



Original article

Scand J Work Environ Health [1997;23\(4\):299-307](#)

doi:10.5271/sjweh.223

Test-retest reliability of an upper-extremity discomfort questionnaire in an industrial population

by [Franzblau A](#), [Salerno DF](#), [Armstrong TJ](#), [Werner RA](#)

The following articles refer to this text: [2001;27\(1\):49-56](#); [2001;27 suppl 1:1-102](#); [2007;33\(1\):58-65](#); [2011;37\(4\):298-306](#)

Key terms: [ergonomics](#); [kappa coefficient](#); [MSD](#); [musculoskeletal disorder](#); [reliability](#); [surveillance](#)

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/9322821



This work is licensed under a [Creative Commons Attribution 4.0 International License](http://creativecommons.org/licenses/by/4.0/).

Test-retest reliability of an upper-extremity discomfort questionnaire in an industrial population

by Alfred Franzblau, MD,^{1,2} Deborah F Salerno, MS,¹ Thomas J Armstrong, PhD,^{1,2}
Robert A Werner, MD^{1,2,3}

Franzblau A, Salerno DF, Armstrong TJ, Werner RA. Test-retest reliability of an upper-extremity discomfort questionnaire in an industrial population. *Scand J Work Environ Health* 1997;23(4):299—307.

Objectives Efforts to understand or to monitor upper-extremity musculoskeletal disorders among workers have usually involved the use of questionnaires. The goal of this study was to assess the test-retest reliability of an upper-extremity discomfort questionnaire among industrial workers.

Methods Test-retest agreement among 148 workers was analyzed using the kappa coefficient for categorical outcomes. Values of kappa greater than 0.75 are considered excellent, values between 0.40 and 0.75 are fair to good, and values of less than 0.40 represent poor agreement beyond chance alone. Test-retest results of continuous measures (eg, visual analogue scale responses) were compared with paired t-tests.

Results The test-retest reliability of the questionnaire used to elicit demographic information, medical history, exercise participation, and information on musculoskeletal symptoms among industrial workers appears to be good to excellent in most instances.

Conclusions These results suggest that most results of this discomfort questionnaire are reliable and suitable for use in epidemiologic studies. For reassurance of the robustness of these findings, similar studies should be carried out in other worker populations with this, and other, questionnaire instruments.

Key terms ergonomics, kappa coefficient, musculoskeletal disorders, reliability, surveillance.

Various reports have indicated that upper-extremity musculoskeletal disorders related to work increased dramatically in the 1980s and early 1990s (1, 2). More recently, it appears that incidence rates may have leveled off (3). Although the temporal trends may be changing, this category of disorders remains financially costly to workers and employers and is a significant cause of morbidity among afflicted workers.

Many efforts to understand or to monitor such disorders among workers have involved the use of questionnaires, sometimes in conjunction with a variety of medical screening procedures or ergonomic assessments of job tasks. Such survey tools have typically included questions pertaining to demographics, past medical

history, avocational activities, and musculoskeletal discomfort in the neck and upper extremities. One of the most widely used such symptom survey tools is the Nordic survey (4). Unfortunately, there have been few reports which describe the test-retest reliability of this or any other musculoskeletal survey instrument when applied within an employed population, and such reports have typically involved small numbers of subjects (5, 6).

As part of an ongoing medical study of active workers, we had the opportunity to readminister the questionnaire to workers at 1 study site. The present report describes the results of the test-retest analyses derived from the readministration of the questionnaire.

¹ Occupational Health Program, Department of Environmental and Industrial Health, University of Michigan, School of Public Health, Ann Arbor, Michigan, United States.

² Center for Ergonomics, Department of Industrial and Operations Engineering, University of Michigan, School of Engineering, Ann Arbor, Michigan, United States.

³ Physical Medicine and Rehabilitation Service, Ann Arbor Veterans Administration Hospital, Ann Arbor, Michigan, and Department of Physical Medicine and Rehabilitation, University of Michigan School of Medicine, Ann Arbor, Michigan, United States.

Subjects and methods

Subject recruitment

The participants in this study were recruited from among workers at a unionized plant manufacturing spark plug and engine components in midwestern United States. The plant was selected as part of a larger ongoing investigation of the relationship between workplace ergonomic exposures and upper-extremity musculoskeletal disorders. For the purposes of this ongoing investigation, certain jobs in the plant were selected on the basis of the frequency of repetitive hand movements ["low", "medium" and "high" (7)]. All the workers in selected jobs with at least 6 months' tenure in such jobs and employed on the day or afternoon shifts were invited to participate in the medical survey. For logistical reasons, only workers with "medium" repetition jobs on the nightshift were invited to participate.

There were 263 subjects eligible for, and invited to participate in, the first round of this study, and 202 (77%) were ultimately included. Eighteen of the round 1 participants were employed on the nightshift; for logistical reasons these 18 persons were not solicited to participate in round 2. One hundred and forty-eight (80%) of the remaining 184 round 1 participants completed questionnaires in round 2. Table 1 displays the demographic characteristics for the 184 subjects who completed round 1 and were eligible for round 2, and table 2 shows job-

related hand repetition categories, seniority, and symptom prevalences for the same group.

The study participants provided written informed consent which had been approved by the Human Subjects Review Committee of the University of Michigan School of Public Health. No personally identifiable results were provided to the company or union. Each participant was sent a confidential summary of their personal medical survey test results, an interpretation of the results, and recommendations for medical follow-up, if indicated. All letters to subjects were sent after completion of the data collection in round 2 of the study.

Survey procedures

This study included 2 rounds of data collection from the same cohort of workers. The round 1 medical survey procedures included a self-administered questionnaire, a physical examination focused on the upper extremities, ulnar and median sensory nerve conduction studies in both wrists, and measurement of anthropometric factors. The examiners were masked to the data collected by other members of the study team. The physical examination procedures, electrodiagnostic tests, and anthropometric factors have been described elsewhere (8); these results were not used in the present analyses.

The self-administered questionnaire focused on demographic information, prior medical conditions, current health status, and symptoms potentially related to

Table 1. Demographic characteristics of the study group.

	Age (years)		Gender		Hand dominance				Race				Education													
	Mean	SD	Male	Female	Right	Left	White	Black	Hispanic	Un-known	Less than high school	High school	Beyond high school	Unknown												
			N	%	N	%	N	%	N	%	N	%	N	%	N	%										
Round 1 (N = 184)	44.2	10.2	100	54	84	46	170	92	14	8	160	87	11	6	10	5	3	2	8	4	126	68	47	26	3	2
Round 1 only (N = 36)	44.7	10.5	21	58	15	42	36	100	x	x	27	75	5	14	2	6	2	6	3	8	23	64	9	25	1	3
Round 2 ^a (N = 148)	44.0	10.1	79	53	69	47	134	91	14	9	133	90	6	4	8	5	1	1	5	3	103	70	38	26	2	1

^a Analyses of subjects who participated in round 2 based on round 1 responses.

Table 2. Job-related hand repetition categories, seniority, and symptom prevalences of the study group.

	Hand repetition								Seniority (years)		Prevalence of symptoms					
	Low		Medium		High		Not rated		Mean	SD	Neck/shoulder/upper arm		Elbow/forearm		Wrist/hand/finger	
	N	%	N	%	N	%	N	%			N	%	N	%	N	%
Round 1 (N = 184)	79	43	18	10	85	46	2	1	18.7	9.8	80	43	57	31	78	42
Round 1 only (N = 36)	15	42	5	14	15	42	1	3	19.5	10.4	19	53	8	22	17	47
Round 2 ^a (N = 148)	64	43	13	9	70	47	1	1	18.5	9.6	61	41	49	33	61	41

^a Analyses of subjects who participated in round 2 based on round 1 responses.

upper-extremity musculoskeletal disorders. The subjects were queried about previous medical conditions in the following manner: "Have you ever been told by a medical doctor that you have any of the following problems? Please check 'No' or 'Yes'. If you don't recognize a name or are not sure, check 'No'.

A variety of medical conditions potentially related to upper-extremity musculoskeletal disorders was listed: diabetes, rheumatoid arthritis, thyroid dysfunction, ruptured cervical disk, carpal tunnel syndrome, ulnar neuropathy, other peripheral neuropathy, tendinitis, sprains, broken bones in the upper extremities, thoracic outlet syndrome, gout, rotator cuff injury, ganglion, degenerative arthritis, use of medication, or history of any surgery involving the neck or upper extremities.

The subjects were instructed to report a symptom if it had been present in at least 3 separate episodes, or 1 episode had lasted more than 1 week, in the 12 months preceding the survey. The survey queried subjects about 9 symptoms (burning, stiffness, pain, cramping, tightness, aching, soreness, tingling and numbness) in each of 15 body locations (neck and right or left of each of the following: shoulder, upper arm, elbow, forearm, wrist, hand and fingers). For purposes of the analyses, if one or more symptoms were checked for a body location, then the finding was "positive" (eg, stiffness and pain in the neck); if none of the symptoms was checked for a location, then the finding was "negative". In essence, checking 1 symptom or 9 symptoms resulted in a "positive" response for a body location. If any symptoms were present, then the subjects were asked to rate the frequency of discomfort, the impact of such discomfort on work standard or quality, whether he or she had sought a job change related to the symptoms, and whether medical treatment had been sought. The questionnaire also instructed the subjects to rate the severity of their current discomfort and their worst discomfort in the 30 days preceding the survey, in each of the 3 body regions (neck, shoulders or upper arms, elbows or forearms, and wrists, hands or fingers) using 10-cm visual analogue scales. The left and right verbal anchors for the scales were, respectively, "no discomfort" and "worst discomfort imaginable".

The women were asked to complete additional questions related to reproductive hormone status (currently pregnant, current use of birth control pills, history of double oophorectomy, history of hysterectomy, and history of natural menopause).

Because of concern and interest related to the carpal tunnel syndrome, results for symptoms in the wrists, hands, and fingers were grouped into various clinically relevant combinations. In addition to analyzing results for "all symptoms" in the wrists, hands and fingers, symptoms potentially more indicative of carpal tunnel syndrome (numbness, tingling, burning or pain in the

wrists, hands, or fingers) were also analyzed separately. The subjects were also asked to complete hand diagrams reflecting these 4 symptoms, which were then scored for the likelihood of carpal tunnel syndrome (9).

The round 1 medical screening examinations were performed by University of Michigan personnel in the plant medical department during each worker's normal work hours; no one was required to arrive early or stay late. In particular, workers were relieved from their job duties for the time required to complete the entire survey protocol (mean: approximately 70 min for the entire protocol — typically about 30 to 40 min to complete the self-administered questionnaire). Note that this time was separate from, and in addition to, regularly scheduled breaks and lunch. All data from round 1 were collected in 3 consecutive days.

The round 2 survey consisted solely of the readministration of a slightly abridged version of the questionnaire. This procedure was completed in 1 day for all subjects 21 to 23 days after the first round of data collection. During the interval between rounds, there were no modifications in job tasks, and no significant changes in work load or production rates. In order to avoid additional disruption to the work process, the workers who participated in round 2 completed the questionnaire during one of their regularly scheduled morning or afternoon breaks, or their lunch break (about 22 min each). In particular, they were not relieved of regular job duties to participate in the second round, but they also did not have to arrive early or stay late. Most participants in round 2 completed the survey in less than 10 min.

The instrument used in round 2 was identical to the questionnaire used in round 1, except for the deletion of all psychosocial questions and some of the demographic questions (eg, gender and race). Since the time available to subjects for completing the questionnaires in round 2 was more limited, it was decided to reduce the number of questions so that the participants would be more likely to finish in the time allotted. The psychosocial questions were largely derived from previously existing survey tools whose reliability had been studied (10, 11, unpublished report: Karasek RA, Pieper C, Schwartz J. Job content questionnaire and user's guide. Revision 1.1. Department of Work Environment, University of Massachusetts-Lowell, Lowell, MA 1985), so it was also felt to be less critical to include these items in round 2.

Statistical analyses

The analyses were performed using STATA for Windows, version 4.0 (12). The tests were considered statistically significant if $P \leq 0.05$. Test-retest agreement was analyzed using the kappa coefficient. Where appropriate, weighted kappa coefficients were calculated using quadratic weights included with the program, which corresponded to the intraclass correlation coefficient. Values

of kappa greater than 0.75 are considered excellent, values between 0.40 and 0.75 are fair to good, and values of less than 0.40 represent poor agreement beyond chance alone (13). It is known that the value of kappa depends, in part, on the prevalence of "positive" findings: kappa approaches zero when the true prevalence approaches zero or 100% (14). Therefore, to aid the interpretation of the kappa coefficients, the prevalences of "positive" findings are presented for rounds 1 and 2, and these findings are compared using the McNemar χ^2 test statistic. The McNemar χ^2 test statistic (and corresponding P-value) is shown only for the items which were significantly different between rounds ($P < 0.05$).

Results

All values of kappa related to prior medical conditions were in the "good" to "excellent" range. (See table 3.) Questions related to female hormone status demonstrated excellent test-retest reliability, except for current pregnancy. (See table 3.) The kappa coefficient for current pregnancy is undefined due to $P_2 = 0$. Table 3 also lists the results for hand dominance, any participation in regular exercise (yes, no), frequency of participation in regular exercise (< 1 time a month, 1 time a month but < 1

time a week, 1—2 times a week, and ≥ 3 times a week), and level of education (less than high school, high school, and more than high school). The test-retest results for these latter items were all good or excellent. Only "history of surgery involving the neck or upper extremities" demonstrated a significant change in the overall prevalence between rounds 1 and 2 ($P_1 = 31\%$ and $P_2 = 23\%$; McNemar $\chi^2 = 12$, $P = 0.0005$).

The results pertaining to symptoms involving the neck, shoulders, or upper arms are shown in table 4. The overall agreement on presence or absence of symptoms anywhere in the neck, shoulders, or upper arms was excellent (kappa = 0.76, 95% CI 0.60—0.92). Similarly, the test-retest reliability for the presence or absence of symptoms separately in the neck, right shoulder, or left shoulder was excellent or nearly so. The kappa coefficients for upper-arm symptoms were good, but the results were lower than for the neck and shoulders. When the subjects queried about difficulty in keeping up with production or one's usual standard of quality because of neck, shoulder or upper-arm symptoms, the kappa equaled 0.64 (95% CI 0.48—0.80). A similar question inquiring if the worker had asked to change to a different job in the last year because of such symptoms yielded poor results (kappa = 0.39, 95% CI 0.23—0.55). It should be noted that the prevalences of positive answers to this question were low ($P_1 = 1\%$ and $P_2 = 2\%$); this result

Table 3. Medical history questions, hobbies, exercise and education. (95% CI = 95% confidence interval, NA = not applicable)

Question or condition	Observed agreement (%)	Expected agreement (%)	Kappa coefficient	95% CI of kappa ^a	Prevalence		McNemar χ^2	P-value of χ^2
					Round 1	Round 2		
Hand dominance	99	83	0.92	0.76—1	0.09	0.09		
Diabetes	100	90	1.00	0.84—1	0.05	0.05		
Rheumatoid arthritis	99	93	0.79	0.63—0.95	0.03	0.03		
Thyroid dysfunction	100	87	1.00	0.83—1	0.07	0.07		
Ruptured disk, neck	97	82	0.85	0.69—1	0.10	0.10		
Carpal tunnel syndrome	98	76	0.92	0.75—1	0.14	0.14		
Ulnar neuropathy	99	95	0.85	0.69—1	0.03	0.02		
Other peripheral neuropathy	99	98	0.66	0.51—0.82	0.01	0.01		
Tendinitis	88	66	0.64	0.48—0.80	0.20	0.24		
Sprain	86	66	0.58	0.42—0.74	0.23	0.20		
Broken bones	93	68	0.79	0.63—0.95	0.22	0.18		
Thoracic outlet syndrome	99	98	0.66	0.51—0.82	0.01	0.01		
Gout	99	97	0.80	0.64—0.95	0.02	0.01		
Rotator cuff injury	98	95	0.56	0.42—0.71	0.03	0.01		
Ganglion	99	90	0.87	0.71—1	0.05	0.05		
Degenerative arthritis	96	89	0.65	0.48—0.81	0.07	0.05		
Use of medication	86	52	0.71	0.55—0.87	0.43	0.39		
Surgery	92	60	0.80	0.64—0.95	0.31	0.23	12	0.0005
Current Pregnancy ^b	99	99	Undefined		0.01	0		
Current use birth control pills ^b	100	89	1.00	0.76—1	0.06	0.06		
Double oophorectomy ^b	99	76	0.94	0.70—1	0.13	0.15		
Hysterectomy ^b	100	63	1.00	0.76—1	0.25	0.25		
Menopause ^b	94	70	0.79	0.55—1	0.19	0.19		
Exercise	89	50	0.77	0.61—0.93	0.52	0.57		
Exercise amount ^c	86	52	0.72 ^d	0.58—0.86	NA	NA		
Level of education ^e	98	78	0.89 ^d	0.75—1	NA	NA		

^a 95% confidence interval for $\kappa = \kappa \pm 1.96 (SE_{\kappa})$.

^b Kappa computed for women only. When prevalence₁ = 0 or prevalence₂ = 0, the kappa coefficient is undefined as indicated for "current pregnancy".

^c Exercise amount = none, < 1 time/month, < 1 time/week, 1—2 times/week, > 2 times/week.

^d Weighted kappa.

^e Level of education = less than high school, high school graduate, more than high school.

may help to explain the poor test-retest reliability. The Subjects demonstrated excellent reliability in reporting whether they felt that their symptoms were related to a particular workstation or work activity ($\kappa = 0.86$, 95% CI 0.69–1.00), but reliability was lower when asked whether such symptoms were attributable to an accident or acute injury ($\kappa = 0.55$, 95% CI 0.39–0.71). Questions pertaining to the “worst” body location (5 choices), frequency, and duration of episodes in the “worst” location (5 ordered categories), the occurrence of symptoms in the “worst” area in the last week (yes/no), current symptoms (yes/no), and having sought medical treatment (yes/no) all yielded similar (good) kappa values.

The results pertaining to the elbows and forearms are shown in table 5. The kappa coefficients for symptoms in the elbows and forearms were excellent, or nearly so. The kappa coefficient for the question eliciting symptoms in the week preceding the survey ($\kappa = 0.68$, 95% CI 0.52–0.84) was significantly better than for the question about current symptoms ($\kappa = 0.46$, 95% CI 0.30–0.62).

As shown in table 6, the repeatability of the symptoms potentially consistent with carpal tunnel syndrome (numbness, tingling, burning or pain in the wrist, hand or fingers) was excellent ($\kappa = 0.81$, 95% CI 0.65–0.97 and $\kappa = 0.78$, 95% CI 0.61–0.93 for right and left hands, respectively). The test-retest reliabilities for any

Table 4. Agreement between rounds 1 and 2 with respect to the neck, shoulder, and upper-arm symptoms. (95% CI = 95% confidence interval, NA = not applicable)

Question or condition	Observed agreement (%)	Expected agreement (%)	Kappa coefficient	95% CI of kappa ^a	Prevalence		McNemar χ^2	P-value of χ^2
					Round 1	Round 2		
Any neck shoulder or upper-arm problems	89	52	0.76	0.60–0.92	0.41	0.36		
Any neck symptoms	91	59	0.79	0.63–0.95	0.31	0.26		
Any right shoulder symptoms	93	67	0.77	0.61–0.93	0.23	0.18	4.45	0.0348
Any left shoulder symptoms	91	66	0.74	0.58–0.90	0.22	0.22		
Any right upper-arm symptoms	93	83	0.61	0.44–0.77	0.09	0.09		
Any left upper-arm symptoms	93	82	0.58	0.42–0.74	0.11	0.09		
Impact production or quality	95	87	0.64	0.48–0.80	0.06	0.08		
Asked for different job	98	97	0.39	0.23–0.55	0.01	0.02		
Related to work activity	95	67	0.86	0.69–1	0.22	0.20		
Worst symptom area	81	46	0.64	0.54–0.75	NA	NA		
Frequency of episodes ^b	90	68	0.68 ^c	0.56–0.81	NA	NA		
Duration of episodes ^d	91	66	0.73 ^c	0.60–0.86	NA	NA		
Particular inciting event	92	83	0.55	0.39–0.71	0.10	0.08		
Symptoms in last week	88	57	0.73	0.56–0.89	0.34	0.28	4.76	0.0290
Current symptoms	86	65	0.61	0.45–0.77	0.25	0.21		
Sought treatment	89	61	0.72	0.56–0.88	0.29	0.25		

^a 95% confidence interval for $\kappa = \kappa \pm 1.96 (SE_{\kappa})$.

^b Frequency of episodes in last year = 3–12, 13–36, 37–52, 53–150, > 150.

^c Weighted kappa.

^d Duration of episodes in last year: = < 1 hour, < 1 day, < 1 week, < 1 month, > 1 month.

Table 5. Agreement between rounds 1 and 2 with respect to the elbow and forearm symptoms. (95% CI = 95% confidence interval, NA = not applicable)

Question or condition	Observed agreement (%)	Expected agreement (%)	Kappa coefficient	95% CI of kappa ^a	Prevalence		McNemar χ^2	P-value of χ^2
					Round 1	Round 2		
Any elbow or forearm symptoms	93	58	0.82	0.66–0.98	0.33	0.26	11.00	0.0009
Any right elbow symptoms	93	73	0.75	0.59–0.91	0.18	0.14		
Any left elbow symptoms	96	72	0.86	0.69–1	0.18	0.16		
Any right forearm symptoms	95	81	0.75	0.59–0.91	0.12	0.09		
Any left forearm symptoms	96	83	0.76	0.61–0.92	0.11	0.07	6.00	0.0143
Impact production or quality	96	91	0.55	0.39–0.71	0.05	0.05		
Asked for different job	99	96	0.66	0.51–0.82	0.03	0.01		
Related to work activity	95	73	0.80	0.64–0.96	0.18	0.14	4.50	0.0339
Worst symptom area	88	58	0.71	0.61–0.82	NA	NA		
Frequency of episodes ^b	92	72	0.71 ^c	0.58–0.84	NA	NA		
Duration of episodes ^d	93	72	0.74 ^c	0.61–0.86	NA	NA		
Particular inciting event	95	91	0.44	0.29–0.59	0.06	0.03		
Symptoms in last week	90	70	0.68	0.52–0.84	0.20	0.16		
Current symptoms	88	77	0.46	0.30–0.62	0.14	0.12		
Sought treatment	92	78	0.62	0.47–0.78	0.15	0.10	5.33	0.0209

^a 95% confidence interval for $\kappa = \kappa \pm 1.96 (SE_{\kappa})$.

^b Frequency in last year = 3–12, 13–36, 37–52, 53–150, > 150.

^c Weighted kappa.

^d Duration of episodes = < 1 hour, < 1 day, < 1 week, < 1 month, > 1 month.

symptoms in the wrists, hands or fingers were all good. The test-retest agreement was good for the hand diagrams (using weighted kappa coefficients). The questions inquiring about nocturnal occurrence of wrist, hand and finger symptoms yielded near-excellent or excellent test-retest results.

The results of the visual analogue scale scores are summarized in table 7. The subjects rated their current discomfort significantly worse in round 2 than in round 1 for the neck, shoulders or upper arms and the wrists, hands or fingers. However, the 30-day discomfort ratings were indistinguishable for the 2 rounds for all the

body regions. Most of the subjects reported no discomfort whatsoever in all the body regions in both rounds, a finding which might tend to overstate the concordance of results between rounds. Hence the analyses for each body region (ie, paired t-tests) were restricted to subjects who reported symptoms in at least 1 round. The results based on this "symptomatic" subset of the study group again demonstrated that "current" discomfort ratings are less stable than 30-day discomfort ratings.

One possible approach to analyzing and presenting data from the questionnaire used in this study would be to present individual symptoms for each body location,

Table 6. Agreement between rounds 1 and 2 with respect to the wrist, hand, and finger symptoms. (95% CI = 95% confidence interval, NA = not applicable, N/T/B/P = symptoms of numbness, tingling, burning, or pain in the wrist, hand or fingers)

Question	Observed (%)	Expected (%)	Kappa ^a coefficient	95% CI of kappa ^b	Prevalence		McNemar χ^2	P-value
					Round 1	Round 2		
Wrist, hand or finger symptoms	93	52	0.84	0.68—1.00	0.41	0.36	4.45	0.0348
Right distal N/T/B/P	92	57	0.81	0.65—0.97	0.32	0.30		
Left distal N/T/B/P	91	61	0.78	0.61—0.93	0.28	0.26		
Right wrist symptoms	91	71	0.69	0.53—0.85	0.18	0.16		
Left wrist symptoms	90	74	0.61	0.45—0.77	0.16	0.14		
Right hand symptoms	89	67	0.65	0.49—0.81	0.22	0.20		
Left hand symptoms	92	69	0.74	0.58—0.90	0.20	0.18		
Right finger symptoms	86	62	0.65	0.49—0.81	0.28	0.23		
Left finger symptoms	88	66	0.64	0.48—0.80	0.24	0.20		
Left hand diagram ^c	87	69	0.59 ^d	0.48—0.70	0.18	0.18		
Right hand diagram ^c	82	63	0.52 ^d	0.41—0.63	0.22	0.21		
Impact production or quality	96	91	0.55	0.39—0.71	0.05	0.04		
Asked for different job	98	98	Undefined		0	0.02		
Related to work activity	89	70	0.62	0.46—0.78	0.18	0.19		
Nocturnal occurrence	92	63	0.78	0.62—0.94	0.27	0.22		
Right	93	69	0.78	0.62—0.94	0.21	0.17		
Left	93	74	0.74	0.58—0.90	0.17	0.14		
Worst symptom area	76	44	0.56	0.47—0.65	NA	NA		
Frequency of episodes ^e	73	42	0.54 ^d	0.45—0.63	NA	NA		
Duration of episodes ^f	78	41	0.63 ^d	0.54—0.72	NA	NA		
Particular inciting event	95	92	0.44	0.28—0.59	0.05	0.03		
Symptoms in last week	91	58	0.79	0.63—0.95	0.30	0.31		
Current symptoms	88	71	0.60	0.44—0.76	0.18	0.17		
Sought treatment	93	81	0.60	0.45—0.76	0.12	0.09		

^a When the prevalence = 0, the kappa coefficient is undefined, as indicated for "asked for different job".

^b 95% confidence interval for $\kappa = \kappa \pm 1.96 (SE_{\kappa})$.

^c Hand diagram categories = classic, probable, possible, unlikely.

^d Weighted kappa.

^e Frequency of episodes in last year = 3—12, 13—36, 37—52, 53—150, > 150.

^f Duration of episodes in last year = < 1 day, < 1 week, < 1 month, > 1 month.

Table 7. Upper-extremity discomfort among the subjects according to the visual analogue scale ratings.

Question	All subjects						Subjects with symptoms ^a					
	Number	Round 1		Round 2		Paired t P-value	Number	Round 1		Round 2		Paired t P-value
		Mean	SD	Mean	SD			Mean	SD	Mean	SD	
Neck, shoulders or upper arms												
Current discomfort	145	1.08	2.06	1.43	2.44	0.0171	57	2.75	2.49	3.64	2.68	0.0163
Worst discomfort in last 30 days	145	2.24	3.23	2.22	3.44	0.9101	61	5.33	2.89	5.28	3.45	0.9108
Elbows or forearms												
Current discomfort	148	0.63	1.41	0.68	1.83	0.6494	47	1.98	1.89	2.15	2.73	0.6528
Worst discomfort in last 30 days	148	1.31	2.50	1.33	2.67	0.9197	45	4.31	2.78	4.37	3.20	0.9206
Wrists, hands or fingers												
Current discomfort	147	0.85	2.01	1.16	2.23	0.0383	53	2.38	2.76	3.21	2.70	0.0375
Worst discomfort in last 30 days	147	2.03	3.06	2.03	3.15	0.9967	59	5.05	2.82	5.05	3.06	0.9967

^a Analyses restricted to subjects who reported symptoms in round 1 or round 2.

rather than "lumping" responses together. (See the methods section.) A summary of itemized symptoms for the neck are shown in table 8. Although the kappa coefficients for each of the 9 symptoms was significant (ie, all confidence intervals exclude 0), they varied from 0.26 (poor) to 0.78 (excellent). Such a wide range of kappa values was demonstrated for all the other body locations (shoulders, upper arms, elbows, forearms, wrists, hands, and fingers — data not shown). There was no consistent pattern of the kappa coefficients for any of the symptoms among the body locations (ie, "soreness" did not consistently have the highest kappa value, and "tingling" did not always have the lowest kappa value).

A concern, particularly when questionnaires are used in epidemiologic studies, is whether, or how, the test-retest reliability of survey results may be related to demographic or ergonomic exposure covariates. The issue is whether any of these covariates might differentially impact on the agreement of responses across trials. As an example, consider the variable "any neck-shoulder-upper arm problems" (table 4). We constructed a new variable as follows: "neck, shoulder or upper-arm problems" = 1 if there was agreement of responses from rounds 1 and 2; "neck, shoulder or upper-arm problems" = 0 if there was disagreement of the responses from rounds 1 and 2. "Neck, shoulder or upper-arm problems" was then used as the dependent variable in a multiple logistic regression model with the following covariates: age, gender, repetition category, level of education (as reported in round 1), and seniority. This model was not significant, and none of the individual odds ratios were either (not shown). This finding suggests that agreement between rounds for "any neck, shoulder or upper-arm problems" was not related significantly to any of the covariates in the logistic model. We constructed all similar symptom response variables and examined each in a multiple logistic regression model using the same covariates (age, gender, repetition category, level of education, and seniority). None of the models were significant, except for 1 isolated finding for gender and age as covariates for right upper-arm symptoms (not shown).

Discussion

The focus of this study was the assessment of the reliability of a discomfort survey employed in field studies of upper-extremity musculoskeletal disorders. "Reliability" in this instance refers to the extent to which a measurement procedure yields the same result on repeated trials. Obviously, with more consistent results obtained on repeated trials, it is implied that the measurement procedure is more reliable.

The approach employed in this study, direct re-testing of the same subjects with essentially the same survey

Table 8. Agreement between rounds 1 and 2 with respect to the questions concerning individual symptoms for the neck. (95% CI = 95% confidence interval)

Question	Observed (%)	Expected (%)	Kappa coefficient	Kappa 95% CI ^a
Burning	91	85	0.43	0.27—0.59
Stiffness	86	65	0.63	0.46—0.78
Pain	89	69	0.63	0.47—0.79
Cramping	95	90	0.51	0.35—0.67
Tightness	89	70	0.64	0.48—0.80
Aching	90	68	0.68	0.52—0.84
Soreness	93	69	0.78	0.62—0.94
Tingling	86	81	0.26	0.13—0.40
Numbness	97	95	0.27	0.11—0.43

^a 95% confidence intervals for $\kappa = \kappa \pm 1.96 (SE_{\kappa})$.

instrument, is intuitively appealing. However, there are potential problems and limitations with this approach. First, during the time interval between survey administrations, the underlying concepts being measured may have changed (ie, subjects may have developed, or resolved, upper-extremity discomfort) (15). Hence, what appears to be a lack of consistency between trials may actually reflect a true change of the state of a subject, rather than measurement error. One solution to this problem would be to minimize the interval between test administrations, which would reduce the likelihood of any changes in the underlying concepts being measured.

A second issue in test-retest trials is "memory", or the extent to which subjects remember their responses from the initial trial, and simply reproduce them on subsequent trials (15). Hence a short interval between trials may yield inflated estimates of the consistency of responses.

There is no survey procedure which can eliminate the influence of "memory" or a true change of state between trials, but one can attempt to minimize them. We chose to readminister the survey after 3 weeks. During this interval, there were no changes in the production processes, or production rates, and none of the workers changed job titles. Therefore, we believe that these circumstances minimized the likelihood that study participants may have experienced a true change in state with regard to upper-extremity musculoskeletal discomfort. This interval also seems reasonable for reducing the influence of "memory". In essence, there is a trade-off between using a shorter or longer interval between test administrations, and the 3-week interval appears to balance the potential effects of "memory" and change of state.

Another factor that can influence test-retest trials is "reactivity", or the extent to which the measurement process can influence the initial measurement, or future measurements (15). So, for example, the initial measurement may "sensitize" a subject, and such sensitization may influence future responses (eg, subjects may be more or less likely to report musculoskeletal symptoms in the

second trial, perhaps related to increased awareness). Sensitization may contribute to over or underreporting in follow-up trials; it is difficult to predict how a longer or shorter interval between trials influences the impact of sensitization on our results. By presenting, and comparing, the prevalence of positive responses in each trial (P_1 and P_2), we attempted to gain insight into the degree of "reactivity" or "sensitization". It should also be noted that the circumstances of questionnaire administration differed slightly between rounds 1 and 2. The subjects were on "company time" during round 1, while the study participants were on their personal break time during round 2, and it was observed that the subjects completed the questionnaire much more rapidly in the second round. Overall, few of the comparisons between P_1 and P_2 differed significantly. This finding suggests that these factors (ie, circumstances of questionnaire administration, "reactivity", or "sensitization") did not have much influence on the observed results. For all the items in which P_1 and P_2 differed significantly, there was a decline in the prevalence of the item reported (ie, $P_1 > P_2$). It is our impression that the circumstances of survey administration which led to more rapid completion of the questionnaires in round 2 probably contributed to these observed findings.

Overall, kappa coefficients for the presence or absence of symptoms were all good or excellent. The results of the discomfort ratings on the 10-cm visual analogue scale were very stable when the subjects were asked to rate their "worst" discomfort in the 30 days preceding the survey. Ratings based on "current" discomfort at the time of the survey appeared to change significantly between rounds; these latter findings may be related to "reactivity" or "sensitization", as previously discussed, or possibly reflect a true "change of state" since the intensity of musculoskeletal symptoms can fluctuate over time. Overall, the results of the visual analogue scale suggest that quantitative discomfort ratings should be "averaged" over some time period, rather than based on current symptoms alone. In the present study, 30 days appeared to work well. The responses concerning presence or absence of symptoms in the week preceding the completion of the survey also yielded good to excellent results. (See tables 4, 5, and 6.)

The test-retest agreement of the responses for individual symptoms (ie, numbness, stiffness, etc) demonstrated considerable variability. These findings suggest that questions for eliciting individual symptoms should be combined, as was done in the present instance, rather than analyzed separately.

The logistic regression models indicate that ergonomic and common demographic covariates do not appear to have any impact on the reliability of responses. Thus the consistency of reporting symptoms would appear to be unconfounded by these factors.

Dickinson et al (5) described results of the "repeatability" of the Nordic musculoskeletal questionnaire among 44 subjects (all were cashiers). The percentage of the subjects who completed both rounds, the demographic information on subjects, and other circumstances of administration (eg, was test administration performed during normal workhours) were not stated. The nonidentical answers ranged from 0% to 26%. Kuorinka et al (4) also described results of repeat administrations of the Nordic questionnaire to various groups of workers. In 1 trial using the "general questionnaire" and involving 29 safety engineers, 17 medical secretaries, and 22 railway maintenance workers, the nonidentical answers ranged from 0% to 23%. The interval between test administrations was not stated. Another trial using the "neck-shoulder questionnaire" involved 27 female clerical workers who completed the questionnaire twice in a 3-week interval. The percentages of nonidentical answers ranged from 0% to 30%. A third trial assessed the repeatability of the "low-back questionnaire" among 25 nurses (gender not stated). The interval between test administrations was 15 days. The nonidentical answers ranged from 0% to 25%, with a mean of 4.4%. Aside from being small and providing limited information on the study groups, neither of these studies made adjustments for chance agreement (ie, use of kappa coefficient or similar technique), and so it is not possible to make direct comparisons with results in the present study. On the basis of these studies, it is also not possible to determine the extent to which the test-retest reliability of Nordic survey(s) is better than what would occur by chance alone (13). We are not aware of any other published studies which describe results of the readministration of a discomfort survey to workers.

Concluding remarks

The test-retest reliability of the questionnaire used to elicit demographic information, medical history, exercise participation, and musculoskeletal symptom-related information among industrial workers appears to be good to excellent. These results suggest that most of the results of this discomfort questionnaire are reliable and suitable for use in epidemiologic studies. For reassurance of the robustness of these findings, similar studies should be replicated in other worker populations with this, and other, questionnaire instruments.

Acknowledgments

We would like to thank James Albers, Robin Baver, Cynthia Garcia, Christin Grant, Elizabeth Johnston,

Joseph Kearns, Wendi Latko, Bryan Nakfoor, Susan Nalepa, Caroline Serna, Jill Sheiman, and Patricia Strasser for their assistance with the medical field studies. We also thank the workers and managers at the plant who made this study possible.

This study was supported by grant number 1 R01 OH02941—02 from the National Institute for Occupational Safety and Health (NIOSH) and also by a gift from the Office Ergonomics Research Committee (OERC). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIOSH or OERC.

References

1. Bureau of Labor Statistics. Occupational injuries and illnesses in the United States by industry, 1989. Washington (DC): US Government Printing Office, 1991. US Department of Labor Bulletin 2379.
2. Hanrahan LP, Higgins D, Anderson H, Haskins L, Tai S. Project SENSOR: Wisconsin surveillance of occupational carpal tunnel syndrome. *Wis Med J* 1991;90(2):80, 82—83.
3. Brogmus GE. Reporting of cumulative trauma disorders of the upper extremities may be leveling off in the US. In: Proceedings of the Human Factors and Ergonomics Society, 39th Annual Meeting; 1995. Santa Monica (CA). Human Factors and Ergonomics Society, 1995:591—5.
4. Kuorinka I, Jonsson B, Kilbom A, Vinterberg H, Biering-Sorensen F, Andersson G, et al. Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms. *Appl Ergon* 1987;18(3):233—7.
5. Dickinson CE, Campion K, Foster AF, Newman SJ, O'Rourke AMT, Thomas PG. Questionnaire development: an examination of the Nordic musculoskeletal questionnaire. *Appl Ergon* 1992;23(3):197—201.
6. Viikari-Juntura E. Interexaminer reliability of observations in physical examinations of the neck. *Phys Ther* 1987;67(10):1526—32.
7. Latko WA, Armstrong TJ, Foulke JA, Herrin GD, Rabourn RA, Ulin SS. Development and evaluation of an observational method for assessing repetition in hand tasks. *Am Ind Hyg Assoc J* 1997;58:278—85.
8. Franzblau A, Werner R, Valle J, Johnston E. Workplace surveillance for carpal tunnel syndrome: a comparison of methods. *J Occup Rehabil* 1993;3(1):1—14.
9. Franzblau A, Werner RA, Albers JW, Grant CL, Olinski D, Johnston E. Workplace surveillance for carpal tunnel syndrome using hand diagrams. *J Occup Rehabil* 1994;4(4):185—98.
10. Karasek RA, Baker D, Marxer F, Ahlbom A, Theorell T. Job decision latitude, job demands, and cardiovascular disease: a prospective study of Swedish men. *Am J Public Health* 1981; 71:694—705.
11. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. *J Health Soc Behav* 1983;24(4):385—96.
12. STATA Corporation. STATA for Windows, version 4.0. College Station (TX): STATA Corporation, 1995.
13. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York (NY): John Wiley & Sons, 1981.
14. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol* 1988;41(10):949—58.
15. Carmines EG, Zeller RA. Reliability and validity assessment. Beverly Hills, London: Sage Publications, 1979. Sage University paper series on quantitative applications in the social sciences, series no 07—017.

Received for publication: 22 July 1996