JHT READ FOR CREDIT #039

The Upper Limb Functional Index: Development and Determination of Reliability, Validity, and Responsiveness

C. Philip Gabel, MSc, PT
Lori A. Michener, PhD, PT, ATC
Virginia Commonwealth University
Medical College of Virginia
Richmond, Virginia

Brendan Burkett, PhD
Anne Neller, PhD
University of the Sunshine Coast
Queensland, Australia

ABSTRACT: Purpose. Current upper limb regional self-report outcome measures are criticized for poor clinical utility, including length, ease, and time to complete and score, missing responses, and poor psychometric properties. To address these concerns a new measure, the Upper Limb Functional Index (ULFI), was developed with reliability, validity, and responsiveness being determined in a prospective study.

Methods. Patients from nine Australian outpatient settings completed the ULFI and two established scales, the Disabilities of the Arm, Shoulder, and Hand (DASH) ($n = 214$) and the Upper Extremity Functional Scale (UEFS) ($n = 64$) concurrently to enable construct and criterion validity to be assessed. Two subgroups were used to assess test–retest reliability at 48-hour intervals ($n = 46$) and responsiveness through distribution-based methods ($n = 29$). Internal consistency, change scores, and missing responses were calculated. Practical characteristics of the scale were assessed.

Results. The ULFI correlated with the DASH ($r = 0.85$; 95% CI) and UEFS ($r = 0.78$; 95% confidence interval [CI]), demonstrated test–retest reliability (intraclass correlation coefficient = 0.96; 95% CI) and internal consistency (Cronbach alpha = 0.89). The change scores of the ULFI with standard error of the measurement was 4.5% or 1.13 ULFI-points and minimal detectable change at the 90% CI was 10.4% or 2.6 ULFI-points. Responsiveness indices were standardized response mean at 1.87 and effect size at 1.28. The ULFI demonstrated an impairment range of 0–100%, with no missing responses and a combined patient completion and therapist scoring time of less than 3 minutes.

Conclusions. The ULFI demonstrated sound psychometric properties, practical characteristics, and clinical utility thereby making it a viable clinical outcome tool for the determination of upper limb status and impairment. The ULFI is suggested as the preferred upper limb regional tool due to its superior practical characteristics and clinical utility, and comparable psychometric properties without a tendency toward item redundancy.

J HAND THER. 2006;19:328–49.

The use of standardized Self-report Outcome Measures (SROMs) for the determination of functional impairment and the monitoring of change over time has gained increasing favor over the last decade. They are defined as "… a questionnaire completed by the patient to indicate the status of functional loss in a specific area or condition …."[1–6] Musculoskeletal SROMs include condition-specific tools for the various joints and diseases and more recently, region-specific tools have emerged as the preferred option due to their greater application across a variety of clinical and research situations.[7–11] Region-specific tools consider the body in single kinetic chains of the spine and upper and lower extremities[8,9] and provide a means of clarifying clinical status and any subsequent changes that may result from treatment or intervention. They are more practical and easier to implement and administer than objective clinical measures,[7,12,13] responsive to significant improvements over time,[10,11] and require fewer patient numbers to detect an effect.[14] Evidence indicates that only

Correspondence and reprint requests to C. Philip Gabel, MSc, PT, PO Box 760, Coolum, Queensland 4573, Australia; e-mail: <cp.gabel@bigpond.com>.

0894-1130/$ – see front matter © 2006 Hanley & Belfus, an imprint of Elsevier Inc. All rights reserved.

doi:10.1197/j.jht.2006.04.001

a modest correlation exists between impairment status, functional loss, and subsequent participation.[15–19] highlighting the importance of SROM tool use. Professional organizations and third-party insurance payers emphasize the need for tools that are both valid and clinically relevant.[8,16,20,21]

Only four region-specific upper limb tools developed for use in general populations were found in the literature: the Disabilities of the Arm, Shoulder, and Hand (DASH),[9] the Upper Extremity Functional Scale (UEFS),[14] the Upper Extremity Functional Index (UEFI),[10] and the Neck and Upper Limb Index (NULI).[22] Each of these tools uses item statements to test different constructs that can be broadly categorized into four themes: upper limb specific function—such as writing, holding, using utensils, and over head activity; general function—such as driving, work, hobbies, and house chores; health-related quality of life (HRQOL)—such as sleeping, social contact, anxiety, and irritability; and pain specific—such as intensity, duration, and ease of provocation.

Specific deficiencies have been identified in each of these tools that limit their adoption within the clinical setting. These involve four areas. First, ''comprehensiveness,'' relating to both adequacy of the item domains covered by the tool[7,23–25] and generalization from a specific sample population, such as workers in the UEFS and NULI.[14,26,27] Second, ''relevance,'' pertaining to small samples with a high average age as in development of the UEFI.[10,25] Third, ''practical characteristics,'' limiting clinical utility due to excessive time or errors in completion and scoring, particularly missing responses (items not marked or not applicable) as noted for each of the four published tools, as a Likert scale is used,[7,8,14,25] or redundancy (the presence of several similar items) as in the DASH and NULI.[7,25] Finally, ''psychometric properties,'' including the reliability or stability a tool has and whether it can measure change, as is the case with the UEFS[7,10] and significant variations in the level of responsiveness, as found with the DASH.[7,28,29]

Since the effectiveness of any intervention strategy can be demonstrated best by accurate measurement, the tools used to assess outcome must consequently reflect this by being valid, reliable, responsive, and representative of the target population.[30–34] In determining if there is a genuine need for a new tool, three factors must be considered: the goals of the intended measure, whether existing measures meet these goals, and if they have deficiencies in either their practical characteristics or psychometric properties.[10,35] These fundamental requirements for any SROM tool have been supported by numerous publications over the past two decades.[7,35–38] The inadequacies of the existing tools noted above indicate the need for a SROM that is clinically utile with brevity and ease of completion; is simple to score with no serious floor or ceiling effects in general clinical populations; has minimal missing responses; and demonstrates the essential psychometric properties through validity, reliability, and responsiveness.[7,11] If the decision to develop a new tool is made, then such a tool should be consistent with the accepted definition of activity limitation defined by the World Health Organization's International Classification of Impairments, Activities, and Participation as ''… difficulties an individual may have in executing activities.''[39] This definition places emphasis on neutral or positive terms and the avoidance of negative terms such as disability and handicap present in the original definition.

The present study has three aims. First, to develop and validate the initial core quantitative component of the Upper Limb Functional Index (ULFI) through a single-page, three-part SROM with an initial quantitative 25 items focused on assessing HRQOL and upper extremity dysfunction, a subsequent Patient Specific Index (PSI) section for interpreting qualitative information,[40,41] and an 11-point Visual Analogue Scale (VAS) for current ''overall status'' compared to the preinjury or ''normal'' level (see Appendix A). Second, to concurrently compare the performance of the ULFI with the criteria standards of the DASH and UEFS in terms of psychometric properties and practical characteristics. The DASH and UEFS were chosen as they represent advocated and established standards from the peer reviewed literature,[8,9,42] their measurement properties have been formally demonstrated and independently validated,[10,14,29] and they represent both ends of the practicality scale in terms of length, ease of administration and scoring during patient interaction.[43] The UEFI and NULI were not selected as both lacked published independent investigation at the initiation of this research and were developed from specific population samples that lacked broad demographics in terms of age and occupation, respectively. Third, to make recommendations on a preferred clinical tool and requirements for subsequent investigations.

MATERIALS AND METHODS

Study Design

The development, construction, and final validation of a new regional upper limb questionnaire require a methodological process that is systematic and follows established protocols. The ''Guyatt Model'' of questionnaire development[30,44,45] achieves this through its systematic three-stage process as demonstrated in Figure 1. This model and process incorporate the literature search and review strategy instituted by Michener and Leggins[7] and mirror the

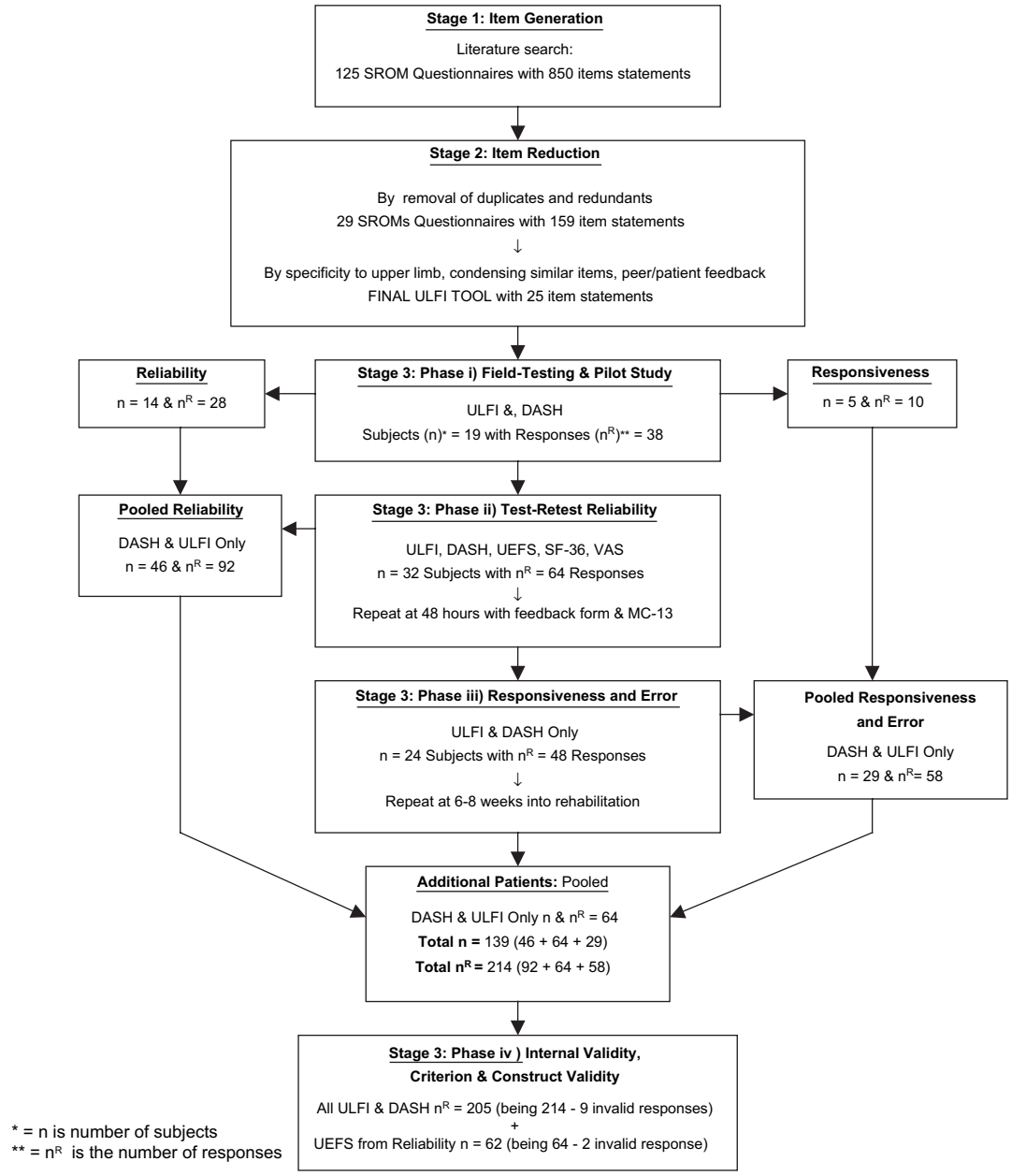

FIGURE 1. *Flow chart of the processes of Stages 1–3 of Upper Limb Functional Index (ULFI) development and testing. * = n is number of subjects; ** = n^R is the number of responses.*

development model of the DASH[38] and comparative investigation of other SROMs.[46] The current study uses a repeated measures and correlation design with a patient sample of convenience and was approved by the University Human Research Ethics Committee.

Stage 1 ''item generation'' and Stage 2 ''item reduction'' involved the collection of possible items from previously published upper extremity scales and subsequent reduction of the bank of items to the final 25 for the ULFI. Stage 3 concerned the validity, reliability, and responsiveness designed to concurrently evaluate the psychometric properties of the new ULFI instrument and the established tools.

Stages 1 and 2: Item Generation and Reduction

The tool development phase was initiated by a review of the literature using the electronic databases, Pubmed Online—MEDLINE and CINAHL from 1980 to 2002, performed with the key words of outcomes, self report, function, disability, and upper limb. In addition, a peer reviewed search by means of direct contact with physiotherapy, occupational therapy, and hand therapy clinicians and researchers enabled a further clarification of the available ''grey or soft literature'' that included tools from conference proceedings, under current research or unpublished.

Stage 1 ''item generation'' produced a total of 71 SROMs that provided 850 potential items for an upper limb questionnaire tool. Stage 2 ''item reduction'' reduced the list to 29 SROMs and 159 items through removal of duplications and redundancies. These were in turn reduced to the final 25 items for Stage 3 by item exclusion based on specificity to the upper limb or HRQOL, condensing similar items into a single item[10,38] and peer and patient feedback. The process also determined the tool format for the essential quantitative item statements plus the qualitative and VAS status sections, thereby ensuring that both face and content validity were accounted for in the development phase.[9,34,47,48]

Stage 3: Reliability and Validity Testing

Subjects

Stage 3 involved four phases: Phase i—an initial pilot investigation ($n = 19$ subjects providing 38 responses), for preliminary assessment of reliability, responsiveness, and the presence of floor capacity and ceiling effect. Phases ii–iv of Stage 3 were designed to assess the psychometric properties of the SROMs using a further 120 subjects to provide 176 responses. Specifically, Phase ii, test–retest reliability ($n = 32$ with 64 responses); Phase iii, responsiveness ($n = 24$ with 48 responses); and Phase iv, the internal consistency, criterion, and construct validity ($n = 64$ with 64 responses), see Figure 1 and Table 1. All phases in Stage 3 involved a total of 139 different subjects providing 214 responses sampled from nine locations in three Australian states and Territories representing public health, private health (medical, rehabilitation, and physiotherapy), and Department of Defence establishments. Subject inclusion criteria were any patients, at least 18 years old, with upper limb symptoms under medical or allied health management. Exclusion criteria were inability to read English or respond to the questionnaire. All subjects completed an intake form for demographic data, the newly developed ULFI and the DASH. Varying subsamples of subjects in Stage 3 completed an 11-point VAS for status, the Short-Form 36 (SF-36) general health measure,[49] and the Marlowe–Crowne Short-Form 13 (MC-13) psychological screening tool for social desirability and the UEFS.

The DASH is a SROM with three parts. The first section consists of instructions, the second contains the essential 30 items, and the third is an optional module for the assessment of sports and performing arts. A 1–5 Likert scale is used for each item with a subsequent score range of 30–150. This raw score is recalculated for expression on a 100-point scale by subtracting 30 then dividing by 1.2 and allows for up to three missing responses by adding the number of missing factors multiplied by the average of the initial score per item. The DASH was developed using the previously described three-stage model,[38] validated,[29] then published with supporting research on its test–retest reliability, construct validity, and responsiveness.[9]

The UEFS is a practical SROM, with one section containing eight items using a 0–10 Likert scale format. This provides a potential summated score range from 0 to 80 and multiplied by 1.25 to be converted to a 100-point scale. One missing response is permitted and accounted for by interpolation into the final score as the average of the other responses. The UEFS was specifically developed to measure the impact of upper extremity disorders on function in a workers' compensation population. The item list was selected from an original set of 12 obtained by means of discussion groups using physicians, occupational therapists, and patients. This tool has a functional focus with only one item being oriented to HRQOL.[7,8,10,14]

Phase i of Stage 3: The pilot investigation of the ULFI and DASH was performed to determine a preliminary assessment of reliability ($n = 14$ subjects with 28 responses) and responsiveness ($n = 5$ subjects with 10 responses). In addition, the capacity to

***TABLE 1.* Patient Numbers and Types of Data Collected in Stage 3**

	Patients (n)	*Responses (n^R)*	*ULFI*	*DASH*	*UEFS*
Pilot test–retest	19	38	Yes	Yes	No
Main test–retest	32	64	Yes	Yes	Yes
Responsiveness	24	48	Yes	Yes	No
Additional patients for comparison	64	64	Yes	Yes	No
Total	139	214			
Missing responses	58	78	0	72 (33.6%)	6 (9.4%)
Invalid responses	4	4	1 (0.5%)	3 (1.4%)	2 (3.1%)
Others excluded	5	5	5 from use of ½ marks	0	0
Data analysis total	130*	205*	208	211	63†

ULFI = Upper Limb Functional Index; DASH = Disabilities of the Arm, Shoulder, and Hand; UEFS = Upper Extremity Functional Scale.
*Total $n = 130$ and $n^R = 205$ for ULFI and DASH criterion and construct comparison as nine patients and their responses were excluded.
†Total $n = 31$ and $n^R = 62$ for UEFS as one patient's repeated responses were excluded.

measure floor response was determined in a test sample of asymptomatic ''normal'' subjects ($n = 8$). The floor response is ''0,'' the lowest possible score, while the ceiling response is ''100,'' the highest possible score. This pilot investigation was also designed to ascertain the practicality of the intended sampling methodology and provide data for calculating the required minimum sample size. To account for the potential of missing response items within a dichotomous tool with a single response option and no confirmatory negative, participants in the pilot study were asked to use a ''✔'' or ''X'' to, respectively, indicate their positive or negative response. No missing responses were found.

Sample size required for the reliability phase was determined by calculating the ''power of the sample'' using the Dawson and Trapp Method[50] where required sample size (n) is:

$$n = \frac{\{(Za - Zb) \times SD\}^2}{\{U1 - U0\}}$$

$(U1 - U0)$ = clinically important difference between the means; SD = standard deviation in the population; Za = two tailed; and Zb = lower tail as defined from Tables of significance levels.

The minimum required sample size for reliability was $n = 28$ patients to ensure an 80% confidence level in determining actual change. This estimate converts to a minimum $n = 32$ when an attrition rate of 12.5% is anticipated to provide a pooled sample of $n = 46$ with the pilot sample included. The values compare favorably with the sample size used by Stratford et al.[10] in developing the UEFI where $n = 28$ was estimated for test–retest reliability and $n = 47$ for cross-sectional and longitudinal validity. In determining the required total sample Stratford et al.[10] estimated a minimum of 196 responses for criterion investigation as calculated using Meng's test of significance and solving for n.[51,52] This number compares favorably with the total of 214 responses in this study for comparative investigation of the ULFI and DASH.

Phase ii of Stage 3: the assessment of test–retest reliability of the ULFI, DASH, and UEFS involved a subgroup of subjects ($n = 32$) who completed all three SROMs at baseline then again after 48 hours. No treatment was administered between the two data collection times. A VAS of functional status was used to evaluate if any change occurred between the data collection days[53] with a permitted limit of either ''change = 0,'' or ''0 +/− 1'' as the bounds of acceptance for inclusion.[9] For the DASH and ULFI, the data were combined with the pilot study for the final determination of test–retest reliability.

Phase iii of Stage 3: the assessment of responsiveness and error was determined only for the ULFI and DASH using a subgroup of subjects ($n = 24$) who were retrospectively combined with those of the pilot study ($n = 5$) to provide a pooled sample ($n = 29$). The UEFS was not tested in this phase as test–retest reliability and construct validity indicated that the psychometric properties of the DASH were consistently more favorable, supporting previous upper extremity SROM research.[7,10,54] Responsiveness evaluation requires that some sort of change has occurred and can be verified in some way so that a patient's score on any SROM can be tested against this change to determine if a true response has occurred.[7,9,11] In this study, the model proposed by Beaton et al.[9] for the DASH was followed as the standard where the external inclusion criterion of ''known group difference'' was natural healing times from the initial to final test periods. These natural times included baseline measures of upper limb status that were postoperative, acute posttrauma with fracture or grade 2+ ligament sprain, or initial rehabilitation commencement with a VAS as the criterion measure for overall status. After the completion of a six- to eight-week rehabilitation program the subjects again completed the measures. In addition, the use of a distribution-based method was selected with the known and accepted minimal detectable change (MDC) for the DASH of 10.7% being the criterion.[8,9,29,42,55]

Phase iv of stage 3: the assessment of criterion validity (with the DASH as the criterion measure), construct validity (through distributional analysis, ''known group'' difference and ''response severity order''), and internal consistency required additional subjects ($n = 64$) to ensure adequate sample size. These subjects completed the ULFI and the DASH concurrently on one occasion to give the final pooled sample ($n = 214$).

The data collected from Stage 3 were analyzed for practical characteristics and the accepted psychometric properties described previously.[8–10,42,43] Data distribution was also analyzed using within-limb grouping[9] where each tool's responses were grouped into 10 categories that included the floor and ceiling score and eight even whole number groups with analysis made for all data, the whole limb, and the three within-limb categories. The practical characteristics analyzed include being self-administered, brevity in terms of length as well as time for completion and scoring, application across a variety of conditions and disease severity levels, ease of understanding and relevance to specific populations and conditions.[8,10,35,43] To present a summarized overview of all characteristics a newly formed ''measurement of outcome measures'' dichotomous tool is introduced that provides an indicator of a SROM's potential value through a ''rule of thumb'' method from a summated percentage score.[25]

Test–retest reliability was assessed for all three SROMs using Type 2,1 intraclass correlation coefficients (ICCs) and the corresponding two-sided

95% confidence interval (CI)[56] to provide an estimate of how closely the numeric scores were to each other (concordance), considered a stronger statistic for the description of reliability.[57–59] Data were analyzed from the respective subgroup participants who completed two separate measures 48 hours apart during a period of no treatment. All subjects completed a participants' details form, the ULFI, DASH, UEFS, the SF-36, and an 11-point VAS for functional status on initial testing. This was repeated on the second occasion with the details form replaced with the MC-13 and a further 11-point VAS scale for feedback in four areas: completion ease, confusion, explanation, and the SROM's reflection of their condition. To compensate for any change between the test times, scores were analyzed with respect to the VAS overall status scale.[53] The values on each occasion were categorized into a permitted limit of either "change = 0," or "0 +/ − 1" as the bounds of acceptance for inclusion[9] with analysis made based upon the sample size exceeding that required for power.[50]

Measurement error was determined by calculating the standard error of the measurement (SEM) and MDC at the 90% and 95% CI levels. There are two methods to achieve these values that are dependent on the use of either the Cronbach alpha (CA) coefficient for internal consistency or the test–retest reliability coefficient (Rxx) with both coefficients having a range capacity of 0 through 1. It is argued that the CA method is preferred for a sample size less than 300 where research is clinically based because the CA value moves away from longitudinal stability as the source of the variance and favors a cross-sectional strength analysis for an instrument where high correlation is present between items.[57,60–62] However, since the test–retest sample size exceeded the required $n = 28$ minimum for power and the total sample size for criterion comparison was $n = 214$, which approaches the estimated size of 300 where Rxx will approximate CA, the former was selected as the critical variable for calculating SEM. This method was the same as that advocated by Michener and Leggins[7] and used by Beaton et al.[9] and Stratford et al.[10] in development of their upper limb tools and further supported in SROM investigations of the lumbar spine by Davidson and Keating[11] and the lower limb by Binkley et al.[63] where

$$\mathrm{SEM} = \mathrm{SD}_{100}(\text{at base line}) \times \sqrt{(1 - \mathrm{Rxx})}$$

SEM subsequently provides MDC_{90} and MDC_{95} where the tabled Z values are, respectively, 1.65 and 1.96 at the 90% and 95% CI levels. This calculation can be interpreted as the magnitude of change, expressed in either scale points or on a percentage basis, required to be 90% or 95% confident that the observed change is real and not measurement error.[64]

$$\mathrm{MDC} = \text{Z value} \times \sqrt{2} \times \mathrm{SEM}$$

Criterion validity was determined by concurrent comparison of the respective questionnaires for the sampled group with the total available data pool minus any incomplete responses. A Pearson coefficient was used to determine the correlation between the total score of each tool.

Construct validity was investigated by three methods.

The distributional analysis of responses was performed to determine the presence of a Gaussian or "bell shape" curve, skew or bias toward a floor or ceiling effect that included within-limb components and a comparison of proximal versus distal means, the latter expected to be higher thus showing greater functional loss.[7,9] The responses were considered in 10 categories with a single category for each of floor and ceiling scores, respectively, Categories 1 and 10. The remaining eight categories were whole number groups from the individual scales: DASH, groups of 15 points and a final group of 14 points; ULFI, groups of three points; and UEFS, groups of eight points and a final group of seven points. In this way distribution can be related to an approximated normal or "bell curve" indicating Gaussian distribution. Data response and distribution within the four kinetic chain components of the limb were also investigated: 1) distal being the wrist and hand, 2) central being the forearm and elbow, 3) proximal being the shoulder and upper arm, and 4) general being those conditions that affected the whole limb, e.g., complex regional pain syndrome and lymphedema. This method indicates the presence or potential for skew or bias to a component of the limb via comparison of the distribution, both range and type, as well as the proximal versus distal means.

The "known group" method compared initial and final scores from a period of known natural healing with expected improvement. A simple difference t-test at a CI of 0.95 comparing these prospective subgroup measurements was used.[9]

Response severity order is a new concept, which ranks the mean item score to reflect the order of severity or incidence of impairment, and is determined by calculating the mean score for the individual items within the total data pool. If an item was marked as a maximum score by every respondent, then it would score a "floor effect" (1.0). Conversely, an item never marked would score "ceiling effect" (0.0) with item redundancy present in either case. For the ULFI, the score is the mean of all scores as the maximum value for the scale is 1. For the DASH, the score is calculated by obtaining the mean, subtracting "1," as the true measure is on a scale of four being from 1 to 5, then dividing by four. For the UEFS, the score is achieved by dividing the mean by the item

scale range of 10. Analysis can be made from ranking of the mean score for each item, which reflects severity order rather than simply the order in which it occurs; the total range of the item means as a percentage value calculated from ''1—mean value;'' and the number of 20% increment categories that are represented.[65] Severity scores will indicate the proportion of responses of the total data pool that were made for any item. Items with lower mean values, being a higher percentage score will indicate higher impairment and be the last items indicated, while those with lower percentage scores will be first and demonstrate sensitivity to initial impairments in function and HRQOL.

Missing responses are anticipated in all SROMs being noted as ''not applicable'' or left blank in Likert format tools.[7,9,29] There is an acceptable limit of up to 10% of the total available question items. If the number of missing responses exceeds this limit the questionnaire is considered invalid.[43,65] The missing response score is accounted for by the addition of the mean of the remaining item scores for each missed response and is used to calculate the final score—as is the case for both the DASH and UEFS. In dichotomous tools two response boxes may be used to account for a confirmatory negative, but with a single response option the incidence of missing responses should be accounted for in either the development and validation stage or by further investigation using the negative response option.

Responsiveness, defined as ''the ability to detect meaningful change over time when it has occurred''[66,67] has also been referred to as ''sensitivity to change.''[45] It is not a fixed property of an instrument and consequently argued as not always being the most critical.[9] The essential statistics used are standard response mean (SRM) and effect size (ES).[43,68–70] These are defined as:

1. SRM: mean change divided by the SD of change scores.[33,71,72]
2. ES: the mean change divided by the SD of baseline scores.[32,73]

Internal consistency was assessed through the use of CA for the full available sample of SROMs.

Practical characteristics are considered within the context of four of the eight essential areas[35] as noted in Table 7. Each tool considered in this study already had four of these requirements established, these being self-administered, applicable across a variety of conditions, disease severity levels, and relevant to specific designated populations. Of the remaining four characteristics, consideration of tool length was evident from face validity while completion and scoring times were assessed by measuring elapsed time for each tool in a subsample of $n = 20$ subjects. The ease of understanding was determined from the four aspects of the 11-point VAS evaluation scale questionnaire administered during test–retest reliability assessment.

A summarized consideration of the 25 essential methodological, practical, distributional, and general characteristics of each tool, with two additional supplementary characteristics identified, has been made and represented in the ''measurement of outcome measures'' dichotomous tool in Table 7. This tool can be used by the clinician and researcher to quickly ascertain if the required characteristics of any SROM are represented. The summated score, a percentage value, is used for direct comparison to assist determining the recommendations of this study.

RESULTS

The study sample provided 214 responses from 139 subjects for both the ULFI and DASH with nine subjects and their responses being excluded providing a final total of 205 responses from 130 subjects. The UEFS had 64 responses from 32 subjects with one subject and his or her responses excluded to provide a total of 62 responses from 31 subjects (Table 1). The demographic data on all subjects are detailed in Table 2.

Data distribution demonstrated substantial variance in all values for the three SROMs. The DASH and UEFS demonstrated a floor response for one subject but no ceiling response. The ULFI demonstrated a floor response for three subjects and ceiling response for one subject. Maximum scores were 84.5% for the DASH, 73.7% for the UEFS, and 100% or ceiling for the ULFI (Figure 2). The ULFI and DASH demonstrated relatively normal distribution for the data pool as well as for the categories of proximal, central, and general upper extremity injuries. The UEFS demonstrated relatively normal distribution for the central category, but was inconsistent in distribution in the other subcategories and for the pooled data. No tool demonstrated a normalized distribution for the distal region, though the ULFI and DASH showed a

TABLE 2. **Subject Demographics**

Subjects	*n = 139*
Age	48.36 yrs, SD 15.60 yrs
Gender	Male 46%, female 54%
Employer	Government 9.9%, private sector 67.7%, military 11.1%, not stated 2.4%, retired 8.9%
Condition duration	All data: 114.2 wks SD = 415.2; with removal of three outliers: 24.5 wks, SD = 28.8
Dominance	Left hand 21%, right hand 77%, nondominant 2%
Work status	Employed 60.8%, unemployed 39.2%
Injury at work	Yes 40%, No 56%, Unsure 4%
Workers' comp	Yes 19.5%, No 80.5%

SD = standard deviation.

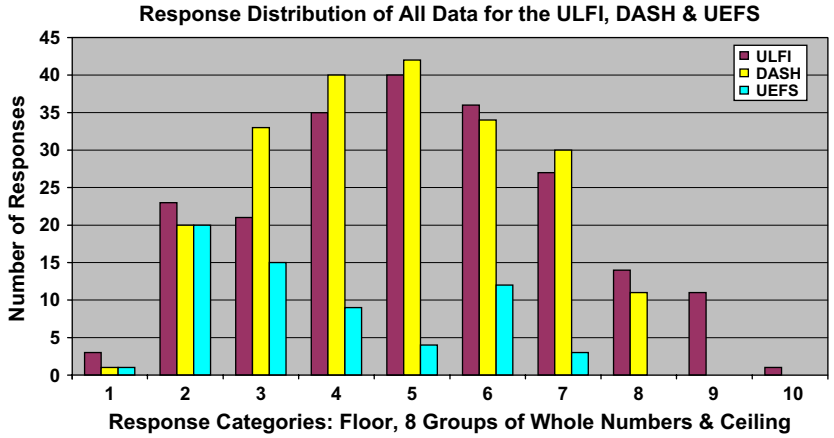

FIGURE 2. *Distribution of all Self-report Outcome Measure (SROM) data in 10 categories. Mean values all data: Upper Limb Functional Index (ULFI) = 44.8; Disabilities of the Arm, Shoulder, and Hand (DASH) = 41.7; Upper Extremity Functional Scale (UEFS) = 26.9. Category 1 = minimum or floor value (0%) and Category 10 = maximum or ceiling value (100%). Categories 2–9 = groups of whole numbers for each scale: ULFI = three points per group: Category 2 = (1–3); 3 = (4–6) … 9 = (22–24). DASH = 15 points per group: Category 2 = (31–45); 3 = (46–60) … 9 = (136–149). UEFS = 10 points per group: Category 2 = (1–8); 3 = (9–16) … 9 = (73–79).*

tendency toward this (Figure 5). The DASH and UEFS pooled data demonstrated a skew tendency toward lower scores, reflecting lower levels of impairment and a limitation on the degree of maximal measurement capacity. The ULFI showed no skew tendency and measurements in all categories (Figure 2). The DASH consistently measured less than the ULFI with a maximum capacity in the eighth category in all subgroups except central, which was at the seventh, with the ULFI measuring, respectively, in the eighth and ninth categories for the same subgroups (Figures 3, 4, and 6).

Test–retest reliability was assessed for all three SROMs from subjects who reported no change in their functional status between test days (change score = "0") where the ICC (2,1) with two-sided 95% CI is 0.98 for the DASH, 0.96 for the ULFI, and 0.92 for the UEFS. This degree of reliability reduced when tolerance was increased to "change = 0 +/ − 1": the DASH mildly decreased to 0.95, the UEFS was unchanged at 0.92, and the ULFI reduced to 0.90 (Table 3). The SF-36 demonstrated that the general health of the participants was comparable to that of the Australian population and showed no change in its reliability of 0.89 at the different change score levels.

Measurement error was determined from SEM (95% CI) and MDC_{90} for each SROM, being, respectively, for the DASH 2.84% and 6.63%, for the ULFI 4.50% and 10.50%, and for the UEFS 5.51% and 12.86%. Higher values were determined for the MDC_{95} by means of the relevant tabled *Z* values (Table 4).

Criterion validity was determined by assessing the relationship between the three SROMs, which were completed concurrently. There was high correlation between the ULFI and DASH with a Pearson's coefficient of 0.87, while with the UEFS the correlation was 0.78 and 0.77, respectively (Table 5).

Construct validity was investigated in three ways. Distribution analysis was demonstrated as relatively normal with limited or no "ceiling" or "floor" effect found for either the ULFI or the DASH. The UEFS did not demonstrate this normal or Gaussian distribution capacity (Figure 2). Further analysis of distribution in the areas of proximal versus distal, general, and then central components of the limb indicates a normalized distribution tendency for the ULFI and the DASH, but again not for the UEFS (Figures 3–6). However, both the ULFI and UEFS had a higher distal mean while the DASH did not. The "known group" method compared baseline and follow-up scores where a simple difference t-test exceeded the 0.95 level supporting the construct validity of the ULFI and DASH. The response severity order (Table 6) has a reference marker at each of the five separate 20% increments. The ULFI has items in each category with a range of 77.0% from 15.5% to 92.5%; the DASH registers in the final four categories only with a range of 53.7% from 26.9% to 80.6%; and the UEFS registers in the final three categories only with a range of 28.8% from 55.5% to 84.3%.

Missing response analysis of the ULFI indicates only one questionnaire (<0.5%) affected in all subjects as there was no provision of a confirmatory negative response. However, 48% of respondents used a negative indicator, such as an "X" as opposed to a "✓", within the box. The DASH had 34% missed responses from 19 different item-questions and question 21, pertaining to "disability with sexual activity," accounted for 27.5% of all missing responses, while question 18, "recreational activities with some force," accounted for 8.5%. A further 1.7% of DASH questionnaires had more than three missed responses

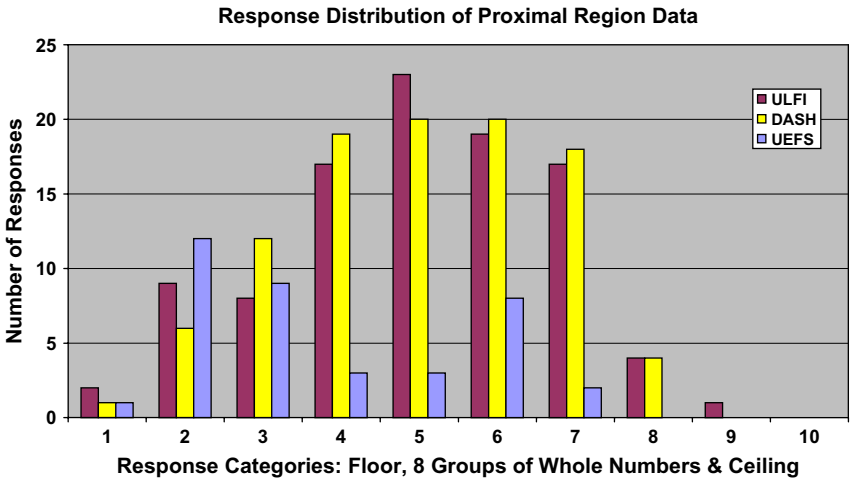

FIGURE 3. *Data distribution for the proximal region: shoulder and upper arm. Mean values proximal region: Upper Limb Functional Index (ULFI) = 43.7; Disabilities of the Arm, Shoulder, and Hand (DASH) = 44.6; Upper Extremity Functional Scale (UEFS) = 27.1. Category 1 = minimum or floor value (0%) and Category 10 = maximum or ceiling value (100%). Categories 2–9 = groups of whole numbers for each scale: ULFI = three points per group: Category 2 = (1–3); 3 = (4–6) … 9 = (22–24). DASH = 15 points per group: Category 2 = (31–45); 3 = (46–60) … 9 = (136–149). UEFS = 10 points per group: Category 2 = (1–8); 3 = (9–16) … 9 = (73–79).*

and were excluded from data analysis. The UEFS had 9.4% missed responses affecting four item questions. Of these 3.1% exceeded one missing response and were excluded.

Responsiveness to change using longitudinal measures of disease impact for SRM and ES are shown in Table 4. These values were determined only for the DASH and ULFI from the subgroup, $n = 24$, with the cut-off criteria of 10.7% or 12.8 DASH points.[9,29]

Internal consistency using CA was determined for each SROM with the ULFI and UEFS at 0.89 and the DASH at 0.96 (Table 4).

Social desirability, measured with the MC-13, was found to have no effect on any characteristic or psychometric property of any of the three tools.

Practical characteristics of the three tools are summarized in a dichotomous descriptive form within Table 7. All three tools demonstrated satisfactory self-administration, application across diverse conditions and severity levels, and were relevant to working populations. The ULFI demonstrated all eight characteristics for clinical practicality exceeding the UEFS with seven and the DASH with five. The completion time reflected the tool length. The DASH at four pages required on average 6.5 minutes, the UEFS at a half page length 1.5 minutes, and the ULFI at one page 2.5 minutes. Subsequent scoring time reflected tool format more so than length with the DASH and UEFS requiring substantially longer and the use of a computational aid in the presence of missing responses. They ranged, respectively, from 2 to 6 minutes for the DASH and 20 seconds to 2 minutes for the UEFS. The ULFI required 20 seconds to score and no computational aids. Use of paired t-tests showed a statistical difference between the completion times, scoring times, and combined times of the DASH and both the UEFS and ULFI and a difference in completion times only between the UEFS and ULFI, but this was not significant. Ease of understanding scored similarly for all three tools on the VAS evaluation scale questionnaire with no significant difference found between any of the four assessed components. The summary of 25 essential and two additional characteristics presented in the "quantifying the measurement of outcome measures" (Table 7) scored the ULFI at 96% while the DASH and UEFS scored 68%.

TABLE 3. **Summary of the Upper Limb Self-report Outcome Measures (SROMs) Test–Retest Reliability Intraclass Correlation Coefficient (ICC) Values at Confidence Interval (CI) = 0.95**

Change (between First and Second Test) on VAS	*n*	*DASH*	*ULFI*	*n*	*UEFS*
Change = 0	28	0.9791	0.9567	22	0.9150
Change = 0 +/− 1	41	0.9527	0.8983	29	0.9156
All data (at any change level)	47	0.9011	0.8950	32	0.8978

ULFI = Upper Limb Functional Index; DASH = Disabilities of the Arm, Shoulder, and Hand; UEFS = Upper Extremity Functional Scale; VAS = Visual Analogue Scale.

DISCUSSION

This prospective study investigated the ULFI as a new regional upper limb SROM tool analyzing its concurrent performance compared to advocated measures, the DASH and the UEFS. The results

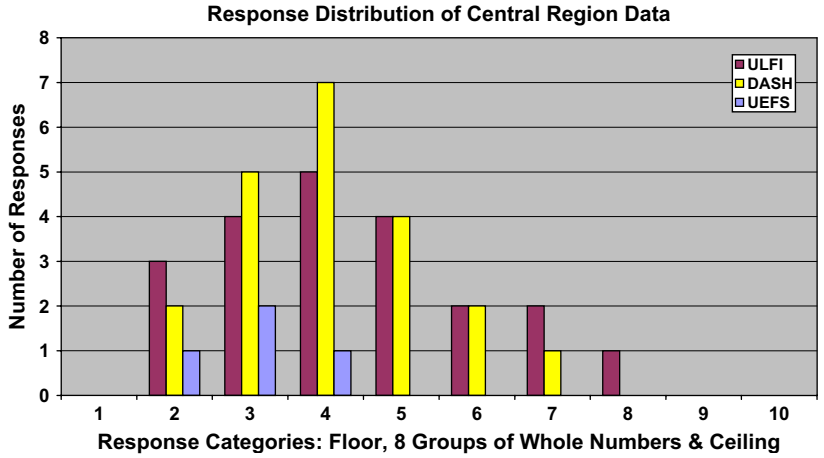

FIGURE 4. *Data distribution, for the central region: forearm and elbow. Mean values central region: Upper Limb Functional Index (ULFI) = 37.4; Disabilities of the Arm, Shoulder, and Hand (DASH) = 32.8; Upper Extremity Functional Scale (UEFS) = 18.4. Category 1 = minimum or floor value (0%) and Category 10 = maximum or ceiling value (100%). Categories 2–9 = groups of whole numbers for each scale: ULFI = three points per group: Category 2 = (1–3); 3 = (4–6) … 9 = (22–24). DASH = 15 points per group: Category 2 = (31–45); 3 = (46–60) … 9 = (136–149). UEFS = 10 points per group: Category 2 = (1–8); 3 = (9–16) … 9 = (73–79).*

demonstrated that the ULFI has both methodological and practical characteristic advantages for measuring upper extremity disorders. Reliability, internal consistency, criterion and construct validity, sample size, error measurement, and responsiveness have been demonstrated to be comparable to the properties of the DASH and UEFS. The ULFI exceeds the distribution range, has a lower missing response rate, and has an improved response severity range and distribution as compared to the DASH and UEFS. The improved practical aspects of tool use in a clinical setting through speed of completion and scoring are further demonstrated. The summation of these important factors, as shown in Table 7, support and validate these advantages, which, along with the strengths of the study and the ULFI's future potential to integrate with other similarly designed regional tools, are discussed in further detail.

The **sample size** in each aspect of the investigation is justified from the power analysis and compares favorably with that of Beaton et al.[9] at $n = 200$ participants and $n = 56$ for test–retest, and the projected numbers calculated by Stratford et al.[10] of $n = 196$ for a statistically valid investigation of psychometric properties for a new upper limb tool and $n = 47$ subjects for test–retest reliability.

Test–retest reliability demonstrated by the ULFI indicates that this tool can consistently monitor physical function and symptoms in proximal, central, and distal disorders and in conditions that affect the entire limb as demonstrated by the multiple musculoskeletal conditions represented in the sample. Furthermore, the increased focus on a reduced tolerance at the change level of "0" resulted in small or no reliability improvement, respectively, for the DASH and UEFS compared to the marked increase for the ULFI from 0.90 to 0.96.

Validity investigated for each tool provides particularly useful data as all SROMs were completed simultaneously and supplements the existing body of knowledge on DASH and UEFS characteristics as well as the comparative performance of the ULFI.

Face and content validity were satisfactorily determined within the initial item generation and reduction stages of tool development.

Construct validity for both the ULFI and DASH were supported by the distributional analysis of both the full data set and the within region subgroups

TABLE 4. **Methodological Characteristics for DASH, ULFI, and UEFS**

Tool	*SD_{100}*	*ICC*	*SEM*	*MDC_{90}*	*MDC_{95}*	*ES*	*SRM*	*ES/SRM %*	*C-Alpha*
$DASH_{100}$	19.67%	0.9791	2.84%	6.63%	7.87%	1.4077	2.1820	35.48	0.9633
$ULFI_{100}$	21.61%	0.9567	4.50%	10.50%	12.47%	1.2841	1.8674	31.24	0.8893
$UEFS_{100}$	18.97%	0.9156	5.51%	12.86%	15.27%	Unavailable	Unavailable	Unavailable	0.8891

SD = standard deviation at baseline score on a 100% scale; ICC = intraclass correlation coefficient for test–retest reliability (one-way random effects model at a 95% CI); SEM = standard error of the measurement; MDC_{90} = minimal detectable change at a 90% CI; MDC_{95} = minimal detectable change at a 95% CI; ES = effect size to measure responsiveness; SRM = standard response mean to measure responsiveness; ES/SRM % = percentage difference between SRM and ES for the same change measurement on the same subjects; C-Alpha = Cronbach's alpha to measure internal consistency; ULFI = Upper Limb Functional Index; DASH = Disabilities of the Arm, Shoulder, and Hand; UEFS = Upper Extremity Functional Scale.

TABLE 5. **Intraclass Correlation Coefficient (ICC) Value for Criterion Comparison between the ULFI, DASH, and UEFS**

	DASH	*ULFI*
ULFI	0.8665—high	–
UEFS	0.7675—medium	0.7759—medium

ULFI = Upper Limb Functional Index; DASH = Disabilities of the Arm, Shoulder, and Hand; UEFS = Upper Extremity Functional Scale.

as each approached a normalized bell curve in all circumstances though the impairment range of the ULFI was larger, specifically at the higher levels. The UEFS was poorly distributed and combined with concerns of inconsistency in other psychometric properties raises doubts on its suitability for clinical or research circumstances. Consequently, other tools, such as the UEFI, which used the UEFS as the primary criterion standard in their development, may also be affected. A regional tool should demonstrate a higher distal than proximal mean as hand injuries cause greater functional impairment than those of the shoulder.[7,9] This was found for ULFI and UEFS but not for the DASH, which is contrary to the findings of Beaton et al.[9] and provides an area of clarification for further research. The ''known group'' method of construct validity determination supported the conclusions of the distributional analysis for both the DASH and ULFI. The introduction of ''response severity order'' further supports the ULFI as the preferred tool as it demonstrated a total range of 77.0% that was, respectively, 1.4 and 2.7 times greater than the DASH and UEFS and covered all five 20% incremental categories compared to four categories for the DASH and the three for the UEFS. It is particularly important that the initiation of the lower impairment mean for the ULFI occurs at 15.5%, 11% sooner than the DASH at 26.9% and 40% sooner than the UEFS at 55.5%. Table 6 can be used to consider the distribution of items with the ULFI approximating a normal Gaussian distribution with minimal skew effect compared to that of the DASH and UEFS. A direct comparison of the order and percentage scores of similar item variables can also be made with some similarity across all three SROMs (transport, jars, and home duties) or between two SROMs (sleep and writing for the DASH and UEFS; light activity and utensil/knife use for the ULFI and DASH; for example, holding a jug for the ULFI and UEFS) and distinct variation with other items. Response severity order offers an alternative method of analyzing both the construct validity and overall performance of a tool through the ability to apportion levels of impairment via item description across a full distribution range. This provides greater scope of impairment measurement and markedly higher capacity to indicate change and the ULFI is shown to be the preferred tool in both these areas.

The determination of **criterion validity** demonstrated a strong relationship between the ULFI and both the DASH and UEFS, which exceeded the minimum requirements.[45,51,74,75]

The analysis of **data distribution** range supports the preference for the ULFI. Measurement of impairment range was greater for the ULFI with a more even distribution pattern overall and within the limb. Division of data distribution into 10 whole number groups enabled a direct comparison between the three SROMs with the presence of a floor measurement capacity demonstrated. The ULFI and DASH had relatively normalized distribution for all data and the within region components providing additional support for their construct validity, though the DASH was mildly skewed toward higher

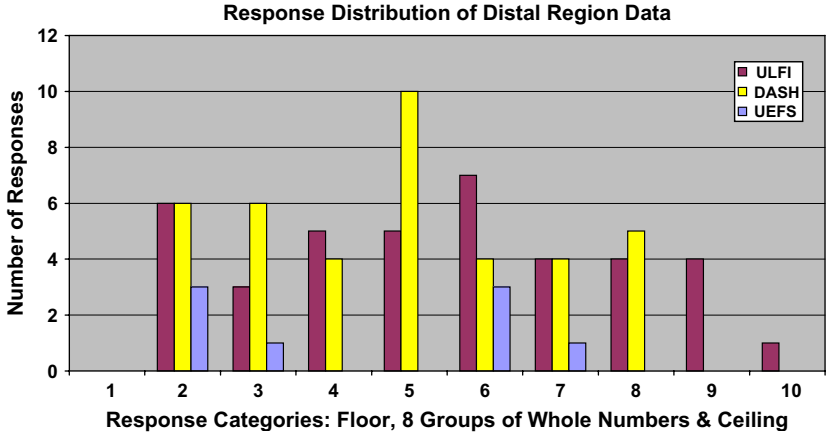

FIGURE 5. Data distribution, for the distal region: wrist and hand. Mean values distal region: Upper Limb Functional Index (ULFI) = 49.6; Disabilities of the Arm, Shoulder, and Hand (DASH) = 42.6; Upper Extremity Functional Scale (UEFS) = 33.1. Categories 1 = minimum or floor value (0%) and Category 10 = maximum or ceiling value (100%). Categories 2–9 = groups of whole numbers for each scale: ULFI = three points per group: Category 2 = (1–3); 3 = (4–6) … 9 = (22–24). DASH = 15 points per group: Category 2 = (31–45); 3 = (46–60) … 9 = (136–149). UEFS = 10 points per group: Category 2 = (1–8); 3 = (9–16) … 9 = (73–79).

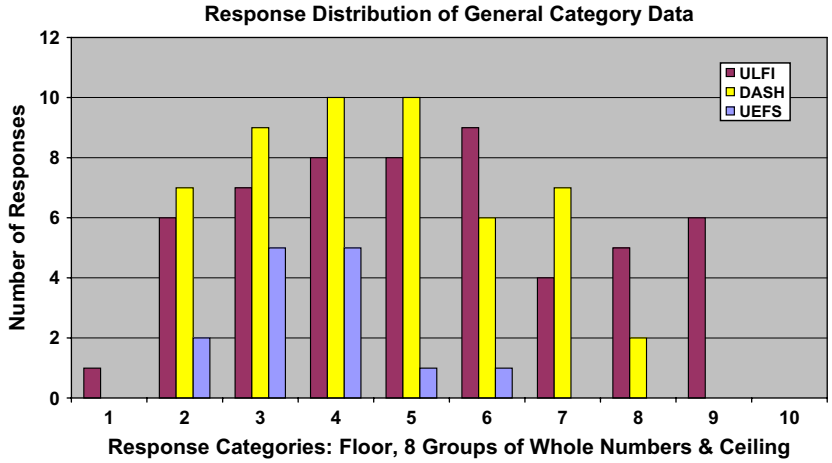

FIGURE 6. *Data distribution, for the general category: conditions affecting the whole limb. Mean values general category: Upper Limb Functional Index (ULFI) = 46.4; Disabilities of the Arm, Shoulder, and Hand (DASH) = 39.1; Upper Extremity Functional Scale (UEFS) = 25.3. Categories 1 = minimum or floor value (0%) and Category 10 = maximum or ceiling value (100%). Categories 2–9 = groups of whole numbers for each scale: ULFI = three points per group: Category 2 = (1–3); 3 = (4–6) … 9 = (22–24). DASH = 15 points per group: Category 2 = (31–45); 3 = (46–60) … 9 = (136–149). UEFS = 10 points per group: Category 2 = (1–8); 3 = (9–16) … 9 = (73–79).*

function or lower scores with an upper limit of 84.5%. This reinforced similar score limitation findings in recent DASH research.[9] The UEFS had a more severe low score skewing tendency reflecting the original authors' findings of floor but no ceiling response capacity.[14] An excess tendency toward a ceiling effect is detrimental,[11] however, a scale is unsuitable if it is unable to detect an impairment level that approaches or achieves the maximal score.

The **measurement error** was determined by means of Rxx at "change = 0" for the VAS status in the formula for SEM and consequently the MDC values. The Rxx value was used in error value calculation as it reflects reliability of use of the entire scale[9–11,14] as opposed to CA,[43,63] which reflects consistency across the items of the scale. The error estimates are thus reduced enabling tool clinical relevance to be increased on the order of 40% for the ULFI, 30% for the DASH, and 20% for the UEFS. This results in the values determined at "change = 0 +/− 1" being directly comparable to those of earlier research with an MDC_{90} of 10.7% for the DASH[9,29] and an SEM of 5.6% for the UEFS[10]. By reducing the level of tolerance for Rxx to "change = 0" the measures of SEM and MDC have improved clinical relevance indicating that they are not stable measurement properties but a reflection of the formula used and thus the study design. The critical importance of these variables is that of assisting the decision process of intervention effectiveness in individual patient care by using the minimal change required to indicate, with 90% confidence, that real change has occurred and not simply measurement error.

The ULFI demonstrated "high" **responsiveness** levels by using Beaton et al.'s[9] and McConnell et al.'s[29] validated standard of "known group difference" as opposed to "transitional scales,"[9] which are a retrospective criterion-based method[68,69,76,77] with questionable accuracy.[41,78–80] A comparison of the variation in the range and magnitude of these responsiveness results concurred with those of Beaton et al.[9] and Wright and Young[66] supporting the criticism of "choice of statistic" for responsiveness description. The results varied markedly for the same change in the same patients with value variation between SRM and ES of 35.5% for the DASH, comparable to the 33% variation noted by Beaton et al.[9] and 31.2% for the ULFI as shown in Table 4. In both circumstances, the values span different classifications of "moderate" and "high," within the same guideline of "Cohen's rule of thumb."[74] This supports the abandoning of this rule as the variation or quality of the classification is due solely to the statistic chosen. The more conservative estimate based on ES is the preferred value.[9,43]

The ULFI had less than 0.5% **missing responses** in all data compared to the UEFS with 9.4% and the DASH with 34% where the single item on sexual function accounted for over one quarter of these. Clinicians using the UEFS and DASH will require additional time and computational aids, which will affect their choice of tool, particularly when coupled with concerns on psychometric properties. With the ULFI it is assumed that, following the instructions, an unmarked box indicated a negative response—though some participants choose to use an "X" as a negative and a "✓" as a positive. From work on similar dichotomous tools, such as the Roland Morris Questionnaire, where a confirmatory negative response option was used, missing responses were noted at 9% being predominantly elderly and non-English speaking participants.[81] Having tested for and finding no missing responses in the pilot sample of $n = 38$, knowing the average age is 48 years, that

TABLE 6. **Comparison of Item Average Severity for the DASH, ULFI, and UEFS**

Percent Markers	*Mean %*	*#*	*DASH Items*	*ULFI Items*	*Mean %*	*#*	*UEFS Items*	*Mean %*	*#*
	80.6%	2	Write	Assistance with washing, hygiene	92.5%	12	Pickup small objects with fingers	84.3%	D
				Appetite affected	87.6%	8	Writing	80.3	B
				Stay at home most of time	81%	1			
80%									
	79.8%	3	Turn a key	Eating: using utensils	79.3%	20	Washing dishes	74.0%	H
	78.4%	17	Recreation activity little effort	Drop things—minor accidents	70.7%	22	Driving > 30 minutes	72.5%	E
	75.8%	16	Knife use to cut food	Transport independence	69.8%	16	Carry jug from fridge	71.4%	G
	75.8%	20	Transport needs	Walking/normal recreation/sport	69.5%	9	Opening a door	67.3%	F
	74.5%	4	Prepare a meal	Arm in shirt sleeve/dressing	67.2%	17	Sleeping	66.6%	A
	72.4%	26	Pins and needles	Writing/using keyboard or mouse	67.2%	18			
	71.3%	21	Sexual activity	Hold or moving dense objects	66.1%	21			
	69.2%	13	Wash, blow dry hair	More irritable/bad tempered	63.5%	14			
	68.0%	9	Make a bed	Painful almost all the time	62.1%	6			
	63.5%	15	Put on a pullover or sweater	Home/family duties and chores	61.5%	10			
	60.9%	29	Sleeping in the last week						
60%									
	58.0%	10	Carry shopping bag/briefcase	Tend to rest more often	58.1%	4	Opening jars	55.5%	C
	57.2%	5	Push open a heavy door	Try get others to do things	54%	5			
	55.7%	28	Stiffness	Difficult button key coins taps	50%	24			
	54.2%	22	Interferes social/family activity	Sleep less well	42.5%	11			
	52.5%	6	Object to shelf above head	Do things at/above shoulder	40.8%	19			
	51.0%	1	Open a tight or new jar						
	50.8%	14	Wash your back						
	48.6%	24	Pain						
	46.6%	12	Change a light-bulb overhead						
	46.3%	27	Weakness						
	45.9%	23	Interferes work/daily activity						
	45.5%	8	Garden or yard work						
	43.1%	7	Heavy household chores						
	40.1%	25	Pain on doing specific activity						
40%									
	38.5%	11	Carry a heavy object (>5 kg)	Change positions frequently	35.6%	2			
	35.5%	30	Feel less capable or useful	Regular daily activity work/social	32.2%	13			
	33.7%	19	Recreation activity + free arm mvt	Open, hold, press, or push	32.2%	25			
	26.9%	18	Recreation activity + force/impact	Feel weaker or stiffer	24.1%	15			
				Lifting and carrying	23.6%	7			
				Use other arm more often	20.7%	23			
20%									
				Avoid heavy jobs	15.5%	3			

Mean scores are represented as percentage values to indicate item average scores. The percentage values reflect the degree of impairment required and are determined by {(1 − mean) × 100}. Items at the top of the list have a higher percent value being marked less often and consequently indicate greater impairment. Items at the bottom of the list indicate less impairment and are more often marked first. ULFI = Upper Limb Functional Index; DASH = Disabilities of the Arm, Shoulder, and Hand; UEFS = Upper Extremity Functional Scale.

48% of respondents used a negative indicator and none of these had missing responses, it would be assumed that missing responses for this tool would be low. The decision not to include a confirmatory negative option was made from peer and patient feedback and ergonomist's advice. The single box response option improved tool efficiency and patient receptivity while additional boxes or altering design to provide this option made the tool appear obtrusive. However, without the inclusion of a confirmatory negative option the assumption of minimal missing responses cannot be confirmed and will require clarification through further investigation.

The presence of **internal consistency** was determined as satisfactory by the use of the CA coefficient with identical values of 0.89 for both the ULFI and the UEFS. This is marginally higher than the 0.85 found by Pransky et al.[14] and mid-range of 0.83–0.93 from subsequent UEFS researchers.[10] However, the DASH, with a CA of 0.96 exceeds the recognized 0.95 upper limit indicating that too many items are too similar or testing the same construct,[82] which

TABLE 7. **Quantifying the Measurement of Outcome Measures**

Number #	*Factors Considered*	*Preferred Response*	*ULFI (Gabel, 2006)*	*DASH (Beaton et al.[9])*	*UEFS (Pransky et al.[14])*
	Methodological				
1	Reliability	Yes	Yes	Yes	Yes
2	Error or change scores (MDC, SEM)	Yes	Yes	Yes	Yes
3	Validity	Yes	Yes	Yes	Yes
4	Responsiveness	Yes	Yes	Yes	Yes
5	Internal consistency ($\alpha =$ 0.80–0.95)	Yes	Yes	No	Yes
6	Adequate sample power and size	Yes	Yes	Yes	No
	Practical				
7	Self-administered	Yes	Yes	Yes	Yes
8	Brevity in length	Yes	Yes	No	Yes
9	Short completion time	Yes	Yes	No	Yes
10	Short scoring time	Yes	Yes	No	No
11	Application across conditions	Yes	Yes	Yes	Yes
12	Diseases severity range	Yes	Yes	Yes	Yes
13	Easy to understand	Yes	Yes	Yes	Yes
14	Relevant to working populations	Yes	Yes	Yes	Yes
	Distributional				
15	All categories 0–100%	Yes	Yes	No	No
16	Even or normalized for all data	Yes	Yes	Yes	No
17	Even distribution for each area	Yes	Yes	Yes	No
18	Different distal mean (i.e., hand $\neq$ shoulder)	Yes	Yes	No	Yes
19	No marked ceiling or floor effect	Yes	Yes	Yes	Yes
20	No floor/ceiling skew tendency	Yes	Yes	No	No
	General				
21	Quantitative data obtained	Yes	Yes	Yes	Yes
22	Qualitative data obtained	Yes	Yes	No	No
23	Represents function and HRQOL	Yes	Yes	Yes	No
24	Independently researched	Yes	No	Yes	Yes
25	Independent statistical analysis	Yes	Yes	Yes	Yes
	Total	25	24	17	17
	Percentage rating	100%	96%	68%	68%
Other	*Future implications*				
A	Other regional SROM consistency	Yes	Yes	No	No
B	Global assessment integration	Yes	Yes	No	No

A summary of 25 important factors in regional SROMs as applied to upper limb tools. ULFI = Upper Limb Functional Index; DASH = Disabilities of the Arm, Shoulder, and Hand; UEFS = Upper Extremity Functional Scale; SROM = Self-report Outcome Measure; HRQOL = health-related quality of life; MDC = minimal detectable change; SEM = standard error of the measurement.

implies item redundancy.[7] The developers of the DASH also found the same high value of 0.96[9,29] with other researchers approximating this[83] which is a likely contributor, along with length, completion, and scoring time for the more recent development of the 11-item ''quick DASH.''[8,84]

The **practical characteristics** of the three tools, summarized in descriptive form in the initial section of Table 7, indicate the ULFI as the preferred option. Tool format improves clinical utility through ease and consistency of use due to brevity, speed of completion, a raw score of 25 being rapidly converted to a 100-point scale without computational aids, minimal missing responses, and subsequent reduced scoring time. The ULFI has the additional benefit of the ''PSI'' individual qualitative component and a VAS of overall status on the one page.

The spontaneous use of the ''½'' mark on the ULFI by 3.6% of patients (see Table 1) suggests that allowance for this response option should be considered in future investigations of existing regional SROMs and the development of new tools. This is supported by the findings of Chansirinukor[85] who showed that ''three category options,'' effectively a dichotomous tool with a ''½ mark'' option, provided optimal psychometric results compared to either the Likert or Dichotomous version of the Roland Morris Questionnaire. Such a format would cause minimal compromise to the practical characteristics advantages and provide definitive ordinal data.

Regional SROMs are the preferred choice for current and future musculoskeletal measurement, but the choice of format and tool template design is an area lacking consensus as highlighted in this study. The practical advantages of the ULFI support the use of dichotomous or three-point Likert scales as advocated by several authors.[69,85–87] Longer Likert scale format is often chosen for new tool development due to either assumption[9,23] or previous comparative studies[11,88] that these scales have higher responsiveness. However, this study has shown that the small reduction in responsiveness is outweighed by the increased practicality without loss of the essential psychometric properties. These practical characteristics are necessary in contributing to improved efficiency in an occupational and therapeutic setting. Treating professionals and their patients have less discretionary time compared to researchers in large centers where patient groups are often students, retirees, and the chronically disabled, who are not reflective of the future users of the developed SROM tools.[37] These considerations are particularly important as tools with low practicality will not be readily accepted by clinicians and integrated into daily clinical practice.[7,77,89] The summarized essential and additional characteristics identified and listed in the ''measurement of outcome measures'' in Table 7 support this premise by providing a quantitative ''rule of thumb'' template for researchers and clinicians to analyze and judge any SROM tool's performance. The determined values of 96% for the ULFI compared to 68% for the DASH and UEFS lend further support to the UEFI as the preferred tool.

This study shows that the ULFI is directly comparable to the DASH and UEFS in its methodological properties but its practicality is higher. Both the DASH and UEFS have demonstrated direct comparison capability with condition-specific scales of disease[9,14,90,91] and joint measurement.[8,28,54,55,92–94] These upper limb SROMs traits imply transferability of the ULFI to these established criteria, however, this will require comparative validation with the condition-specific tools in the designated population groups.

Despite the acceptance and support of the DASH and UEFS, criticisms remain. This study independently established the psychometric properties for the DASH. The values for validity and reliability were in agreement with previous studies[9,83] while performance capacity in error measurement and responsiveness were improved,[29] which may be attributed to the younger, healthier, and less chronic noninstitutional population in the study. The shortcomings of the DASH remain its length, complexity, error in scoring, missing responses, and excess internal consistency. These factors weigh strongly and cause guarded use of the tool due to poor clinical utility and item redundancy.

Criticisms of the UEFS include its development, lack of population diversity (using predominantly injured workers), and the large variations in scores for different groups and different test occasions. The current study had greater population diversity, but also demonstrated large variation in the essential psychometric properties of the UEFS when compared to earlier research. It showed range limitation and that the test–retest reliability and internal consistency were only confirmed in the mid-range of previously reported values.[10,14] The variation and lack of consistency between studies reinforce the guarded use of the UEFS due to underlying methodological insufficiency.

Limitations of the Study

The limitations of statistical power due to sample size must be taken into account by future users of this new tool. It is a study of an Australian population analyzing 205 responses from 130 subjects, 75 of whom were included in subgroups to provide subsequent responses on reliability, error scores, and responsiveness. The conclusions from this study cannot be directly considered to have global implications, though sample diversity from the subjects' geographical areas and ethnic background do minimize this. Similarly, this study did not provide

satisfactory indications of population-based normal values by the demographic areas of gender, age, or occupation, nor did it indicate discriminant validity for different levels of work status, personal activity, or symptom duration. As a regional measure, the implications for the application to individual joint or disease conditions can be made due to the criterion validity with the DASH and UEFS that have had such studies completed. However, without specific validation this remains uncertain being an area of further investigation along with specific discriminant validity relating condition-specific variables and the determination of minimal clinically important difference. Furthermore, this tool has not demonstrated the ability to be predictive of future impairment levels or treatment effectiveness leaving the potential for further investigative research.

Strengths of the Study

The use of concurrent completion of the questionnaires enables their direct performance comparison. The positive results for all three tools as regional measures support the ''single kinetic chain'' model. This study confirms and improves existing methodological characteristics for the DASH and UEFS, specifically in the area of test–retest reliability and the subsequent reduction in the error estimate values of SEM and MDC. The assertions that a standardized single method approach for responsiveness calculation should be used[9,11,66] are confirmed supporting the contention that Cohen's ''rule of thumb'' should be abandoned. It also confirms that the personality trait of social desirability has no influence on the response capacity of regional SROMs supporting the current practice of not accounting for this variable in daily clinical outcome tool use.

Future Research Implications

This study provides a precedent for future investigations of regional outcome tools, for the lower limb and spine as well as generic tools that consider any musculoskeletal area. The balance must be considered between a tool's clinical utility and the quantitative and qualitative focus using predominantly HRQOL item questions and a patient-centered approach. The new emphasis on outcome SROMs will be investigating tool performance in clinical trials and studies through collaborative investigations for the purposes of database formation, including the ongoing creation of population-based normal values and the integration of Item Response Theory. This process will involve researchers from clinical, academic, and institutional areas as well as the health and medical practitioner to provide a base from which comparative data can be further used.[13,38] Existing format patient surveys as well as newly developed SROM tools, with an appropriate research support base that are peer advocated, will become the primary mechanisms for this process of information and e-format collection for databases and storage.[8,16,49,95,96] The information will be used for evidence-based practice (EBP), validation of intervention to third-party payers, clinical reasoning, and intervention implementation as well as in new ways such as global health management. The future for the field of outcome measures and instrument evaluation is extensive and these tools will continue to play an important role in evaluating the different results, interventions, and outcomes achieved in patient treatments. Traditional users of upper limb SROMs will remain the clinician and researcher, both evaluating the effectiveness of interventions to provide EBP and achieve the best possible outcomes for their patients. However, new users such as governments, professional groups, and insurers are likely to be the strongest advocates for their future requirement as a standard component of clinical care.

CONCLUSIONS

This study achieved its two primary objectives. It validated the ULFI and demonstrated that its essential psychometric properties of reliability, validity, responsiveness, error measurement, and internal consistency approximate or exceed those of the DASH and UEFS. The latter advocated tools are shown to be self-limiting in determining maximum impairment and have practical restraints that affect clinical utility, limitations that may also by implication be present in the UEFI. The ULFI's practical characteristics of brevity, ready transferability to a 100-point scale, ease and rapidity of completion and scoring reinforce methodological consistence while the inclusion of a VAS for functional status and a ''PSI'' increase collected information diversity. The summary quantification of all characteristics supports these advantages. Consequently, the ULFI is advocated over the DASH and UEFS as the preferred regional tool for upper extremity outcome measurement.

REFERENCES

1. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low back pain. Spine. 1983;8:141–4.
2. Wiechman SA, Smith RE, Smoll FL, Ptacek JT. Masking effects of social desirability response on relations between psychosocial factors and sport injuries: a methodological note. J Sci Med Sport. 2000;3:194–202.
3. Reynolds WN. Development of reliable and valid short forms of the Marlowe–Crowne Social Desirability Scale. J Clin Psychol. 1982;38:119–25.
4. Denniston OL, Jette A. A functional status assessment instrument: validation in an elderly population. Health Serv Res. 1980;15:21–34.

5. Friedsam HJ, Martin HW. A comparison of self and physician's health ratings in an older population. J Health Human Behav. 1963;4.
6. Williams RGA, Johnston M, Willis LA, Bennet AE. Disability: a model and measurement technique. Br J Prev Soc Med. 1976;30:71–8.
7. Michener LA, Leggins BG. A review of self-report scales for the assessment of functional limitation and disability of the shoulder. J Hand Ther. 2001;14:68–76.
8. Amadio PC. Outcome assessment in hand surgery and hand therapy: an update. J Hand Ther. 2001;14:63–7.
9. Beaton DE, Katz NK, Fossel AH, Wright JG, Tarasuk V, Bombardier C. Measuring the whole of the parts? Validity, reliability, and responsiveness of Disabilities of the Arm Shoulder and Hand outcome measure in different regions of the upper limb. J Hand Ther. 2001;14:128–46.
10. Stratford PW, Binkley JM, Stratford D. Development and initial validation of the Upper Extremity Functional Index. Physiother Can. 2001;52:259–67, 281.
11. Davidson M, Keating JL. A comparison of five low back disability questionnaires: reliability and responsiveness. Phys Ther. 2002;82:8–24.
12. Fries J, Spitz P, Young DW. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. Rheumatology. 1982;9:789–93.
13. MacDermid JC. Editorial—the outcome issue. J Hand Ther. 2001;14:61–2.
14. Pransky G, Feuerstein M, Himmelstein J, Kratz JN, Vickers-Lahti M. Measuring functional outcomes in work-related upper extremity disorders. J Occup Environ Med. 1997;39:1195–202.
15. Sullivan MS, Shoaf LD, Riddle DL. The relationship of lumbar flexion to disability in patients with low back pain. Phys Ther. 2000;80:240–50.
16. Deyo R. Measuring outcomes of low back pain. In: Musculoskeletal Physiotherapy Australia, 13th Biennial Conference, Sydney; 2003.
17. Hazard RG, Haugh LD, Green PA, Jones PL. Chronic low back pain the Relationship between patient satisfaction and pain, impairment, and disability outcomes. Spine. 1994;19:881–7.
18. Levine DW, Simmons BP, Koris MJ. A self-administered questionnaire for the assessment of severity of symptoms in carpal tunnel syndrome. J Bone Joint Surg Am. 1993;75A: 1585–1592.
19. Katz JN, Fossel KK, Simons BP, Swartz RA, Fossel AH, Korris MJ. Symptoms, functional status and neuromuscular impairment following carpal tunnel release. J Hand Surg. 1995;20: 549–55.
20. Schonstein E, Kenny DT, Maher CG. Workcover's physiotherapy forms: purpose beyond paperwork? Aust J Physiother. 2002;48:221–5.
21. Rogerson S, Workcover NSW. Evidence based guidance material for exercise and activity programs in post acute non-red flag musculoskeletal injuries. In: Musculoskeletal Physiotherapy Australia 13th Biennial Conference, Sydney, Australia; 2003.
22. Stock SR, Streiner D, Reardon R, et al. The impact of neck and upper limb musculoskeletal disorders on the lives of affected workers: development of a new functional status index. Qual Life Res. 1995;4:491.
23. Feise RJ, Menke JM. Functional Rating Index. A new valid and reliable instrument to measure the magnitude of clinical change in spinal conditions. Spine. 2001;26:78–86.
24. Stock SR, Cole DC, Tugwell P, Streiner D. Review of applicability of existing functional status measures to the study of workers with musculoskeletal disorders of the neck and upper limb. Am J Ind Med. 1996;29.
25. Gabel CP. Development and initial validation of a new regional outcome measure: the Upper Limb Disability Questionnaire (ULDQ). Masters by Research Thesis, Charles Darwin University, Faculty of Health Science. Darwin: Charles Darwin University, 2003:200.
26. Salerno DF, Copley-Merriman C, Taylor TN, Shinogle J, Schulz RM. A review of functional status measures for workers with upper extremity disorders. Occup Environ Med. 2002;59:664–70.
27. Devereux JJ, Vlachonikolis IG, Buckle PW. Epidemiological study to investigate potential interaction between physical and psychosocial factors at work that may increase the risk of symptoms of musculoskeletal disorder of the neck and upper limb. Occup Environ Med. 2002;59:269–77.
28. Kirkley A, Griffith S, McLintock H, Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability: the Western Ontario Shoulder Instability Index (WOSI). Am J Sports Med. 1998;26: 764–72.
29. McConnell S, Beaton DE, Bombardier C. The DASH Outcome Measure User's Manual. Toronto, Canada: Institute for Work and Health, 1999.
30. Kirshner B, Guyatt GH. A methodological framework for assessing health indices. J Chronic Dis. 1985;38:27–36.
31. Maklan CW, Greene R, Cummings MA. Methodological challenges and innovations in patient outcomes research. Med Care. 1994;32(suppl):JS13–21.
32. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum. 1985;28: 542–7.
33. Liang MH, Fossel AH, Larson MG. Comparison of five health status instruments for orthopaedic evaluation. Med Care. 1990;28:632–42.
34. George K, Batterham A, Sullivan I. Validity in clinical research: a review of basic concepts and definitions. Phys Ther Sport. 2000;1:19–27.
35. Liang MH, Jerre AM. Measuring functional ability in chronic arthritis: a critical review. Arthritis Rheum. 1981;24:80–6.
36. Patrick DL, Deyo RA. Generic and disease specific measures in assessing health status and quality of life. Med Care. 1989; 27:S217–32.
37. Greenfield S, Nelson EC. Recent developments and future issues in the use of health status assessment measures in clinical settings. Med Care. 1992;30:MS23–41.
38. Hudak PL, Amadio PC, Bombardier C. Upper Extremity Collaborative Group (UECG). Development of an upper extremity outcome measure: the DASH (Disabilities of the Arm, Shoulder, and Head). Am J Ind Med. 1996;29:602–8.
39. World Health Organisation (WHO). ICIDH-2: International Classification of Impairments, Activities and Participation. vol. 2004. Geneva: World Health Organisation, 2001.
40. Stratford P, Gill C, Westaway M, Brinkley J. Assessing disability and change on individual patients: a report of a patient specific measure. Physiother Can. 1995;47:258–63.
41. Westaway MD, Stratford PW, Binkley JM. The Patient Specific Functional Scale: validation of its use in persons with neck dysfunction. J Orthop Sports Phys Ther. 1998;27:331–8.
42. Macdermid J. Support for the DASH. In: Proceedings of the Fall Conference of the American Hand Therapy Association 2001; 2001.
43. Michener L, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: reliability, validity and responsiveness. J Shoulder Elbow Surg. 2002;11:587–94.
44. Guyatt GW, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. CMAJ. 1986;134:889–95.
45. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. New York: Oxford Medical Press, 1995.
46. Bennell K, Bartam S, Crossley K, Green S. Outcome measures in patellofemoral pain syndrome: test retest reliability and inter-relationships. Phys Ther Sport. 2000;1:32–41.
47. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther. 1991;14: 409–15.
48. Bolton JE, Breen AC. The Bournemouth Questionnaire: a short-form comprehensive outcome measure. I. Psychometric properties in back pain patients. J Manipulative Physiol Ther. 1999;2:503–10.

49. Ware JE, Snow KK, Kosinski MM, Gandek B. SF-36 Health Survey: Manual and Interpretation Guide. The Health Institute, New England Medical Centre, 1993.
50. Dawson B, Trapp R. Basic and Clinical Biostatistics. Sydney: McGraw-Hill, 2001.
51. Meng X, Rosenthal R, Rubin DB. Comparing correlated correlation coefficients. Psychol Bull. 1992;111:172–5.
52. Stratford PW. Calculation of sample size estimation for dependent samples. In: Gabel CP, ed. Coolum, Qld., Australia; 2002.
53. Jacobsen NS, Follette WC, Revensdorf D. Psychotherapy outcomes research: methods for reporting variability and evaluating clinical significance. Behav Ther. 1984;15:336–52.
54. MacDermid JC. Outcome evaluation in patients with elbow pathology: issues in instrument development and evaluation. J Hand Ther. 2001;14:105–14.
55. MacDermid JC, Richards RR, Donner A, Bellamy N, Roth JH. Responsiveness of the Short Form 36, Disabilities of the Arm, Shoulder and Hand Questionnaire, patient-related wrist evaluation and physical impairment measurements in evaluating recovery after a distal radius fracture. J Hand Surg. 2000;25A: 330–40.
56. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86:420–8.
57. McHorney CA, Taylor AR. Individual patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4:293–8.
58. Dabbagh T, Leaver A, Young S. Musculo-skeletal outcomes measures course: New South Wales Branch of the Private Practitioners Group of the Australian Physiotherapy Association, 2000.
59. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Control Clin Trials. 1991; 12:142S–58S.
60. Nunnally JC, Bernstein IH. Psychometric Theory. New York: McGraw-Hill, 1994.
61. Wyrwich KW, Nienaber MA, Tierney WM, Wolinsky FD. Linking clinical relevance and statistical significance in evaluating intra-individual changes in health-related quality of life. Med Care. 1999;37:469–78.
62. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting standard error of measurement based criterion for identifying meaningful intra-individual change in health related quality of life. J Clin Epidemiol. 1999;52:861–73.
63. Binkley JM, Stratford PW, Lott SA, Riddle DL. The Lower Extremity Functional Scale (LEFS): scale development, measurement properties, and clinical application. Phys Ther. 1999;79: 371–83.
64. Stratford PW, Binkley J, Solomon P, Finch E, Gill C, Moreland J. Defining the minimum level of detectable change for the Roland–Morris Questionnaire. Phys Ther. 1996;76:359–65.
65. Fairbank JCT, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy. 1980; 66:271–3.
66. Wright JG, Young NL. A comparison of different indices of responsiveness. J Clin Epidemiol. 1998;50:239–46.
67. De Bruin AF, Diederiks JPM, De Witte LP, Stevens FCJ, Philipsen H. Assessing the responsiveness of a functional status measure: the Sickness Impact Profile versus the SIP68. J Clin Epidemiol. 1997;50:529–40.
68. Kopec JA, Esdaile JM, Abrahamowicz M, Abenhaim L, Wood-Dauphinee S, Lamping DL. The Quebec Back Pain Disability Scale: measurement properties. Spine. 1995;20:341–52.
69. Beurskens AJHM, de Vet HCW, Koke AJA. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain. 1996;65:71–6.
70. Stratford PW, Binkley JM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. Phys Ther. 1996;76:1109–23.
71. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of shorter and longer health status instruments. Med Care. 1992;30.
72. Katz JN, Gelberman RH, Wright EA, Lew RA, Liang MH. Responsiveness of self-reported and objective measures of disease severity in carpal tunnel syndrome. Med Care. 1994; 32:1127–33.
73. Kazis LE, Anderson JJ, Meenon RF. Effect sizes for interpreting changes in health status. Med Care. 1989;27:S178.
74. Cohen J. Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences. Hillsdale, NJ: Lawrence Erlbaum, 1983.
75. Deyo RA. Comparative validity of the SIP, Sickness Impact Profile, and shorter scales for functional assessment in low back pain. Spine. 1986;11:951–4.
76. Stratford PW, Binkley J, Solomon P, Gill C, Finch E. Assessing change over time in patients with low back pain. Phys Ther. 1994;74:528–33.
77. Stratford PW, Finch E, Solomon P, Binkley J, Gill C, Moreland J. Using the Roland–Morris Questionnaire to make decisions about individual patients. Physiother Can. 1996.
78. Linton SJ, Melin L. The accuracy of remembering chronic pain. Pain. 1982;13:281–5.
79. Ross M. Relation of implicit theories to the construction of personal histories. Psychol Rev. 1989;96:341–57.
80. Norman GR, Stratford PW, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. J Clin Epidemiol. 1997;50.
81. Desousa L, Miles C, Frank AO. Back pain disability: are we learning what we should from the Roland Morris Disability Questionnaire (RMDQ). 11th World Congress on Pain, Sydney. IASP Press, Seattle, 2005 [Abstract 901–P143, P5325 pp.].
82. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297–334.
83. Veehof MM, Sleegers EJ, van Veldhoven NH, Schuurman AH, van Meeteren NL. Psychometric qualities of the Dutch language version of the disabilities of the arm, shoulder, and hand questionnaire (DASH-DLV). J Hand Ther. 2002;15: 347–54.
84. Institute for Work and Health. Development and Testing of the DASH and Quick-DASH Outcome Measure Instruments and the DASH User's Manual. IWH Measurement of Health & Function Projects. Ontario: Institute for Work and Health Ontario Canada, 2003.
85. Chansirinukor W. Development and measurement properties of the multi-level RM24: a modified version of the Roland–Morris Disability Questionnaire. In: Musculoskeletal Physiotherapy Australia: 13th Biennial Conference, Sydney; 2003.
86. Jacobson GP, Newman CW. The development of the Dizziness Handicap Inventory. Arch Otolaryngol Head Neck Surg. 1990;116:424–7.
87. Croft P, Pope D, Zonca M, O'Neill T, Silman A. Measurement of shoulder related disability: results of a validation study. Annu Rheum Dis. 1994;53:525–8.
88. Leclaire R, Blier F, Fortin L, Proulx R. A cross-sectional study comparing the Oswestry and Roland–Morris functional disability scales in two populations of patients with low back pain of different levels of severity. Spine. 1997;22:68–71.
89. Roland M, Morris R. A study of the natural history of back pain. Part II. Development of guidelines for trials of treatment in primary care. Spine. 1983;8:145–50.
90. Navsarikar A, Gladman DD, Husted JA, Cook RJ. Validity assessment of disabilities of arm shoulder and hand questionnaire (DASH) for patients with psoriatic arthritis. J Rheumatol. 1999;26:2191–4.
91. Shutek M, Fremerey RW, Zeichen J, Bosch U. Outcome analysis following open rotator cuff repair: early effectiveness validated using four different shoulder assessment scales. Arch Orthop Trauma. 2000;120:432–6.
92. Turchin DC, Beaton DE, Richards RR. Validity of observer based aggregate scoring systems as descriptors of elbow pain, function and disability. J Bone Joint Surg. 1998;80A:154–62.
93. Jain R, Hudak PL, Bowen CVA. Validity of health status measures in patients with ulnar wrist disorders. J Hand Ther. 2001;14:147–53.
94. Beaton DE, Richards RR. Assessing the reliability and responsiveness of five shoulder questionnaires. J Shoulder Elbow Surg. 1998;7:565–72.

95. Beattie P, Maher C. The role of functional status questionnaires for low back pain. Aust J Physiother. 1997;43: 29–35.
96. Ritchie JE. Using qualitative research to enhance the evidence-based practice of health care providers. Aust J Physiother. 1999;45:251–6.
97. Zimmerman D. Mimicking properties of nonparametric rank tests using scores that are not ranks. J Gen Psychol. 1993;120: 509–16.
98. Tabachnick BG, Fiddell LS. Using Multivariate Statistics. New York: Harper Collins, 1996.
99. Roland M, Fairbank JCT. The Roland–Morris Disability Questionnaire and the Oswestry Disability Questionnaire. Spine. 2000;25:3115–24.
100. Reneman MF, Jorritsma W, Schellekens JM, Goeken LN. Concurrent validity of questionnaire and performance-based disability measurements in patients with chronic nonspecific low back pain. J Occup Rehabil. 2002;12:119–29.

(Print on YELLOW Paper) **UPPER LIMB FUNCTIONAL INDEX (ULFI)** **DATE:_____**

NAME:_____ INJURY_____ ☐ **LEFT ARM** ☐ **RIGHT ARM**

Your upper limb (arm) may make it difficult to do some of the things you normally do. This list contains sentences people often use to describe themselves when they have such problems. Think of yourself over the last few days.

If an item describes you, mark the box. If not, leave the box blank.

DUE TO MY ARM:

☐	1	I stay at home most of the time.
☐	2	I change position frequently for comfort.
☐	3	I avoid heavy jobs eg. cleaning, lifting more than 5kg or 10lbs, gardening etc.
☐	4.	I rest more often.
☐	5.	I get others to do things for me.
☐	6.	I have pain almost all the time.
☐	7.	I have difficulty lifting and carrying (eg bags, shopping up to 5kg or 10lbs).
☐	8.	My appetite is now different.
☐	9.	My walking or normal recreation activity is affected.
☐	10.	I have difficulty with normal home or family duties and chores.
☐	11.	I sleep less well.
☐	12.	I need assistance with personal care eg. washing and hygiene.
☐	13.	My regular daily activities (work, social contact) are affected.
☐	14.	I am more irritable and / or bad tempered.
☐	15.	I feel weaker and / or stiffer.
☐	16.	My transport independence is affected (driving, public transport).
☐	17.	I have difficulty putting my arm into a shirt sleeves or need assistance dressing.
☐	18.	I have difficulty writing or using a key board and / or "mouse".
☐	19.	I am unable to do things at or above shoulder height.
☐	20.	I have difficulty eating and /or using utensils (eg knife, fork, spoon, chop sticks).
☐	21.	I have difficulty holding and moving dense objects (eg mugs, jars, cans).
☐	22.	I tend to drop things and/or have minor accidents more frequently.
☐	23.	I use the other arm more often.
☐	24.	I have difficulty with buttons, keys, coins, taps/faucets, containers or screw-top lids.
☐	25.	I have difficulty opening, holding, pushing or pressing (eg triggers, lever, heavy doors).

ULFI SCORE: To Score the Upper Part – Add the Marked Boxes:

TOTAL ULFI Points = ☐ **100 Scale (x 4) =** ☐ **%**

Patient Specific Index (PSI): Note 5 activities that are important to you and affected by your arm problem. If you cannot think of 5, choose from the ones you have marked above.

Score each activity on a scale range as follows, you may use Half (½) marks if you wish:

0 = BEST: Never affected / Can do activity normally **5 = WORST:** Always affected / Can't do activity at all

	ACTIVITY	*Score*
1.		
2.		
3.		
4.		
5.		

PSI Total = _____

% Score = (Total x 4) = _____

MDC (90% Confidence): 10.5 % or 2.6 ULFI points. Change < this may be due to error

Think of yourself **over the last few days** and **due to your arm** - assess your **Overall Status** compared to your normal or pre-injury level?

0 1 2 3 4 5 6 7 8 9 10

Pre-Injury or Normal — Worst Possible

Parametric tests were used as all data demonstrated or approached a Gaussian or ''bell shape'' distribution. In addition it has been demonstrated that the results of statistical tests are minimally affected by either severely abnormal distributions or violation of the homogeneity of variance assumption, provided the samples are, respectively, from the same population with the same sample size.[97] Consequently, when the ''quality'' of SROM scales is considered using ''Stephen's Typology,'' whether the format is ''Likert'' (with multiple response options) or ''dichotomous'' (with two options of ''yes'' or ''no''), data should be considered ''ordinal.'' This is because the values on the dichotomous scale are not arbitrary or ''nominal'' as the responses are ordered with a ''yes'' defining the presence of impairment.[98] Since the individual item statements are then summated to produce the total score for the respondent the data are determined as being interval or ratio quality. These assumptions and statistical analyses are supported by previous authors who compared data from both Likert and Dichotomous format SROMs.[11,64,99,100] They are further supported by research that analyzed the performance of a standard dichotomous SROM that was modified to a ''Likert'' format and found that three ascending categories of 0–3, 4–7, and 8–10 provided optimal psychometric properties in preference to either the original Dichotomous or the modified Likert format, effectively the original dichotomous choice with the addition of a half (½) mark option.[85]

JHT Read for Credit
Quiz: Article #039

Record your answers on the Return Answer Form found on the tear-out coupon at the back of this issue. There is only one best answer for each question.

#1. The ULFI includes:
   a. a Moberg pick-up test
   b. an SAT
   c. a VALPAR
   d. a VAS

#2. The following stat was used to determine the reliability of the ULFI:
   a. Kappa
   b. ANOVA
   c. ICC
   d. student T test

#3. Missing responses are:
   a. anticipated in all SROMs
   b. not seen in the ULFI
   c. not seen in the DASH
   d. not seen in the UEFS

#4. The authors claim that compared to the DASH and UEFS, the ULFI is:
   a. less practical
   b. higher in terms of practicality
   c. equally practical
   d. none of the above

#5. A potential worry in interpreting the clinical applicability of the ULFI is the:
   a. effect size
   b. statistical power based on the sample size
   c. ANOVA
   d. use of English speaking subjects

When submitting to the HTCC for re-certification, please batch your JHT RFC certificates in groups of 3 or more to get full credit.