlated to the MRC grade ($p = 0.0001$), but not to tone. The differences in variability were probably due to patients with bulbar palsy having less clinically weak muscles.

ESTIMATE OF RELATIVE CONTRIBUTIONS OF FACTORS TO ACCURACY, REPRODUCIBILITY, AND VARIABILITY IN EXPERIMENTS 1 AND 2

Accuracy

The different raters, forces tested, order of reading (within a session), and sessions contributed 34%, 11%, 2%, and 49% respectively to the total inaccuracy of readings in experiment 1. The remaining 4% comprised error rate of the dynamometer (1%) and unaccounted factors (3%). It was impossible to estimate accuracy in experiment 2.

Reproducibility

The variance of % differences for rater 1 in experiment 1 was 49·61, and the overall variance of % differences in experiment 2 was 102·01. Thus the rater was estimated to contribute 49% of the overall differences between readings at sessions 3 hours apart. The factors contributing to this 49% loss of reproducibility due to rater were the different forces tested (22%) and order of reading within a session (11%); 16% were due to unaccounted factor(s) and random error. The remaining 51% contribution to differences in readings were due to intrapatient (24%) and interpatient (19%) factors, with 8% due to unaccounted factor(s) and random error. The 24% of intrapatient factors comprised differences between muscle groups (11%) and forces (2%) tested, and the order of readings (11%) within a session.

Variability

The variance of standardised readings (see data analysis) obtained by rater 1 in experiment 1 was 17·20, and the overall variance in experiment 2 was 45·98. The rater was thus estimated to contribute 37% to the overall variability. A comparison of the mean CV (of groups of three readings, see data analysis) for rater 1 in experiment 1 (4·1%) with the overall mean CV in experiment 2 (9·9%) led to a similar estimate of rater contribution (41%) to overall variability.

Data from experiment 1 indicated that the 37% contribution of rater to overall variability was due mainly to the effects of the order of reading within a session (14%). Different sessions and forces tested made little contribution (both < 1%); the remaining 23% of rater variability were due to unaccounted factor(s) and random error. Data from experiment 2 indicated that the remaining 63% of overall variability was due mainly to the effects of order of reading within a session on patients (62%). Different patients and muscle groups tested made little contribution to overall variability (both < 1%), the remaining 1% being due to unaccounted factor(s) and random error. The dynamometer's contribution (0·5%) was also insignificant.

Composite scores

The composite scores constructed led to improvement in both reproducibility and variability (table 2). These improvements were significant for % differences and CVs for upper and global limb scores ($p < 0.03$) compared with analysing those muscle groups separately.

Discussion

The first experiment showed that overall accuracy of hand-held dynamometry was good, with a mean error rate of 3% for all

*Table 3 Comparison of results from fixed and hand-held dynamometry**

			Muscle group tested							
	Population		*Elbow flexion*		*Elbow extension*		*Hip flexion*		*Knee extension*	
Authors	*Type*	*No*	*r*	*CV*	*r*	*CV*	*r*	*CV*	*r*	*CV*
Fixed:										
Tornvall 1963	Normal	44	–	3·2	–	7·1	–	7·1	–	4·4
Fowler and Gardner 1967	Duchenne	11	0·98	–	0·99	–	0·96	–	0·99	–
	Other dystrophy	11	0·93	–	0·99	–	0·99	–	0·99	–
Hyde and Goddard 1983†	Duchene	12	–	–	–	–	–	–	0·91	–
Wiles and Karni 1983†‡	Normal	6	–	–	–	–	–	–	–	8·5
Andres *et al* 1986	Normal	35	0·97	–	0·96	–	0·97	–	0·97	–
	ALS	10	0·99	–	0·99	–	0·99	–	0·99	–
Hand-held:										
Hyde and Goddard 1983	Duchenne	9	–	–	–	–	0·94	4·6	0·96	9·1
	SMA	2	–	–	–	–	–	–	–	–
	Limb girdle	1	–	–	–	–	–	–	–	–
Wiles and Karni 1983†‡	Neuropathy or polymyositis	3	–	6·3	–	10·1	–	6·3	–	17·9
Stuberg and Metcalf 1988	Normal	14	0·98	–	–	–	–	–	0·98	–
	Duchenne	14	0·98	–	–	–	–	–	0·99	–
Riddle *et al* 1989§	Brain damaged:									
	Paretic	14	0·99	–	0·96	–	0·96	–	0·87	–
	Non-paretic	11	0·76	–	0·93	–	0·78	–	0·77	–
Present study	MND	19	0·96	4·8	0·97	4·1	0·98	4·9	1·00	4·5

*Data for two sets of readings from one side of body, performed within a 24 hour period (unless otherwise stated).
†Data for both sides of body.
‡CV of 13 readings over 5 month period with fixed device, but of five assessments in a single 24 hour period for hand-held device.
§Correlation of two readings taken 48 hours apart.
r = Pearson correlation coefficient; SMA = spinal muscular atrophy; ALS = amyotrophic lateral sclerosis; MND = motor neuron

three raters. The experienced rater was more accurate, achieving an overall error rate of 0%; the less experienced raters tended to overestimate the forces. Overall reproducibility and variability of raters were also good, with a mean % difference of 4·6% and a CV of 5·4%; experience did not affect these variables. The overall correlation coefficient was 0·96. The conditions of this experiment simulated shoulder abduction force assessments in subjects. The figures obtained for variability of rater against a fixed object cannot be assumed to be exactly the same as that occurring when different muscle groups are assessed in patients. They do provide a reasonable estimate, however, of a minimum variability to be expected from a rater when assessing patients.

Many studies in this field have quoted the Pearson correlation coefficient as the only index of reproducibility. Studies with both fixed and hand-held dynamometry have described highly correlated readings in healthy normal subjects and in patients with neurogenic or myopathic disorders (table 3). This coefficient has been criticised as only indicating the closeness to any straight line relation between two sets of readings, irrespective of the differences between them.[16] This point is well exemplified in table 2, which shows that muscles with relatively large repeatability coefficients and % differences had high Pearson correlation coefficients. The use of intraclass correlation coefficient,[17] although more appropriate in this field, has similar constraints. The repeatability coefficient[16] conveys information on the actual differences between two sets of readings, but has the disadvantage that its interpretation depends on the absolute values compared. The % difference is an index of reproducibility that immediately conveys the degree of closeness between two sets of readings, irrespective of their absolute values.

The reproducibility of readings obtained by hand-held dynamometry in this study compare well with those obtained by fixed devices; differences in readings three hours apart in our patients (range 8·7%–12·0%; right side) were similar to those obtained in 10 patients with amyotrophic lateral sclerosis[11] using a fixed device (range 5·5%–12·4%; right side). Using hand-held dynamometry on 100 healthy subjects,[18] a mean week to week difference by muscle group of 8·9% (range 5·1%–14·2%) was found; the equivalent mean difference in our study was 13·2% (range 8·9%–18·4%).

The higher intraclass and Pearson correlation coefficients of repeated readings in paretic limbs compared with non-paretic limbs described in brain damaged patients is of interest,[14] the authors' interpretation being that paretic limbs give more reliable readings. Paretic limbs are limited to performing stereotyped movements, compared with the wider range of movements present in non-paretic limbs; the authors speculated that this may have accounted for the differences. Our data showed the same Pearson correlation coefficient (0·97) for weak muscle groups (MRC grade < 5) and those of normal strength (MRC grade 5); correlation coefficients across the range of MRC grades (4–, 4, 4+ and 5) showed no systematic trend. By contrast, our analysis of % differences suggested that dynamometry readings in weaker muscles were less reproducible.

Previous workers have used different test indices resulting from three replications. The use of the maximum is attractive (reflecting the maximal force that can be exerted by a muscle); the variability of this reading can be affected by differences in patient effort at different sessions. The mean is prone to the effects of abnormally high or low readings in one assessment; this may arise in easily fatigued patients, as seen in motor neuron disease.[19 20] The median will not be affected so much by outliers. We found that the choice made no differences to variability.

The variability of readings obtained by hand-held dynamometry compares well with that obtained by fixed devices (table 3), with a mean CV of 9·9%. Agre *et al*[7] found lower limb muscle groups more variable than upper groups in eight healthy subjects; this may have resulted from testing muscle groups that were technically difficult to stabilise (for example, hip abduction and extension). By contrast, we found similar variability in upper and lower limb muscle groups.

Previous studies on hand-held dynamometry in neurological and healthy subjects have suggested that employing different raters may adversely affect variability.[8 9] Others, by performing an ANOVA on the CV of readings obtained by four raters on four subjects, estimated the rater to contribute 2%–4% to the overall variability, whereas the subjects were responsible for 79%–86%.[6] Our study, by employing a spring loaded device that eliminates patient related variability, confirms that the combined readings from three raters increase variability by 53% when compared with a single rater. These findings are important when planning serial assessments of patients in clinical trials or undergoing natural history studies, particularly in calculating the sample sizes required, as numerous raters may be employed. Further, we estimated that the rater was responsible for 37% of the variability of readings in patients with motor neuron disease (see appendix). The larger component of variability due to the patients is probably related to the effect of fatigue. A comparison of differences in readings across sessions (reproducibility) suggests an equal contribution of rater and patient.

The effects of using composite scores in hand-held dynamometry have not been considered before. These scores were designed to obtain an overall view of total muscle strength for long term assessments. The benefits of their use on data reduction are immediately apparent. The use of such scores led to significant improvements in reproducibility and variability compared with muscle groups being tested separately.

We conclude that the variability of readings

obtained by hand-held dynamometry can be minimised by the use of a single rater. Experience in use of the dynamometer is not required to obtain readings with good reproducibility and variability. If multiple raters are used for longitudinal studies of muscle force, individual patients should be assessed by the same rater throughout. If this is not feasible, the variability of the combined readings of all raters should be established beforehand; a procedure such as the one described in this work could be used. Otherwise, estimates of the sample size required to detect specified changes will be inaccurate.

Appendix

Let the variability of a single rater in experiment 1 = V1 and the total variability in experiment 2 = V2

$$V_1 = V_D + V_{Fr} + V_{Or} + V_{Sr} \quad (1)$$

where V_D = dynamometer's variability; V_{Fr} = variability at different forces attributable to the rater; V_{Or} = variability due to the order of force reading (within a session) due to the rater; V_{Sr} = intersession variability attributable to the rater

$$V_2 = V_D + V_F + V_O + V_S + V_P + V_M \quad (2)$$

where V_F = variability between different forces tested; V_O = variability due to the order of force reading (within a session); V_S = intersession variability; V_P = variability between patients; V_M = variability between different muscle groups. Also:

$$V_F = V_{Fr} + V_{Fp} \quad (3)$$

$$V_O = V_{Or} + V_{Op} \quad (4)$$

$$V_S = V_{Sr} + V_{Sp} \quad (5)$$

where V_{Fp} = variability at different force levels attributable to the patient; V_{Op} = variability due to the order of force reading (within a session) due to the patient; V_{Sp} = intersession variability attributable to the patient.

An estimate of V_1 is obtained by calculating the variance of the standardised force readings (see data analysis) of the rater in experiment 1 who subsequently participated in experiment 2. Similarly, an estimate of V_2 is the variance of the standardised force readings obtained in experiment 2.

Multifactor ANOVAs were performed on these standardised variables (which were normally distributed) by force tested, order of reading within a session, and different sessions in experiment 1; and by force level tested, order of reading within a session, different sessions, patients, and muscle groups in experiment 2. Resultant mean squares are proportional to their relative contributions to the overall variability in experiments 1 and 2. A value of 0·5% was taken for V_D (manufacturer's figure).

Using the estimate for V_1, the ratios of mean squares from a multifactor ANOVA on experiment 1 can be used with equation (1) to obtain estimates for V_{Fr}, V_{Or}, and V_{Sr}. Similarly, using the estimate for V_2, the ratio of mean squares from a multifactor ANOVA on experiment 2 can be used with equation (2) to obtain estimates for V_F, V_O, V_S, V_P and V_M. Because V_F, V_O, V_S, V_{Fr}, V_{Or}, and V_{Sr} have now been determined, the remaining factors (V_{Fp}, V_{Op}, and V_{Sp}) may then be calculated from equations (3), (4), and (5) respectively.

We are grateful to Mr A J Overill and the Department of Medical Physics for designing and constructing the spring-loaded device, to Mr J Emami for statistical advice, and to Miss J Roberts for secretarial help. This work was funded by the Motor Neuron Disease Association of the United Kingdom and the Special Trustees of Westminster and Roehampton Hospitals.

1 Beasley WC. Influence of method on estimates of normal knee extensor force among normal and postpolio children. *Phys Ther Rev* 1956;**36**:21–41.
2 Lovett RW, Martin EG. Spring balance muscle test. *Am J Orthop Surg* 1916;**14**:415–24.
3 Edwards RHT, McDonnell M. Hand-held dynamometer for evaluating voluntary-muscle function. *Lancet* 1974;**ii**:757-8.
4 Hyde SA, Goddard CM. The myometer: the development of a clinical tool. *Physiotherapy* 1983;**69**:424–7.
5 Hosking GP, Bhat US, Dubowitz V, Edwards RHT. Measurement of muscle strength and performance in children with normal and diseased muscle. *Arch Dis Child* 1976;**51**:957–63.
6 Ploeg RJO van der, Oosterhuis HJGH, Reuvekamp J. Measuring muscle strength. *J Neurol* 1984;**231**:200–3.
7 Agre JC, Magness JL, Hull SZ, *et al.* Strength testing with a portable dynamometer: reliability for upper and lower extremities. *Arch Phys Med Rehab* 1987;**68**:454–8.
8 Rheault W, Beal JL, Kubik KR, *et al.* Intertester reliability of the hand-held dynamometer for wrist flexion and extension. *Arch Phys Med Rehab* 1989;**70**:907–10.
9 Wiles CM, Karni Y. The measurement of muscle strength in patients with peripheral neuromuscular disorders. *J Neurol Neurosurg Psychiatry* 1983;**46**:1006–13.
10 Bohannon RW. Test-retest reliability of hand-held dynamometry during a single session of strength assessment. *Phys Ther* 1986;**66**:206–9.
11 Andres PL, Hedlund W, Finison L, *et al.* Quantitative motor assessment in amyotrophic lateral sclerosis. *Neurology* 1986;**36**:937–41.
12 Bohannon RW, Andrews AW. Interrater reliability of hand-held dynamometry. *Phys Ther* 1987;**67**:931–3.
13 Stuberg WA, Metcalf WK. Reliability of quantitative muscle testing in healthy children and in children with Duchenne muscular dystrophy using a hand-held dynamometer. *Phys Ther* 1988;**68**:977-82.
14 Riddle DL, Finucane SD, Rothstein JM, Walker ML. Intrasession and intersession reliability of hand-held dynamometer measurements taken on brain-damaged patients. *Phys Ther* 1989;**69**:182-9.
15 Guiloff RJ, Eckland DJA, Demaine C, *et al.* Controlled acute trial of a thyrotrophin releasing hormone analogue (RX77368) in motor neuron disease. *J Neurol Neurosurg Psychiatry* 1987;**50**:1359–70.
16 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–11.
17 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;**86**:420–8.
18 Ploeg RJO van der, Fidler V, Oosterhuis HJGH. Hand-held myometry: reference values. *J Neurol Neurosurg Psychiatry* 1991;**54**:244–7.
19 Denys EH, Norris FH. Amyotrophic lateral sclerosis: impairment of neuromuscular transmission. *Arch Neurol* 1979;**36**:202–20.
20 Guiloff RJ, Eckland DJA. Observations on the clinical assessment of patients with motorneuron disease: experience with a TRH analogue. *Neurol Clin* 1987;**5**: 171–92.
21 Tornvall G. Assessment of physical capabilities. *Acta Physiol Scand* 1963;**58**(Suppl 201):4–102.
22 Fowler WM, Gardner GW. Quantitative strength measurement in muscular dystrophy. *Arch Phys Med Rehab* 1967;**48**:629–44.