Test-Retest Reliability of the Upper Extremity Questionnaire Among Keyboard Operators

Deborah F. Salerno, PhD,[1] Alfred Franzblau, MD,[1,2]* Thomas J. Armstrong, PhD,[2,3] Robert A. Werner, MD,[1,2,4,5] and Mark P. Becker, PhD[6]

Background *Questionnaires are often used in research among workers although few have been tested in the working population. The Upper Extremity Questionnaire is a self-administered questionnaire designed for epidemiological studies and tested among workers. This study assessed reliability of the instrument.*
Methods *A two-part assessment was conducted among 138 keyboard operators as part of a large medical survey. Test-retest reliability was analyzed using the kappa statistic, paired t-test, and intraclass correlation coefficient (ICC). Logistic regression models were used to test the effect of demographic and work-related factors on reliability.*
Results *The average respondent was a white woman, age 35 years, with some college education, in permanent employment with tenure of 1.4 years. Overall, reports of symptoms were stable from Round 1 to 2. Most kappa values for symptom reports were between 0.60 and 0.89. Kappa values for right and left hand diagrams were 0.57 and 0.28, respectively. Among psychosocial items, Perceived Stress and Job Dissatisfaction Scales were most reliable (ICC = 0.88); co-worker support was least reliable (ICC = 0.44).*
Conclusion *Reliability of items on the Upper Extremity Questionnaire were generally good to excellent. Reports of symptom severity and interference with work were less stable. Demographic and work-related factors were not statistically significant in modeling the variation in reliability. Repeated use of the questionnaire with similar results suggests findings are applicable to a larger working population.* Am. J. Ind. Med. 40:655–666, 2001. *© 2001 Wiley-Liss, Inc.*

KEY WORDS: kappa statistic; occupational medicine; questionnaire; reliability; upper extremity disorders

INTRODUCTION

Questionnaires are often used in research of upper extremity disorders [Silverstein et al., 1987; Homan et al., 1999; Punnett and van der Beek, 2000]. Yet, the reliability of questionnaires is seldom tested directly among workers. One reason for the lack of reliability studies is the difficulty in arranging repeated access to workers. Also, in contrast to

[1]Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, Michigan
[2]Center for Ergonomics, College of Engineering, University of Michigan, Ann Arbor, Michigan
[3]Department of Industrial and Operations Engineering, College of Engineering, University of Michigan, Ann Arbor, Michigan
[4]Physical Medicine and Rehabilitation, Veterans Administration Hospital, Ann Arbor, Michigan
[5]Department of Physical Medicine and Rehabilitation, School of Medicine, University of Michigan, Ann Arbor, Michigan
[6]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan

Website for University of Michigan Upper Extremity Questionnaire: http://umrerc.engin.umich.edu

This research was completed during Dr. Salerno and Dr. Becker's tenure at the University of Michigan School of Public Health. Dr. Salerno is currently at Pfizer Global Research and Development-Ann Arbor Laboratories and Dr. Becker is currently Dean, School of Public Health, University of Minnesota.
Contract grant sponsor: Johns Hopkins University Center for VDT and Health Research.
*Correspondence to: Alfred Franzblau, MD, Department of Environmental Health Sciences, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: afranz@umich.edu

Accepted 6 September 2001

DOI 10.1002/ajim.10024

questionnaires designed for clinical populations, instruments relevant to active workers are hard to find [Stock et al., 1996].

Notable exceptions are reliability reports on the Nordic Musculoskeletal Questionnaire [Kuorinka et al., 1987; Dickinson et al., 1992]. They found acceptable reliability of questionnaires among cashiers, clerical workers, nursing staff, railway maintenance workers, and safety engineers. However, only percentage agreement was reported without agreement adjusted for chance.

The high prevalence and cost of musculoskeletal disorders in the workplace show the importance of measures for workers [Feuerstein et al., 1999; Herbert et al., 1999; United States Department of Labor, 2000, 2001]. Since reliability is context-dependent, if used in the workplace, measures need to be tested there. High quality data are contingent on valid and reliable measures being used in the appropriate population [Bergner and Rothman, 1987; Gerr et al., 1991; Baron et al., 1996; McHorney, 1999].

This report summarizes test-retest reliability of items on the Upper Extremity Questionnaire. A previous report [Franzblau et al., 1997] focused on reliability among industrial workers. This study evaluated reliability among keyboard operators. Repeated testing among different types of workers with similar results indicates the extent to which reliability results are applicable to the larger working population.

METHODS

Description of the Instrument

The Upper Extremity Questionnaire is a 15-page, self-administered instrument designed for epidemiologic studies of workers (http://umrercdev.engin.umich.edu; Resources, Symptom Survey, UE questionnaire). The first part consists of demographic items, medical history (including exercise), and a series of questions relating to symptoms in three body regions:

1. Neck/shoulder/upper arm,
2. Elbow/forearm, and
3. Wrist/hand/finger, including a hand diagram, which subjects are instructed to complete by shading in affected areas of the wrists, hands, and/or fingers if they experienced numbness, tingling, burning, or pain.

Questions probe whether symptoms have occurred more than three times in the past year or lasted more than 1 week, whether symptoms have interfered with production rate and/or usual standard of quality in the last 7 days, the number and duration of episodes of the problem in the past year, severity of discomfort right now and in the last 30 days, and treatment by a medical professional or the subject. The questionnaire continues with questions regarding overall problems, additional jobs, and if so, how many additional hours worked per week, height and weight, membership in a union or employee association, annual income, and a section on tobacco smoking [Ferris, 1978].

The second part of the questionnaire consists of two subtests of psychosocial conditions at work: (1) the Job Content Scale [Karasek et al., 1981] explores whether the job requires creativity, involves repetitive work, decision latitude, and work relations (e.g., exposure to hostility, friendliness, encouragement), relations with the supervisor, and job security; (2) the Perceived Stress Scale [Cohen et al., 1983] explores stress, perceived locus of control, and coping mechanisms.

Survey Protocol

For this study, a two-part reliability assessment was conducted as part of a larger medical survey of keyboard operators at a data coding center in the midwestern US. Active workers employed on the day or afternoon shifts were eligible to participate. Round 1 of the survey consisted of the Upper Extremity Questionnaire; a functional activity questionnaire, physical examination, and electrodiagnostic testing. Round 2, 3 weeks later, included the questionnaire, and slightly modified medical procedures.

Subjects completed questionnaires on company time during their normal work hours in a semi-private area to minimize group interaction. The mode of administration was identical between Rounds. Subjects provided informed consent, which had been approved by the Human Subjects Review Committee at the University of Michigan School of Public Health. No personally identifiable information was provided to the union or management.

Statistical Analysis

The reliability of categorical data was assessed with the kappa statistic [Cohen, 1960], a measure for testing whether agreement exceeds chance levels: $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is the proportion of observed agreement; and p_e is the proportion of agreement expected by chance, the joint probabilities of the marginal proportions.

As noted by Feinstein and Cicchetti [1990], we define observed and expected agreement as $p_o = (a + d)/N$,

		Round 2		
		Yes	No	Total
Round 1	Yes	a	b	g_1
	No	c	d	g_2
		f_1	f_2	N

and $p_e = (f_1 g_1 + f_2 g_2)/N^2$ where *a* and *d* represent positive and negative agreement, respectively, in a 2×2 table.

Weighted kappa statistics [Fleiss and Cohen, 1973] were used, when appropriate. Values of kappa greater than 0.75 were considered excellent; values between 0.40 and 0.75 were fair to good; and values less than 0.40 represented poor agreement beyond chance [Fleiss, 1981].

To enhance interpretation, the prevalence of positive answers was reported. With very low or high prevalence (i.e., as prevalence approaches 0 or 1), kappa is attenuated [Thompson and Walter, 1988]. McNemar χ^2 statistics were reported for items with statistically significant differences in prevalences between Rounds 1 and 2.

In testing the reliability of hand diagrams, first, diagrams were separated into stacks of marked and unmarked diagrams. Then, two raters independently scored the diagrams for the likelihood of underlying carpal tunnel syndrome (CTS) according to the following rating system [Franzblau et al., 1994]:

Classic: Numbness, tingling, burning or pain in at least two of digits I, II, or III. The palm and dorsum of the hand were excluded while wrist pain or radiation proximal to the wrist was allowed.

Probable: Same as for classic, except palmar symptoms allowed unless confined solely to ulnar aspect.

Possible: Numbness, tingling, burning or pain in at least one of digits I, II, or III.

Unlikely: No symptoms in digits I, II, or III.

After initial scoring, raters resolved discrepancies by consensus. Reliability was tested with the weighted kappa statistic.

Test-retest reliability of the continuous data was assessed with an intraclass correlation coefficient (ICC), which combines a measure of correlation with a test in the difference of means [Kramer and Feinstein, 1981]. We chose the computing formula for a random-effects ANOVA model [Müller and Büttner, 1994]. Although the present study involved two fixed observers, this was justifiable since the number of subjects was large. Thus, the variances for fixed and random observers were practically interchangeable.

DEMOGRAPHIC OR WORK-RELATED FACTORS

Further analyses were performed to model the variation in reliability due to demographic or work-related factors. We examined the relationship between reliability and age, sex, education [1, high school; 2, some college; 3, ≥ 16 years (Bachelor's degree or higher)], employment status (temporary versus permanent), tenure, keystrokes per hour, and total number of hours keying since starting work. Although these variables could be associated with *symptoms*, none were expected to have an association with *reliability.*

As subjects who are always symptomatic are probably different than asymptomatic subjects, we checked to see if any factors differentiated subjects consistently with symptoms versus without symptoms. [Note: this analysis was more relevant to issues of validity rather than reliability].

The main interest was in those whose status changed, as they provided unreliable reports. For these subjects, we tested whether any factors could differentiate the subjects without symptoms in Round 1, but with symptoms in Round 2, from the subjects with symptoms in Round 1, but without symptoms in Round 2. For this exploratory analysis, new dichotomous variables were generated for each condition (e.g., ‘‘Neck Symptom Fluctuation,’’ was defined as: Neck Symptom Fluctuation = 0 if subject had no neck symptoms in Round 1, but symptoms in Round 2; Neck Symptom Fluctuation = 1 if subject had neck symptoms in Round 1, but no symptoms in Round 2). Fourteen body areas were tested: neck; shoulder, upper arm, elbow, forearm, bilaterally; neck/shoulder/upper arm problems, elbow/forearm problems, wrist/hand/finger problems; and distal numbness, tingling, burning or pain, bilaterally.

Each new variable was tested with the chi-squared statistic, Fisher's exact or Student *t*-test, to examine associations for those with consistent or fluctuating symptoms and the demographic or work-related factors.

Finally, multivariate logistic models were constructed using the maximum likelihood method to explore reliability of symptom reporting for the same 14 body regions, including all relevant covariates. Additional new variables were generated similar to the previous report in an industrial population [Franzblau et al., 1997], such as ‘‘Neck Problems’’ defined as: Neck Problems = 0 if symptoms fluctuated between Rounds; Neck Problems = 1 if symptoms were stable between Rounds.

The method for logistic regression involved two steps: (1) Variable selection, and (2) Assessing the fit of the model. The selection process began with univariate analysis of each variable. Any variable with a *P*-value less than 0.25 was included in the first model. The best subset of variables was chosen by a process of deleting, refitting, and verifying variables, including all that were clinically meaningful or statistically important. Goodness of fit was assessed with the Pearson chi-square statistic and the Hosmer–Lemeshow goodness of fit statistic [Hosmer and Lemeshow, 1989].

Statistical analyses were performed using Stata for Windows, Release 5.0 (StataCorp, College Station, TX). Other than in the univariate regression routines, tests were considered statistically significant at the $\alpha = 0.05$ level.

Nonresponse

Overall, item nonresponse was very low (generally 5% or less for questionnaire items), with few exceptions [weeks of pregnancy (50%), area of worst symptoms (12–14%), date when subject first experienced the problem (10–14%), problems being associated with a particular work station or work activity (7%), and supervisor support (7%)]. No imputations for missing data were performed.

Subject Participation

Of the 161 subjects in Round 1 of the study, 138 (86%) subjects returned in Round 2. (There were 476 subjects who completed the Questionnaire in Round 1 or 2 of the medical survey, however, for logistical reasons, not all subjects were able to participate in the two-part reliability study).

Subject Characteristics and Disposition

The average worker was a 35-year-old white woman with 13–15 years of education, in a permanent position with 1.4 years' tenure as a keyboard operator (Table I). There were no significant demographic differences between the subjects who completed Round 1 only and those who completed Round 2. Overall, reports of wrist/hand/fingers symptoms had the highest prevalence (62%), followed by neck/shoulder/upper arm symptoms (50%), and elbow/forearm symptoms (35%).

RESULTS

For the "Medical History and Exercise" section (Table II), items related to gynecological history and surgery were more reliable (and prevalent) than accounts of arthritis, diabetes, or gout (κ range: 0.48–1).

For reports of "Localized Injury or Disease" (Table III), kappa values ranged from 0.30 to 1, with excellent reliability for thoracic outlet syndrome, right shoulder tendinitis, CTS, and fractures in the fingers, hand, wrist, forearm, or elbow.

Kappa values ranged from 0.40 to 1 for neck/shoulder/upper arm reports (Table IV). In the elbow/forearm area (Table V), the majority of reports exhibited excellent reliability ($\kappa > 0.75$). For reports of wrist/hand/fingers symptoms, most kappa values ranged from 0.60 to 0.86 (Table VI).

For right and left hand diagrams ($n = 138$), weighted kappa values were 0.57 and 0.28, respectively (Table VI). In Round 1, there were 82 marked diagrams with 96% agreement between the two raters ($\kappa = 0.85$, right hand; $\kappa = 0.82$, left hand). In Round 2, there were 78 marked diagrams with

TABLE I. Subject Characteristics in Test-Retest Reliability Study Among Keyboard Operators

	Round 1 (n = 161)	Round 1 only (n = 23)	Round 2* (n = 138)
Age (years) (SD) [range]	35.5 (9.1) [20–58]	36.9 (8.6) [22–51]	35.2 (9.2) [20–58]
Sex			
Female (%)	147 (91)	21 (91)	126 (91)
Male (%)	14 (9)	2 (9)	12 (9)
Hand dominance			
Right (%)	147 (91)	22 (96)	125 (91)
Left (%)	14 (9)	1 (4)	13 (9)
Race			
White (%)	148 (92)	21 (92)	127 (92)
Black (%)	8 (5)	1 (4)	7 (5)
Other (%)	5 (3)	1 (4)	4 (3)
Education			
High school (%)	51 (32)	6 (26)	45 (33)
13–15 years (%)	84 (52)	13 (57)	71 (51)
≥ 16 years (%)	26 (16)	4 (17)	22 (16)
Tenure (years) (SD) [range]	1.4 (0.4) [0.4–1.6]	1.5 (0.2) [0.7–1.6]	1.4 (0.4) [0.4–1.6]
Employment status			
Temporary (%)	56 (35)	11 (48)	45 (33)
Permanent (%)	105 (65)	12 (52)	93 (67)

*Round 2 data based on Round 1 responses. No statistically significant difference (*P*-value $\leq$ 0.05) between Round 1 only and Round 2 subjects.

TABLE II. Reliability of Medical History and Exercise Items Among Keyboard Operators

Condition or question	Kappa	Kappa 95% CI[a]	Observed % agreement	Expected % agreement	P_1[b]	P_2[b]
Hysterectomy[c]	1	0.82, 1	100	91	.05	.05
Double oophorectomy[c]	1	0.82, 1	100	95	.02	.02
Birth control pills[c]	0.95	0.78, 1	98	66	.22	.22
Menopause[c]	0.88	0.71, 1	99	93	.03	.04
Surgery on neck or upper extremity	0.76	0.59, 0.92	98	91	.05	.04
Thyroid dysfunction	0.66	0.49, 0.82	98	94	.03	.04
Relative with nerve/ muscle disease	0.64	0.47, 0.81	92	78	.13	.12
Diabetes	0.49	0.35, 0.64	99	97	.01	.02
Rheumatoid arthritis	0.49	0.35, 0.64	99	97	.01	.02
Gout	0.49	0.35, 0.64	99	97	.01	.02
Fibromyalgia	0.49	0.35, 0.64	99	97	.01	.02
Degenerative arthritis	0.49	0.35, 0.64	99	97	.01	.02
Current pregnancy[c]	0.48	0.31, 0.66	97	94	.03	.03
Chronic renal failure	undefined	—	99	99	0	.01
Exercise	0.84	0.67, 1	93	53	.60	.64
Exercise amount[d]	0.67	0.45, 0.89	94	82	n/a[e]	n/a[e]

[a]95% Confidence Intervals for $\kappa = \kappa \pm 1.96\,(SE_\kappa)$.
[b]P_1 = Prevalence of condition reported in Round 1; P_2 = Prevalence of condition reported in Round 2. When $P_1 = 0$ or $P_2 = 0$, then kappa is undefined, (see Chronic renal failure). Unless noted, differences in the prevalences of signs reported in Round 1 and 2 are not statistically significant at $\alpha = 0.05$ level with the McNemar χ^2 test for independent proportions.
[c]Kappa computed for women only.
[d]Weighted kappa (quadratic weight).
[e]n/a = Not applicable.

97% agreement between raters ($\kappa = 0.90$, right hand; $= 0.84$, left hand).

Regarding treatment for upper extremity problems, kappa values for responses ranged from 0.64 to 0.78. Rest and anti-inflammatory medications were most reliable (κ range: 0.58–0.84).

In each body region, the item with the highest prevalence was whether problems were "associated with a particular work station or work activity," which had excellent reliability in the neck and elbow regions ($\kappa \geq 0.75$), but not in the wrist/hand/fingers region ($\kappa = 0.45$).

For the item assessing "interference of symptoms with production rates and/or usual standard of quality in the last 7 days" [1 = strongly agree to, 5 = strongly disagree], kappa values ranged from 0.31 to 0.56 (Tables IV–VI).

Items regarding the "number of separate episodes" and "duration of episodes" had moderate reliability (κ range: 0.44–0.70). The item asking whether this problem was "caused by a particular event such as an accident or injury" had higher reliability (κ range: 0.65–1.00), indicating workers consistently attributed problems to a particular event, when a cause was identified.

Discomfort ratings on the 0–10 visual analog scale indicated wrist/hand/fingers was most reliable of the three body regions (Table VII). Among all subjects, ratings of "discomfort in the past 30 days" were slightly more reliable (ICC range: 0.77–0.80) than "current discomfort" (ICC range: 0.59–0.79). When analyses were restricted to those subjects with symptoms in Round 1 or 2, reliability was poor to good (ICC range: 0.32–0.74). "Current discomfort" in the wrist/hand/fingers region was most reliable (ICC = 0.74) among subjects with symptoms, even with a statistically significant increase in mean ratings between Rounds 1 and 2.

Reports of moonlighting had excellent reliability ($\kappa = 0.79$), as did reports of the number of hours per week in outside employment (ICC = 0.96). Tobacco smoking questions had perfect reliability, querying whether respondents ever/never smoked, and current/non-smoker status. The ICC was 0.94 for reports of pack years, which involved recollection of how many years the subject smoked, and how many cigarettes were smoked per day, on average, during the entire time the subject smoked.

Among the psychosocial scales (Table VIII), job dissatisfaction, perceived stress, intellectual discretion, and creative skill had excellent reliability (ICC range: 0.78–0.88). Co-worker support was least reliable (ICC = 0.44).

TABLE III. Reliability of History Localized Injury or Disease Items Among Keyboard Operators

Localized injury or disease	Kappa	Kappa 95% CI[a]	Observed % agreement	Expected % agreement	P1[b]	P2[b]	*P*-value
Right thoracic outlet syndrome	1	0.83, 1	100	99	.01	.01	
Left thoracic outlet syndrome	1	0.83, 1	100	99	.01	.01	
Right tendinitis, shoulder	0.90	0.74, 1	99	92	.04	.04	
Left tendinitis, shoulder	0.49	0.32, 0.65	97	94	.04	.02	
Right fracture, fingers, hand, wrist, forearm or elbow	0.78	0.61, 0.95	95	77	.12	.15	
Left fracture, fingers, hand, wrist, forearm or elbow	0.85	0.68, 1	96	76	.14	.14	
Right carpal tunnel syndrome	0.76	0.59, 0.92	98	91	.05	.04	
Left carpal tunnel syndrome	0.48	0.32, 0.65	97	94	.03	.03	
Right tendinitis, fingers, wrist, forearm or elbow	0.72	0.55, 0.89	93	77	.14	.13	
Left tendinitis, fingers, wrist, forearm or elbow	0.49	0.33, 0.66	93	87	.08	.06	
Right ganglion	0.66	0.51, 0.82	99	98	.01	.01	
Left ganglion	Undefined	—	99	99	.01	0	
Right muscle strain, upper arm or shoulder	0.64	0.48, 0.81	96	88	.07	.06	
Left muscle strain, upper arm or shoulder	0.30	0.15, 0.44	91	88	.09	.04	0.02
Right rotator cuff injury	0.56	0.41, 0.71	98	95	.04	.01	
Left rotator cuff injury	Undefined	—	99	99	.01	0	
Right cervical radiculopathy	0.56	0.39, 0.73	98	95	.03	.02	
Left cervical radiculopathy	0.66	0.49, 0.82	98	94	.04	.03	
Right muscle strain, hand, wrist, forearm or elbow	0.39	0.23, 0.56	86	76	.17	.11	
Left muscle strain, hand, wrist, forearm or elbow	0.39	0.24, 0.55	89	82	.13	.06	0.02
Right fracture, upper arm	Undefined	—	99	99	.01	0	
Left fracture, upper arm	0.32	0.16, 0.48	97	96	.01	.03	
Right ulnar neuropathy	Undefined	—	100		0	0	
Left ulnar neuropathy	Undefined	—	100		0	0	

[a]95% Confidence Intervals for $\kappa = \kappa \pm 1.96\,(SE_\kappa)$.
[b]P_1 = Prevalence of condition reported in Round 1; P_2 = Prevalence of condition reported in Round 2. When $P_1 = 0$ or $P_2 = 0$, then kappa is undefined (see Fracture, upper arm, right). Unless noted, differences in the prevalences of signs reported in Round 1 and 2 are not statistically significant at $\alpha = 0.05$ level with the McNemar χ^2 test for independent proportions.

DEMOGRAPHIC OR WORK-RELATED FACTORS

In testing demographic and work-related variables, as expected, analysis of subjects with stable symptom reports (either with symptoms or without symptoms in both Rounds 1 and 2 indicated age, education, employment, keystrokes per hour, and/or total hours worked were significantly related to symptoms for at least one body area (data not shown). Although this analysis was interesting, it provides no information on the variation in symptom reports (since reports in both rounds were consistent—either symptomatic or asymptomatic). Thus, it relates to an issue of validity, not reliability.

Our main interest in this reliability study was in those whose status changed, as they are the people who provide unreliable reports. Overall, very few reports changed from symptomatic to asymptomatic (or vice versa) between Rounds 1 and 2 (Table IX). Univariate analysis with fluctuating reports identified age, education, and keystrokes per hour, as significant factors for reports of right and left elbow symptoms, and wrist/hand/finger problems. However, low cell frequencies precluded meaningful statistical inference.

For the logistic regression analysis, all fluctuating symptom reports were combined into one category, in contrast to the stable symptom reports. None of the regression models was significant, indicating no statistically significant association between demographic or work-related variables and reliability (data not shown).

DISCUSSION

Our primary objective was to assess test-retest reliability of individual items on the Upper Extremity Questionnaire among keyboard operators. Test-retest reliability of items was generally good to excellent, similar to findings in an industrial population [Franzblau et al., 1997]. Repeated use of the questionnaire with similar results suggests

TABLE IV. Reliability of Neck/Shoulder/Upper Arm Conditions Among Keyboard Operators

Question	Kappa	Kappa 95% CI[a]	Observed % agreement	Expected % agreement	P_1[b]	P_2[b]	*P*-value
Any neck/shoulder/upper arm symptoms	0.75	0.59, 0.92	88	50	.50	.52	
Neck symptoms	0.78	0.61, 0.95	89	51	.44	.45	
Right shoulder/upper arm symptoms	0.68	0.51, 0.85	85	52	.38	.41	
Left shoulder/upper arm symptoms	0.62	0.46, 0.79	83	54	.35	.37	
Right shoulder symptoms	0.65	0.48, 0.81	83	53	.35	.41	
Left shoulder symptoms	0.61	0.44, 0.77	83	56	.33	.33	
Right upper arm symptoms	0.48	0.31, 0.65	85	71	.18	.17	
Left upper arm symptoms	0.58	0.42, 0.75	88	72	.16	.17	
Caused by particular event	0.90	0.65, 1	98	83	.08	.10	
Association with work station or work activity	0.76	0.51, 1	90	58	.72	.68	
Duration of episodes[c,g]	0.63	0.38, 0.88	94	85	n/a	n/a	
Number of separate episodes[d,g]	0.61	0.36, 0.86	89	73	n/a	n/a	
Worst symptom area[e]	0.53	0.34, 0.72	72	40	n/a	n/a	
Interference with production and/or quality[f,g]	0.45	0.21, 0.69	90	82	n/a	n/a	
Treatment	0.78	0.61, 0.94	91	57	.33	.28	0.05
Steroid injection	1	0.83, 1	100	96	.02	.02	
Anti-inflammatory medication	0.66	0.50, 0.83	91	74	.14	.16	
Rest	0.58	0.41, 0.75	86	67	.21	.20	
Physical therapy	0.48	0.34, 0.63	96	92	.06	.02	0.01
Icing	0.40	0.23, 0.56	94	90	.04	.05	
Surgery	Undefined	—	99	99	0	0	

[a] 95% Confidence Intervals for $\kappa = \kappa \pm 1.96\ (SE_\kappa)$.
[b] P_1 = Prevalence of condition reported in Round 1; P_2 = Prevalence of condition reported in Round 2. When $P_1 = 0$ or $P_2 = 0$, then kappa is undefined (see Surgery). Unless noted, differences in the prevalences of signs reported in Round 1 and 2 are not statistically significant at $\alpha = 0.05$ level with the McNemar χ^2 test for independent proportions. n/a = not applicable.
[c] Duration of episodes: 1 hr or less, more than an hour but less than a day, more than a day but less than a week, more than a week but less than a month, or more than a month.
[d] Number of separate episodes: 3–12 episodes, 13–36 episodes, 37–52 episodes, 53–150 episodes, more than 150 episodes, or continuous.
[e] Worst symptom area: neck, right shoulder, left shoulder, right upper arm, left upper arm.
[f] Interference: 1 = strongly agree to 5 = strongly disagree.
[g] Weighted kappa (quadratic weight).

that findings are applicable to a larger working population of industrial workers and keyboard operators.

In general, reports of injury or disease on the right side had higher reliability than the left side (Table III), attributed, in part, to the salience of conditions on the right side. Given the smaller number of events on the left side, one could expect them to be easier to recall, however, the current study suggested otherwise. The salience of issues on the dominant side (90% of subjects were right-handed), may have contributed to right-sided ailments being easier to recall, and therefore more reliable. This was clearly evident for shoulder tendinitis, and upper arm/shoulder muscle strain, with higher right-sided kappa values and non-overlapping confidence intervals.

Reliability for items regarding ‘‘interference of symptoms with production rates and/or usual standard of quality’’ (in Tables IV–VI) was consistently low in every body region. One factor that may have contributed to the low reliability is the compound nature of the item [in the last 7 days, my (body region) symptoms have interfered with my production rates and/or my usual standard of quality]. Respondents may have focused on production rates in Round 1, and quality in Round 2, or vice versa.

In comparison to ratings of ‘‘current discomfort,’’ ratings with a larger frame of reference (the last 30 days) were generally more reliable (Table VII). These findings exactly paralleled what was found in our previous study [Franzblau et al., 1997]. Ratings of subjects with symptoms had lower ICCs than ratings among all subjects, most likely a statistical artifact of restricting analyses to exclude subjects with perfect agreement (no symptoms in Round 1 and 2).

Reliability of Hand Diagrams

For the hand diagrams, reliability was a function of two variables: (1) the subjects who completed the diagrams (test-retest reliability), and (2) the raters who scored them

TABLE V. Reliability of Elbow/Forearm Conditions Among Keyboard Operators

Question	Kappa	Kappa 95% CI[a]	Observed % agreement	Expected % agreement	P_1[b]	P_2[b]
Any elbow/forearm symptoms	0.89	0.72, 1	95	55	.35	.33
Right elbow or forearm symptoms	0.80	0.63, 0.97	95	75	.15	.14
Left elbow or forearm symptoms	0.87	0.70, 1	95	61	.27	.25
Right elbow symptoms	0.80	0.63, 0.97	95	75	.15	.14
Left elbow symptoms	0.82	0.65, 0.98	96	80	.12	.10
Right forearm symptoms	0.79	0.63, 0.96	92	61	.26	.26
Left forearm symptoms	0.85	0.68, 1	94	62	.26	.24
Caused by particular event	1	0.69, 1	100	95	.03	.03
Association with work station or work activity	0.75	0.44, 1	92	69	.80	.82
Worst symptom area[c]	0.75	0.53, 0.97	84	35	n/a	n/a
Duration of episodesf[d,g]	0.61	0.30, 0.93	95	86	n/a	n/a
Number of separate episodes[e,g]	0.44	0.13, 0.75	84	72	n/a	n/a
Interference with production and/or quality[f,g]	0.31	0.01, 0.61	90	85	n/a	n/a
Treatment	0.73	0.57, 0.90	92	70	.18	.20
Physical therapy	1	0.79, 1	100	94	.03	.03
Icing	1	0.79, 1	100	96	.02	.02
Steroid injection	1	0.79, 1	100	98	.01	.01
Anti-inflammatory medication	0.84	0.65, 1	97	81	.12	.09
Rest	0.76	0.56, 0.95	96	83	.10	.08
Surgery	Undefined	—			0	0

[a]95% Confidence Intervals for $\kappa = \kappa \pm 1.96\,(SE_\kappa)$.
[b]P_1 = Prevalence of condition reported in Round 1; P_2 = Prevalence of condition reported in Round 2. When $P_1 = 0$ or $P_2 = 0$, then kappa is undefined (see Surgery). Unless noted, differences in the prevalences of signs reported in Round 1 and 2 are not statistically significant at $\alpha = 0.05$ level with the McNemar χ^2 test for independent proportions. n/a = not applicable.
[c]Worst symptom area: right elbow, left elbow, right forearm, left forearm.
[d]Duration of episodes: 1 hr or less, more than an hour but less than a day, more than a day but less than a week, more than a week but less than a month, or more than a month.
[e]Number of separate episodes: 3–12 episodes, 13–36 episodes, 37–52 episodes, 53–150 episodes, more than 150 episodes, or continuous.
[f]Interference: 1 = strongly agree to 5 = strongly disagree.
[g]Weighted kappa (quadratic weight).

(inter-rater agreement). Each of these operations contributed to the overall variability. Theoretically, it is possible to have poor inter-rater agreement but high test-retest reliability. Conversely, it is possible to have excellent inter-rater agreement but low test-retest reliability.

Katz and Stirrat [1990] reported inter-rater agreement of 84% (with no adjustment for chance). Franzblau et al. [1994] reported excellent inter-rater agreement for scoring hand diagrams in the dominant hand ($\kappa = 0.93$), with similar results in the non-dominant hand. The previous reports are similar to results of the present study, and suggest that inter-rater agreement for scoring hand diagrams is excellent.

In the study among an industrial population, Franzblau et al. [1997] showed comparable bilateral test-retest reliability of hand diagrams ($\kappa = 0.52$ right and $\kappa = 0.59$ left). In contrast, the present study showed right hand diagrams had better test-retest reliability ($\kappa = 0.57$) than left hand diagrams ($\kappa = 0.28$). The current results indicate that reliability of hand diagrams may be more stable in the right or dominant hands.

Limitations of the Test-Retest Method

Prevalence is a critical statistical issue to consider in studies of reliability, especially when comparing studies. As true prevalence approaches zero, kappa also approaches zero [Thompson and Walter, 1988]. The test-retest method is also associated with certain problems that may affect the interpretation of results, namely: sensitization, an actual change in the underlying condition being studied, and the effects of memory [Carmines and Zeller, 1979]. The general stability of prevalences, measured by the McNemar test, indicated that sensitization was not a major issue in this study. Changes in prevalence, when statistically significant, were typically in the negative direction from Round 1 to 2. This was the same pattern observed in our previous study [Franzblau et al., 1997], and could indicate a fear of re-

TABLE VI. Reliability of Wrist/Hand/Fingers Conditions Among Keyboard Operators

Question	Kappa	Kappa 95% CI[a]	Observed % agreement	Expected % agreement	P_1[b]	P_2[b]
Any wrist/hand/finger symptoms	0.86	0.69, 1	93	53	.62	.64
Right wrist symptoms	0.83	0.66, 0.99	91	50	.47	.50
Left wrist symptoms	0.84	0.67, 1	92	51	.41	.43
Right distal numbness/tingling/burning/pain[c]	0.78	0.61, 0.95	89	50	.45	.47
Left distal numbness/tingling/burning/pain[c]	0.64	0.47, 0.81	83	51	.41	.42
Nocturnal occurrence	0.70	0.49, 0.92	86	53	.37	.36
Right nocturnal symptoms	0.60	0.44, 0.77	89	73	.15	.17
Left nocturnal symptoms	0.77	0.60, 0.93	94	72	.14	.14
Right fingers symptoms	0.75	0.58, 0.91	88	51	.41	.42
Left fingers symptoms	0.65	0.49, 0.82	83	52	.37	.42
Right hand symptoms	0.64	0.47, 0.80	83	52	.37	.43
Left hand symptoms	0.63	0.46, 0.79	83	53	.35	.38
Right hand diagram[f,g]	0.57	0.47, 0.74	93	83	n/a	n/a
Left hand diagram[f,g]	0.28	0.11, 0.45	90	86	n/a	n/a
Number of separate episodes[d,g]	0.70	0.49, 0.92	93	75	n/a	n/a
Duration of episodes[e,g]	0.69	0.47, 0.91	95	84	n/a	n/a
Caused by particular event	0.65	0.45, 0.86	97	93	.05	.03
Interference with production and/or quality[g,h]	0.56	0.34, 0.77	91	79	n/a	n/a
Association with work station or work activity	0.45	0.11, 0.67	82	68	.84	.77
Worst symptom area[i]	0.35	0.23, 0.46	49	21	n/a	n/a
Treatment	0.64	0.42, 0.87	83	51	.57	.59
Steroid injection	0.80	0.62, 0.97	99	96	.03	.02
Anti-inflammatory medication	0.76	0.59, 0.92	92	67	.19	.22
Icing	0.60	0.43, 0.77	96	89	.06	.06
Rest	0.59	0.43, 0.76	86	66	.19	.24
Physical therapy	0.59	0.42, 0.75	97	93	.04	.03
Surgery	Undefined	—			0	0

[a]95% Confidence Intervals for $\kappa = \kappa \pm 1.96\ (SE_\kappa)$.
[b]P_1 = Prevalence of condition reported in Round 1; P_2 = Prevalence of condition reported in Round 2. When $P_1 = 0$ or $P_2 = 0$, then kappa is undefined. Unless noted, differences in the prevalences of signs reported in Round 1 and 2 are not statistically significant at $\alpha = 0.05$ level with the McNemar χ^2 test for independent proportions.
[c]Distal symptoms in wrist, hand, or finger.
[d]Number of separate episodes: 3–12 episodes, 13–36 episodes, 37–52 episodes, 53–150 episodes, more than 150 episodes, or continuous.
[e]Duration of episodes: 1 hr or less, more than an hour but less than a day, more than a day but less than a week, more than a week but less than a month, or more than a month.
[f]Hand diagram categories: classic, probable, possible, or unlikely median mononeuropathy.
[g]Weighted kappa (quadratic weight).
[h]Interference: 1 = strongly agree to 5 = strongly disagree.
[i]Worst symptom area: right wrist, left wrist, right hand, left hand, right fingers, left fingers.

porting, misunderstanding the question, resolution of the problem, or overlooking the item.

Another concern in reliability testing is that the underlying condition changed. That is, respondents may have had a true change (for better or worse). In such a case, what appears to reflect poor reliability may be a true change in state. For example, regarding the "co-worker support" scale, the possibility exists that the items exhibited poor reliability due to the high turnover rate in the cohort under study with an actual change in co-workers.

Memory is an unavoidable problem in re-test situations. Subjects may remember how they answered questions during Round 1, and attempt to reproduce those answers during Round 2. A 3-week interval between Rounds was chosen, in part, to minimize disruption of the work process, and to optimize the trade-off between too short an interval, which might overestimate the reliability (due to the influence of memory), and too long an interval, which might underestimate reliability (due to an actual change in the underlying condition).

TABLE VII. Reliability of Upper Extremity Discomfort Ratings Among Keyboard Operators

	All subjects					Subjects with symptoms[b]				
Visual analog scale ratings[a]	**Number**	**Round 1 mean (SD)**	**Round 2 mean (SD)**	**Paired t *P*-value**	**Intraclass correlation**	**Number with Sx**	**Round 1 mean (SD)**	**Round 2 mean (SD)**	**Paired t *P*-value**	**Intraclass correlation**
Neck/shoulder/upper arm current discomfort	138	1.1 (1.9)	1.5 (2.3)	0.04	0.59	79	2.0 (2.2)	2.6 (2.6)	0.04	0.44
Discomfort in last 30 days	138	2.8 (3.3)	2.9 (3.3)	0.68	0.77	79	4.9 (2.9)	5.0 (2.9)	0.68	0.48
Elbow/forearm current discomfort	136	0.5 (1.4)	0.6 (1.4)	0.47	0.66	48	1.4 (2.1)	1.6 (2.0)	0.47	0.54
Discomfort in last 30 days	136	1.9 (3.0)	1.7 (2.9)	0.31	0.77	48	5.1 (2.8)	4.7 (3.0)	0.31	0.32
Wrist/hand/fingers current discomfort	137	1.2 (2.1)	1.6 (2.4)	< 0.01	0.79	90	1.8 (2.4)	2.4 (2.6)	< 0.01	0.74
Discomfort in last 30 days	137	3.3 (3.4)	3.3 (3.3)	0.64	0.80	90	5.0 (3.0)	4.9 (2.9)	0.64	0.59

[a]Ratings from 0 (no discomfort) to 10 (worst discomfort imaginable).
[b]Analyses restricted to subjects who reported symptoms in Round 1 or 2.

In the interim, there were no modifications in job tasks, however part of the month preceding Round 1 included the busiest season of the year for the data coding operation. The month preceding Round 2 had more typical work demands, which may have affected reporting rates. Even though workflow was higher at baseline compared to 3 weeks later, the psychosocial scales related to job content (job dissatisfaction, intellectual discretion, and creative skill), and the Perceived Stress Scale were relatively stable. Furthermore, the self-reported ratings of peak discomfort during the 30 days before each survey were remarkably stable, and suggested that work demands did not significantly influence the outcomes.

Implications for Research

In comparison to other research methods among workers, items on the questionnaire were often as reliable as, or more reliable than certain median and ulnar nerve conduction tests [Salerno et al., 1999], and more reliable than results from physical examination of the upper extremity [Salerno et al., 2000]. In addition, the questionnaire allowed identification of information not obtainable or difficult to obtain via clinical signs (e.g., pain or discomfort). For example, in this study, although less than 25% had abnormal clinical findings (data not shown), over 50% of workers indicated they had symptoms.

Clearly, the study shows that questionnaire design merits careful attention since not all items were reliable. For example, the number and duration of episodes of specific problems were not the most reliable items reported. This has important implications since frequency and duration of symptoms are useful to gauge upper extremity disorders [Verbrugge and Ascione, 1987; Hales et al., 1994; Harrington et al., 1998; Norlander and Nordgren, 1998]. Good data refinement can optimize time, and allow easier comparison of findings across studies through retention of items with higher prevalence and reliability.

Age, sex, education, employment status, tenure, keystrokes per hour, and total hours worked were not statistically significant factors in modeling variation of reliability. Workforce characteristics, such as high turnover, may limit the reliability of certain psychosocial factors. Repeated use of the questionnaire with similar results among industrial workers and keyboard operators suggests the extent to which results are applicable to a larger working population. The high kappa and ICC values provide support for the reliability of the instrument.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the workers and management who participated in the study, and express our appreciation to Milt Brouwer, Mike Gerard, Mari Hagen,

TABLE VIII. Reliability of Job Content and Perceived Stress Scales Among Keyboard Operators

Scale (n = 138)	Round 1 Mean (SD) [Range]	Round 2 Mean (SD) [Range]	ICC[a]
Job dissatisfaction (n = 137)	0.3 (0.2) [0–0.8]	0.3 (0.2) [0–0.8]	0.88
Perceived stress	23 (8) [4–51]	23 (9) [6–51]	0.88
Intellectual discretion	21 (6) [8–40]	21 (6) [12–44]	0.82
Creative skill	6 (2) [1–12]	6 (2) [3–12]	0.78
Supervisor support (n = 123)	11 (2) [5–16]	11 (2) [5–18]	0.69
Decision amount	20 (8) [12–48]	20 (7) [12–40]	0.63
Job insecurity	6 (2) [3–13]	6 (2) [3–13]	0.58
Co-worker support	11 (2) [6–16]	11 (1) [5–14]	0.44

[a]ICC, Intraclass correlation coefficient.

TABLE IX. Status of Symptoms Among Keyboard Operators

	Stable		Fluctuating	
Body area	Symptoms (%)	No symptoms (%)	.lo Sx Rd 1 – Sx Rd 2[a] (%)	Sx Rd 1 – No Sx Rd 2[b] (%)
Neck	54 (39)	69 (50)	8 (6)	7 (5)
Right shoulder	41 (30)	74 (54)	15 (11)	8 (5)
Left shoulder	34 (24)	80 (58)	12 (9)	12 (9)
Right upper arm	14 (10)	103 (75)	10 (7)	11 (8)
Left upper arm	15 (11)	107 (78)	9 (6)	7 (5)
Right elbow	17 (12)	114 (83)	3 (2)	4 (3)
Left elbow	13 (9)	120 (87)	1 (1)	4 (3)
Right forearm	31 (22)	96 (70)	5 (4)	6 (4)
Left forearm	31 (22)	99 (72)	3 (2)	5 (4)
Right distal symptoms	57 (41)	66 (49)	9 (6)	6 (4)
Left distal symptoms	45 (33)	69 (50)	13 (9)	11 (8)
Neck/shoulder/upper arm problems	62 (45)	59 (43)	10 (7)	7 (5)
Elbow/forearm problems	43 (31)	88 (64)	2 (1)	5 (4)
Wrist/hand/finger problems	82 (60)	47 (34)	6 (4)	3 (2)

[a]No symptoms (Sx) in Round 1, but symptoms in Round 2.
[b]Symptoms in Round 1, but no symptoms in Round 2.

Wendi Latko, and Randy Rabourn, for assistance in the medical field studies. For expert statistical advice, we thank Ronán Conroy, and Laura H. Weiss, PhD.

REFERENCES

Baron S, Hales T, Hurrell J. 1996. Evaluation of symptom surveys for occupational musculoskeletal disorders. Am J Ind Med 29:609–617.

Bergner M, Rothman ML. 1987. Health status measures: an overview and guide for selection. Annu Rev Public Health 8:191–210.

Carmines EG, Zeller RA. 1979. Reliability and validity assessment. Beverly Hills and London: Sage Publications, Series No. 07-017.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46.

Cohen S, Kamarck T, Mermelstein R. 1983. A global measure of perceived stress. J Health Soc Behav 24:385–396.

Dickinson CE, Campion K, Foster AF, Newman SJ, O'Rourke AMT, Thomas PG. 1992. Questionnaire development: an examination of the Nordic Musculoskeletal Questionnaire. Appl Ergo 23:197–201.

Feinstein AR, Cicchetti DV. 1990. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 43(6):543–549.

Ferris BG. 1978. Epidemiology standardization project. American thoracic society. Am Rev Respir Dis 118(6):7–53.

Feuerstein M, Burrell LM, Miller VI, Lincoln A, Huang GD, Berger R. 1999. Clinical management of carpal tunnel syndrome: a 12-year review of outcomes. Am J Ind Med 35:232–45.

Fleiss JL. 1981. Statistical methods for rates and proportions. 2nd edn. New York: John Wiley and Sons.

Fleiss JL, Cohen J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 33:613–19.

Franzblau A, Werner RA, Albers JW, Grant CL, Olinski D, Johnston E. 1994. Workplace surveillance for carpal tunnel syndrome using hand diagrams. J Occup Rehabil 4:185–198.

Franzblau A, Salerno DF, Armstrong TJ, Werner RA. 1997. Test-retest reliability of an upper extremity discomfort questionnaire in an industrial population. Scand J Work Environ Health 23:299–307.

Gerr F, Letz R, Landrigan PJ. 1991. Upper-extremity musculoskeletal disorders of occupational origin. Annu Rev Publ Health 12:543–566.

Hales TR, Sauter SL, Peterson MR, Fine LJ, Putz-Anderson V, Schleifer LR, Ochs TT, Bernard BP. 1994. Musculoskeletal disorders among visual display terminal users in a telecommunications company. Ergonomics 37:1603–1621.

Harrington JM, Carter JT, Birrell L, Gompertz D. 1998. Surveillance case definitions for work related upper limb pain syndrome. Occup Environ Med 55(4):264–271.

Herbert R, Janeway K, Schechter C. 1999. Carpal tunnel syndrome and workers' compensation among an occupational clinic population in New York State. Am J Ind Med 35:335–342.

Homan MM, Franzblau A, Werner RA, Albers JW, Armstrong TJ, Bromberg MB. 1999. Agreement between symptom surveys, physical examination procedures and electrodiagnostic findings for the carpal tunnel syndrome. Scand J Work Environ Health 25:115–124.

Hosmer DW Jr, Lemeshow S. 1989. Applied logistic regression. New York: John Wiley and Sons.

Karasek R, Baker D, Marxer F, Ahlbom A, Theorell T. 1981. Job decision latitude, job demands, and cardiovascular disease: a prospective study of Swedish men. Am J Public Health 71(7): 694–705.

Katz JN, Stirrat CR. 1990. A self-administered hand diagram for the diagnosis of carpal tunnel syndrome. J Hand Surg 15A(2):360–363.

Kramer MS, Feinstein AR. 1981. Clinical biostatistics LIV. The biostatistics of concordance. Clin Pharmacol Ther 29(1):111–123.

Kuorinka I, Jonsson B, Kilbom A, Vinterberg H, Biering-Sorensen F, Andersson G, Jorgensen K. 1987. Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms. Appl Ergo 18(3): 233–37.

McHorney CA. 1999. Health status assessment methods for adults: past accomplishments and future challenges. Annu Rev Public Health 20:309–335.

Müller R, Büttner P. 1994. A critical discussion of intraclass correlation coefficients. Stat Med 13:2465–76.

Norlander S, Nordgren B. 1998. Clinical symptoms related to musculoskeletal neck-shoulder pain and mobility in the cervico-thoracic spine. Scand J Rehabil Med 30:243–251.

Punnett L, van der Beek AJ. 2000. A comparison of approaches to modeling the relationship between ergonomic exposures and upper extremity disorders. Am J Ind Med 37:645–655.

Salerno DF, Werner RA, Albers JW, Becker M, Armstrong TJ, Franzblau A. 1999. Reliability of nerve conduction studies among active workers. Muscle Nerve 22:1372–1379.

Salerno DF, Franzblau A, Chung K, Schultz S, Becker M, Werner RA, Armstrong TJ. 2000. Reliability of physical examination of the upper extremity among keyboard operators. Am J Ind Med 37:423–430.

Silverstein BA, Fine LJ, Armstrong TJ. 1987. Occupational factors and carpal tunnel syndrome. Am J Ind Med 11:343–358.

Stata Corporation. 1997. Stata Statistical Software: Release 5.0 1997.

Stock SR, Cole DC, Tugwell P, Streiner D. 1996. 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 29:679–688.

Thompson WD, Walter SD. 1988. A reappraisal of the kappa coefficient. J Clin Epid 41(10):949–958.

United States Department of Labor. 2000. Workplace injuries and illnesses in 1999. Bureau of Labor Statistics, USDL 00-357.

United States Department of Labor. 2001. Lost-worktime injuries and illnesses: characteristics and resulting time away from work, 1999. Bureau of Labor Statistics, USDL 01-71.

Verbrugge LM, Ascione FJ. 1987. Exploring the iceberg: common symptoms and how people care for them. Med Care 25:539–569.