ABSTRACT: Nerve conduction studies play an important role in clinical practice and research. Given their widespread use, reliability of tests merits careful attention. We assessed interexaminer and intraexaminer reliability of median and ulnar sensory nerve measures of amplitude, onset latency, and peak latency. In a two-phase cross-sectional study, two examiners tested 158 workers. Reliability was assessed with intraclass correlations (ICC) and kappa statistics. Median nerve measures were more reliable (ICC range, 0.76 to 0.92) than ulnar measures (ICC range, 0.22 to 0.85). Ulnar-onset latencies had the worst reliability. The median-ulnar peak latency difference was a particularly stable measure (ICC range, 0.79 to 0.92). The median-ulnar peak latency difference had high interexaminer reliability (κ range, 0.71 to 0.79) for normal tests defined by cut points of 0.8 ms and 0.5 ms. Intraexaminer reliability was higher with the 0.8-ms cut point ($\kappa = 0.90$ and $\kappa = 0.85$ for examiners 1 and 2, respectively). Rather than absolute cut points to describe normality, a more rational interpretation of results can be made with ordered categories or continuous measures.
 Muscle Nerve **22:** 1372–1379, 1999

RELIABILITY OF NERVE CONDUCTION STUDIES AMONG ACTIVE WORKERS

DEBORAH F. SALERNO, PhD,[1] ROBERT A. WERNER, MD,[1–4] JAMES W. ALBERS, MD, PhD,[1,5] MARK P. BECKER, PhD,[6] THOMAS J. ARMSTRONG, PhD,[2,7] and ALFRED FRANZBLAU, MD[1,2]

[1] Department of Environmental and Industrial Health, School of Public Health, The University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029 USA
[2] Center for Ergonomics, School of Engineering, The University of Michigan, Ann Arbor, Michigan, USA
[3] Physical Medicine and Rehabilitation, Veterans Administration Hospital, Ann Arbor, Michigan, USA
[4] Department of Physical Medicine and Rehabilitation, School of Medicine, The University of Michigan, Ann Arbor, Michigan, USA
[5] Department of Neurology, The University of Michigan Medical Center, Ann Arbor, Michigan, USA
[6] Department of Biostatistics, School of Public Health, The University of Michigan, Ann Arbor, Michigan, USA
[7] Department of Industrial and Operations Engineering, School of Engineering, The University of Michigan, Ann Arbor, Michigan, USA

Accepted 7 June 1999

Nerve conductions studies (NCS) are a standard procedure for evaluation of peripheral neuropathy. Increased use of NCS in clinical trials and research, and attention to quality in health care has heightened interest in the reliability of results.[20,26] Results may be used as a basis for diagnosis and in preplacement examinations for work restrictions, such as those related to carpal tunnel syndrome (CTS).[4] A variety of tests may be performed to detect sensory abnormalities, which are typical findings in early stages of disease.[6] Although methods for evaluating nerve function have evolved since the 1940s,[10,22,23] reliability has rarely been assessed, particularly among workers.

In studies of 20 diabetic patients with varying degrees of neuropathy, Dyck et al.[14] found better reliability among measures of motor and sensory nerve amplitude compared with conduction velocity and distal latency measures. Valensi et al.[31] studied a group of 132 diabetic patients, and found good re-

Abbreviations: CTS, carpal tunnel syndrome; ICC, intraclass correlation coefficient; NCS, nerve conduction studies; SNAP, sensory nerve action potential

Key words: carpal tunnel syndrome; electrodiagnostic tests; nerve conduction studies; reliability; upper extremity disorders

Correspondence to: A. Franzblau at the Department of Environmental and Industrial Health

CCC 0148-639X/99/101372-08

liability among measures of conduction velocity and F-wave latency but poor reliability for amplitudes. Chaudhry et al.[7] tested the reliability of NCS in small samples of healthy subjects ($n = 7$) and diabetic patients ($n = 6$)[8] with mixed results. More recently, Bril et al.[3] found that adherence to strict protocols in core monitoring laboratories improved reliability in multicenter trials.

We are aware of no reliability assessments of NCS among a large sample of active workers, as in this study. We assessed interexaminer and intraexaminer reliability of median and ulnar sensory nerve measures of amplitude and of onset and peak latency.

METHODS

As part of a large, two-part medical survey among keyboard operators, interexaminer and intraexaminer reliability was assessed for measures of median and ulnar sensory nerves. In round 1, tests were performed on both wrists of each subject twice, once by each examiner. In round 2, 3 weeks later, tests were performed by one of the two examiners in the dominant wrist only.

Sensory conduction studies involved stimulation of the median and ulnar nerves at the wrist, recording from digits II and V, respectively. Antidromic stimulation was applied 14 cm proximal to standard ring recording electrodes, separated by a distance of 3 cm. The latency to both the initial deflection (onset latency) and negative peak (peak latency) and the amplitude of the sensory nerve action potential (SNAP) were recorded for all subjects in accordance with the guidelines outlined by the American Association of Electrodiagnostic Medicine.[1]

Midpalm temperature was recorded at the beginning of testing, and subjects with cool hands were warmed to at least 32.0°C, when possible. No needle or surface electromyography was performed. (Although the distinction is sometimes blurred, needle electromyograph abnormalities are late findings in CTS.[5]) Tests were conducted with standard techniques on Teca TD-20 (Pleasantville, New York) or Nicolet Compass (Madison, Wisconsin) equipment.

The examiners included a board-certified neurologist (JWA) and physiatrist (RAW), both of whom were also board-certified in electrodiagnostic medicine. Two electrodiagnostic technologists assisted in testing, each working under the direct supervision of one examiner. The technologists prepared the subjects, applied electrodes, and made distance and temperature measurements.

On all test occasions, examiners were masked to each other's results and results obtained in other parts of the survey. The complete medical survey included physical examination of the neck and upper extremities, self-administered questionnaires of symptoms and functional activities, and anthropometric measurements.

All subjects in the medical survey provided written informed consent that had been approved by the University of Michigan Human Subjects Review Committee. All subjects were active workers, and all aspects of the survey were performed on company time during normal work hours for each subject. No personally identifiable information was provided to the union or management.

Statistical Analysis. Initial assessment of reliability was performed with Pearson product-moment correlations as measures of association, in contrast to measures of agreement,[25] as it is known that observations may disagree sharply and yet still be highly correlated. Paired *t*-tests were also used to compare examiners, not as a measure of agreement, but to see whether overall, they obtained the same mean measurement. Intraclass correlation coefficients (ICC) were used as true measures of agreement between examiners. The ICC combines a measure of correlation with a test of the difference of means.[24]

In addition, analyses were performed for reliability of normal ratings. We constructed two-by-two tables of the median-ulnar peak latency difference, defining normal with two cut points, 0.5 ms and 0.8 ms, respectively. Values less than the cut point were classified as normal. For the dichotomous data, reliability was assessed with the kappa statistic,[9] a measure of agreement corrected for chance, defined as:

$$\kappa = (p_{\text{observed}} - p_{\text{expected}})/(1 - p_{\text{expected}}),$$

where p_{observed} is the observed proportion of agreement, and p_{expected} is the expected proportion of agreement. The McNemar chi-squared statistic was used to compare proportions. Values of kappa greater than 0.75 were considered excellent; values between 0.40 and 0.75 were fair to good; and values of less than 0.40 represented poor agreement beyond chance.[15]

Statistical analyses were performed using Stata Statistical Software.[30] Tests were considered statistically significant at the $\alpha = 0.05$ level.

Temperature Correction. Temperature is one of the most important factors explaining variation in nerve conduction and the most frequent cause of misclassification of false-positive or borderline findings.[12] Even with control of temperature within an acceptable range, it is an important covariate of latency.[29] There is a decrease in latency by 0.3 ms for

every 1°C increase in temperature,[23] with an approximately linear relationship between velocity and temperature from 18°C to 36°C.[11]

Examiners were not assessed for reliability of skin temperature measurements. Instead, after the initial analyses were completed, a temperature correction was applied to latency measurements. Onset and peak latencies were adjusted for temperature differences according to the following formula, taking 35°C as the standard skin temperature:

$$\text{latency}_{\text{corrected}} = \text{latency}_{\text{initial}} - 0.3 \text{ ms } (35°\text{C} - \text{temp}°\text{C}).$$

Temperature correction was not applied to measurements of amplitude, as no significant change in amplitude has been reported with change in skin temperature.[13]

RESULTS

One hundred sixty-one workers completed the medical survey in round 1, for a participation rate of 76%. Three weeks later, 138 subjects returned for round 2. Interexaminer reliability was analyzed with data from 158 subjects in round 1 (three subjects' data were lost for one examiner in round 1). Four subjects declined NCS in round 2, so data from 134 subjects were available for analysis. Intraexaminer reliability was analyzed with data from 58 subjects tested by examiner 1 and 76 subjects tested by examiner 2.

The average age of subjects was 35 years (range, 20 years to 58 years). Most subjects were right-handed (91%) and women (91%). The majority of participants in the medical survey were permanent employees, with a mean of 1.4 years (range, 0.4 years to 1.6 years) of seniority. There were no statistically significant demographic differences between subjects who completed both rounds compared with those who participated in round 1 only.

Between the two examiners, 36 (22%) of the subjects studied in round 1 had abnormal sensory studies in the dominant or nondominant hand using the 0.5-ms cut point; 16 (10%) of the subjects had abnormal studies using the 0.8-ms cut point. On physical examination (n = 160), 57 (36%) subjects had positive Phalen, Tinel, or carpal compression tests. Ninety-nine (61%) respondents reported wrist, hand, or finger symptoms; 86 (53%) had numbness, burning, tingling, or pain; 39 (24%) had nocturnal symptoms; and 23 (14%) indicated classic or probable indications of CTS on a hand diagram.[17] Thirty-one (19%) subjects reported substantial decrements in functional activity.

Interexaminer Reliability. Overall, between the two examiners, median sensory nerve measures were more reliable (ICC range, 0.82 to 0.91) than ulnar measures (ICC range, 0.33 to 0.85) (Table 1). Amplitude and peak latency had higher interexaminer reliability (ICC range, 0.63 to 0.91) than did onset latency (ICC range, 0.33 to 0.87). The median-ulnar peak latency difference had consistently high reliability (ICC = 0.89, dominant hand). Ulnar onset latency had the poorest reliability (ICC = 0.35, dominant hand). The pattern of ICC results was the same for measures in the nondominant hand.

As noted, it was not a study objective to assess the reliability of temperature measures. Instead, onset and peak latency measures were corrected for temperature differences (amplitudes were not adjusted). Temperature correction improved reliability

Table 1. Interexaminer summary statistics and reliability results for sensory nerve conduction studies.

Parameter	Examiner 1		Examiner 2		Pearson correlation	Paired *t*-test (*P* value)*	Intraclass correlation
	Mean (SD)	Median (range)	Mean (SD)	Median (range)			
Dominant hand (n = 157)							
Median SNAP amplitude (μv)	45.2 (14.7)	44.8 (4.9–97.0)	44.7 (14.6)	43.8 (6.8–95.7)	0.87	NS	0.87
Median SNAP onset latency (ms)	2.5 (0.4)	2.4 (1.8–4.7)	2.6 (0.4)	2.5 (2.0–4.5)	0.87	0.01	0.87
Median SNAP peak latency (ms)	3.2 (0.4)	3.1 (2.6–5.4)	3.2 (0.4)	3.1 (2.6–5.6)	0.91	NS	0.91
Ulnar SNAP amplitude (μv)	41.9 (14.1)	40.5 (10.7–92.9)	42.4 (13.5)	42.0 (11.7–82.0)	0.85	NS	0.85
Ulnar SNAP onset latency (ms)	2.3 (0.2)	2.3 (1.5–2.9)	2.4 (0.2)	2.4 (2.0–3.0)	0.41	<0.01	0.35
Ulnar SNAP peak latency (ms)	3.0 (0.2)	3.0 (2.5–3.8)	3.0 (0.3)	3.0 (2.4–4.0)	0.64	NS	0.63
Median-ulnar peak latency difference (ms)	0.2 (0.4)	0.1 (−0.4–2.1)	0.2 (0.4)	0.1 (−0.6–2.4)	0.89	NS	0.89
Hand temperature (°C)	33.1 (0.9)	33.0 (28.5–35.0)	33.1 (1.0)	33.0 (29.0–35.2)			
Nondominant hand (n = 158)							
Median SNAP amplitude (μv)	51.2 (16.5)	50.8 (7.8–99.6)	49.2 (15.8)	47.8 (9.0–96.7)	0.83	0.01	0.82
Median SNAP onset latency (ms)	2.5 (0.3)	2.4 (1.8–4.8)	2.5 (0.3)	2.4 (2.0–4.9)	0.87	NS	0.87
Median SNAP peak latency (ms)	3.2 (0.4)	3.0 (2.5–6.5)	3.2 (0.4)	3.1 (2.6–6.3)	0.91	NS	0.91
Ulnar SNAP amplitude (μv)	44.7 (14.1)	43.1 (7.3–92.5)	44.1 (14.1)	42.0 (14.1–85.2)	0.78	NS	0.78
Ulnar SNAP onset latency (ms)	2.3 (0.2)	2.3 (1.5–2.7)	2.4 (0.2)	2.4 (1.8–2.8)	0.35	0.01	0.33
Ulnar SNAP peak latency (ms)	3.0 (0.2)	3.0 (2.6–3.6)	3.0 (0.2)	3.0 (2.0–4.0)	0.64	NS	0.63
Median-ulnar peak latency difference (ms)	0.1 (0.4)	0.1 (−0.4–3.1)	0.1 (0.4)	0.1 (−0.4–3.1)	0.88	NS	0.88
Hand temperature (°C) (n = 157)	33.0 (0.9)	33.0 (29.7–35.0)	33.0 (1.1)	33.0 (28.5–35.3)			

**NS, not statistically significant at α = 0.05 level.*

of ulnar onset latency in the dominant hand (ulnar onset latency, ICC = 0.47). Reliability of ulnar peak latency decreased slightly (ICC = 0.56). Interexaminer reliability for median latency measures was still excellent (ulnar onset latency, ICC = 0.80; peak latency, ICC = 0.86). Similar results were found in the nondominant hand.

Intraexaminer Reliability. Results of intraexaminer reliability analyses showed high congruence between examiners 1 and 2 (Table 2), with the same contrast as interexaminer results between median and ulnar nerve measures. Intraexaminer reliability was higher for median measures (ICC range, 0.76 to 0.92) than for ulnar measures (ICC range, 0.22 to 0.80). Again, measures of amplitude and peak latency had higher reliability than did onset latency. Ulnar onset latencies had the poorest reliability (ICC = 0.22 and 0.23 for examiners 1 and 2, respectively). The median-ulnar peak latency difference had excellent intraexaminer reliability (ICC = 0.92 and 0.79 for examiners 1 and 2, respectively).

Intraexaminer reliability with temperature correction reflected the same pattern of higher reliability for median than ulnar measures. For examiner 1, paradoxically, temperature correction reduced reliability of median latency measures (median onset and peak latency ICC = 0.68 and ICC = 0.79) and ulnar onset latency (ICC = 0.13) but slightly improved reliability of the ulnar peak latency measure (ICC = 0.42). For examiner 2, reliability of median latency and ulnar peak latency measures were similar to results without temperature correction (median onset latency ICC = 0.72, median peak latency ICC = 0.77, ulnar onset latency ICC = 0.20), but temperature correction slightly decreased reliability for ulnar peak latency (ICC = 0.28).

Normal Rating Determinations. Two cut points, 0.5 ms and 0.8 ms, defined normal for the median-ulnar peak latency difference (Table 3). Interexaminer reliability of the 0.8-ms cut point had slightly higher reliability ($\kappa = 0.75$ and $\kappa = 0.79$, dominant and nondominant hands, respectively) than the 0.5-ms cut point ($\kappa = 0.71$). With temperature correction, interexaminer reliability for the 0.5-ms cut point remained high ($\kappa = 0.71$); interexaminer reliability for the 0.8-ms cut point was also high ($\kappa = 0.75$ and $\kappa = 0.72$ for dominant and nondominant hands, respectively).

Intraexaminer reliability measures of examiner 1 was excellent for ratings with the 0.80-ms cut point ($\kappa = 0.90$) (Table 4). The 0.5-ms cut point had reduced, though still good, reliability ($\kappa = 0.52$). Examiner 2 had higher reliability with the 0.8-ms cut point ($\kappa = 0.85$) than the 0.5-ms cut point ($\kappa = 0.68$). With temperature correction, intraexaminer reliability had similar patterns. The 0.8-ms cut point had higher reliability ($\kappa = 0.78$ and $\kappa = 0.79$ for examiners 1 and 2, respectively) than the 0.5-ms cut point ($\kappa = 0.51$ and $\kappa = 0.72$ for examiners 1 and 2, respectively).

DISCUSSION

In earlier assessments of NCS, reliability was mixed. In one study, seven examiners (who also served as subjects) each tested four healthy subjects on two occasions.[7] High interexaminer reliability was found for eight of 12 measures (four sensory and eight motor nerve). High intraexaminer reliability was re-

Table 2. Intraexaminer summary statistics and reliability results for sensory nerve conduction studies, dominant hand.

Parameter	Round 1		Round 2		Pearson correlation	Paired *t*-test (*P* value)*	Intraclass correlation
	Mean (SD)	Median (Range)	Mean (SD)	Median (Range)			
Examiner 1 (*n* = 58)							
Median SNAP amplitude (μv)	44.9 (14.9)	44.2 (4.9–97.0)	43.4 (15.4)	42.5 (4.0–83.3)	0.88	NS	0.88
Median SNAP onset latency (ms)	2.5 (0.4)	2.4 (1.8–4.7)	2.6 (0.4)	2.4 (2.1–4.5)	0.92	NS	0.92
Median SNAP peak latency (ms)	3.3 (0.5)	3.1 (2.7–5.4)	3.2 (0.5)	3.1 (2.8–5.5)	0.92	NS	0.92
Ulnar SNAP amplitude (μv)	43.0 (14.6)	41.0 (10.7–81.7)	41.8 (16.0)	41.3 (13.4–90.4)	0.68	NS	0.68
Ulnar SNAP onset latency (ms)	2.3 (0.2)	2.3 (1.5–2.8)	2.3 (0.2)	2.3 (1.8–2.8)	0.22	NS	0.22
Ulnar SNAP peak latency (ms)	3.0 (0.2)	3.0 (2.5–3.4)	3.0 (0.2)	3.0 (2.5–3.6)	0.37	0.01	0.33
Median-ulnar peak latency difference (ms)	0.2 (0.5)	0.1 (−0.3–2.1)	0.3 (0.5)	0.1 (−0.3–2.2)	0.92	0.02	0.92
Hand temperature (°C) (*n* = 55)	33.0 (1.0)	33.0 (28.5–35.0)	33.8 (0.9)	34.0 (32.0–35.5)			
Examiner 2 (*n* = 75)							
Median SNAP amplitude (μv)	46.4 (13.9)	47.9 (11.1–82.0)	49.8 (15.2)	47.9 (17.2–84.2)	0.84	<0.01	0.81
Median SNAP onset latency (ms)	2.5 (0.3)	2.4 (2.0–4.1)	2.4 (0.3)	2.4 (1.9–3.8)	0.79	<0.01	0.76
Median SNAP peak latency (ms)	3.2 (0.4)	3.1 (2.6–4.7)	3.1 (0.4)	3.0 (2.6–5.0)	0.82	<0.01	0.80
Ulnar SNAP amplitude (μv)	43.0 (14.1)	42.0 (15.6–78.6)	44.1 (15.3)	41.4 (15.9–90.9)	0.81	NS	0.80
Ulnar SNAP onset latency (ms)	2.4 (0.2)	2.4 (2.1–3.0)	2.3 (0.2)	2.3 (2.0–3.0)	0.36	<0.01	0.23
Ulnar SNAP peak latency (ms)	3.0 (0.3)	3.0 (2.4–3.6)	2.9 (0.2)	2.9 (2.6–3.7)	0.47	0.01	0.43
Median-ulnar peak latency difference (ms)	0.2 (0.3)	0.1 (−0.6–1.6)	0.2 (0.3)	0.1 (−0.3–2.0)	0.79	NS	0.79
Hand temperature (°C) (*n* = 76)	33.0 (1.0)	33.0 (31.0–35.0)	33.4 (0.7)	33.5 (32.0–35.0)			

**NS = Not statistically significant at $\alpha = 0.05$ level.*

Table 3. Interexaminer reliability in sensory studies: median-ulnar peak latency difference.

	Abnormal ratings by examiner 1	Abnormal ratings by examiner 2	Agreement (%)	Expected Agreement (%)	Kappa (95% CI*)	P_1†	P_2‡	*P* Value§
Cut point 0.5 ms								
Dominant hand (*n* = 157)	25	25	92	73	0.71 (0.56, 0.87)	0.16	0.16	NS
Nondominant hand (*n* = 158)	19	16	94	80	0.71 (0.56, 0.87)	0.12	0.10	NS
Cut point 0.8 ms								
Dominant hand (*n* = 157)	14	12	96	85	0.75 (0.59, 0.90)	0.09	0.08	NS
Nondominant hand (*n* = 158)	7	8	98	91	0.79 (0.63, 0.95)	0.04	0.05	NS

**95% condifence intervals for $\kappa = \kappa \pm 1.96 (SE_K)$.*
†Proportion with abnormal ratings reported by examiner 1.
‡Proportion with abnormal ratings reported by examiner 2.
§NS, differences in the prevalences reported by examiner 1 and examiner 2 are not statistically significant at $\alpha = 0.05$ level with the McNemar χ^2 test for independent proportions.

ported for all measures: sural and median SNAP amplitude and conduction velocity, peroneal and median motor nerve distal latency, compound muscle action potential amplitude, conduction velocity, and the shortest F-wave latency of 10 trials. In a second study, six examiners each studied six patients with diabetic neuropathy.[8] High interexaminer reliability was reported for eight of 12 measures, with consistently lower reliability for sural SNAP and median compound muscle action potential amplitudes. High intraexaminer reliability was found for 11 of 12 measures. Although feasibility is an issue, results suggested that the same examiner perform the repeated NCS in longitudinal studies to minimize interexaminer variability.

Like the two previous studies, results of the current study were mixed, and, at times, the difference in reliability was striking. Interexaminer reliability of median sensory latency measures was excellent, but both interexaminer and intraexaminer ICCs were lower for ulnar latencies (with and without temperature correction). In addition, while interexaminer reliability of the median-ulnar peak latency difference in the dominant hand was clearly excellent, the dichotomous normal ratings were less reliable.

One reason for the difference in reliability between median and ulnar nerve activity may be due to inherent difficulties in measurement. Anatomically, the ulnar nerve is smaller than the median nerve. In consequence, the response of the ulnar nerve can have a relatively small amplitude compared with the median response, although in the present study, the mean amplitudes of the ulnar and median nerves were comparable. Also, a poorly defined takeoff of the evoked response can make it difficult to accurately identify the onset of response. Averaging was not used, although it would have increased the accuracy of determining onset latency. Because averaging was not used, the reliability of onset latency was probably underestimated.

Examiner 1 had computer-generated markings for latency, whereas examiner 2 rendered manual markings. This would not influence intraexaminer reliability, but may have contributed to interexaminer variability of the median and ulnar latency measures. This apparently had minimal effect on interexaminer results, as median latencies were consistently higher than ulnar latencies.

One factor that may have influenced intraexaminer reliability was timing. As mentioned, a 3-week interval separated rounds 1 and 2 of the study. This interval was chosen, in part, to minimize the possi-

Table 4. Intraexaminer reliability in sensory studies: median-ulnar peak latency difference.

	Abnormal ratings in round 1	Abnormal ratings in round 2	Agreement (%)	Expected agreement (%)	Kappa (95% CI*)	P_1†	P_2‡	*P* Value§
Examiner 1								
Cut point 0.5 ms; dominant hand (*n* = 58)	9	11	86	71	0.52 (0.26,0.77)	0.16	0.19	NS
Cut point 0.8 ms; dominant hand (*n* = 58)	6	5	98	83	0.90 (0.64, 1)	0.10	0.09	NS
Examiner 2								
Cut point 0.5 ms; dominant hand (*n* = 75)	12	10	92	75	0.68 (0.46,0.91)	0.16	0.13	NS
Cut point 0.8 ms; dominant hand (*n* = 75)	4	3	99	91	0.85 (0.63, 1)	0.05	0.04	NS

**95% confidence intervals for $\kappa = \kappa \pm 1.96 (SE_K)$.*
†Proportion with abnormal ratings in round 1.
‡Proportion with abnormal ratings in round 2.
§NS = Differences in the prevalences reported in round 1 and round 2 are not statistically significant at $\alpha = 0.05$ level with the McNemar χ^2 test for independent proportions.

bility of change in the underlying physical condition of workers and to minimize disruption of the work process. During that time, there were no modifications in job tasks. However, part of the month preceding round 1 included the busiest season of the year for the keyboard operation. The month preceding round 2 had more typical work demands. The worse intraexaminer reliability compared with interexaminer reliability suggests the possibility that nerve conditions may have been different, reducing apparent intraexaminer reliability. Furthermore, different technicians worked with the examiners between rounds 1 and 2. This could have led to differences in the positioning of the subject's hand during electrode placement and in distance measurement that may have adversely affected reliability.

Although many paired *t*-tests demonstrated significant statistical differences in this study, none was clinically important. For example, the paired *t*-test (Table 1) showed a statistically relevant difference between examiners for the median sensory onset latency, yet the means differed by only 0.1 ms, with little clinical significance.

Effect of Reliability on Diagnostic Criteria. A key question was how examiner differences affected assessment of electrodiagnostic abnormality (i.e., Would examiner 1 rate a subject "normal" where Examiner 2 would rate the subject "abnormal"?). To explore this, we used two cut points for the median-ulnar peak latency difference: 0.5 ms and 0.8 ms. The conventional normal cut point for the median-ulnar peak latency difference is 0.4 to 0.5 ms, to avoid false-positive findings.[27] However, strong empirical evidence has shown that the normative value among workers is 0.8 ms, defined by the upper 95th percentile for the median-ulnar peak latency difference in the dominant hand.[29]

Results from this study showed that classification into normal/abnormal categories amplified inconsistencies in ratings around the lower cut point of 0.5 ms for the median-ulnar peak latency difference. The higher cut point of 0.8 ms was clearly able to distinguish those with very abnormal findings. Also, there was good to excellent interexaminer agreement using the median-ulnar peak latency difference with the normal criterion of 0.8 ms ($\kappa = 0.75$ on the dominant side), while interexaminer reliability was slightly lower using the 0.5-ms cut point ($\kappa = 0.71$). Figure 1 provides a graphical illustration of these data in relation to the two cut points. More marked differences were seen with intraexaminer reliability. Kappa was 0.90 at the higher cut point of 0.8 ms and 0.52 at the 0.5-ms cut point for examiner 1.

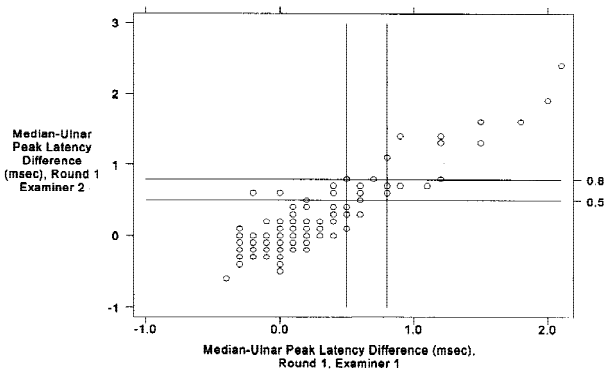

FIGURE 1. Interexaminer reliability of median-ulnar peak latency difference for SNAP in the dominant hand.

For examiner 2, kappa was 0.85 at the higher cut point and 0.68 at the 0.5-ms cut point. Presumably, the 0.8-ms cut point reflected more advanced nerve dysfunction. Subjects with very abnormal responses were well above the 0.8-ms cut point.

Given the instability of the normal ratings and the higher reliability of the median-ulnar peak latency difference, a classification with more gradations (e.g., normal, equivocal, mild, or definite abnormality) or the use of the original continuous measures is indicated. This may better demonstrate the magnitude of abnormality and provide a more rational interpretation of results. Of course, appropriate normative values[29] are critical for a reasonable definition of normal within a worker population. Also, 14-cm antidromic SNAPs are often only a first step towards further electrodiagnostic studies.

Practical Considerations for Diagnosis of CTS. Nerve conduction studies have been considered the gold standard for CTS,[18] but abnormal NCS alone, without symptoms, do not define CTS.[28] Symptoms and physical findings are critical for accurate test interpretation.[21] Clinically, electrodiagnostic studies of suspected CTS that involved normal or equivocal findings for 14-cm antidromic median studies should be evaluated further with short-segment orthodromic studies.

Nerve conduction studies have false-positive rates of perhaps 23% among active workers with the most expanded symptom definition (i.e., numbness, tingling, burning, or pain in the forearm, wrist, or hands).[16] Grundberg[19] reported false-negative rates of perhaps 10%. Barnhart et al.[2] reported that NCS improved specificity of CTS classification, as opposed to case definitions using only physical examination and symptom data among workers.

Still, the significance of abnormal NCS in asymptomatic workers remains unclear. Abnormal NCS are

not predictive of CTS in asymptomatic workers, and there is no justification for asymptomatic workers to receive differential treatment (i.e., work restrictions or special job placement).[32]

Strengths and Limitations. This study benefited from highly trained and seasoned examiners including experienced clinicians, one of whom trained the other. Reliability may be worse among clinicians who do not receive standardized training. In this context, the results are best estimates of what would be found between two experienced examiners with extended training in a common methodology. As noted, strict adherence to protocols improves the reliability of results.[3]

Paradoxically, the strength of this study also limits its generalizability, as it is subject to potential bias due to the number of examiners ($n = 2$), testing two nerves (median and ulnar) and only 14-cm segments (no short-segment orthodromic SNAPs). With the high proportion of women in the study, the results may not apply to men, although we are unaware of any data to suggest differential reliability of electrodiagnostic measurements by gender. Also, given the research design in the work setting, practical considerations restricted the extent of testing. In addition, a measurement issue arises, statistically, with the small proportion of abnormal studies (between 10% and 20%). This report from a field study of active workers represents a worst-case scenario. One would expect better reliability in a clinical setting among subjects with a larger spectrum, or range, of findings, and more severe morbidity.

The data for these analyses were collected in studies supported by the Johns Hopkins University Center for VDT and Health Research and other sources. The authors gratefully acknowledge the workers and management who participated in the study and express appreciation to Kim Augenstein, MD, Lisa Carchidi, Mike Gerard, Mark Gordon, MD, Randy Rabourn, and Teresa Spiegelberg for their assistance in the medical field studies.

REFERENCES

1. American Association of Electrodiagnostic Medicine. Practice parameter for electrodiagnostic studies in carpal tunnel syndrome: summary statement. Arch Phys Med Rehabil 1994;75: 124–125.
2. Barnhart S, Demers PA, Miller M, Longstreth WT, Rosenstock L. Carpal tunnel syndrome among ski manufacturing workers. Scand J Work Environ Health 1991;17:46–52.
3. Bril V, Ellison R, Ngo M, Bergstrom B, Raynard D, Gin H, the Roche Neuropathy Study Group. Electrophysiological monitoring in clinical trials. Muscle Nerve 1998;21:1368–1373.
4. Britt R. Hands and wrists are thrust into the hiring process. New York Times; Sep 21, 1997: Section 3, p 11.
5. Cailliet R. Hand pain and impairment, 4th edition. Philadelphia: FA Davis; 1994. 311 p.
6. Chang C, Lien I. Comparison of sensory nerve conduction in the palmar cutaneous branch and first digital branch of the median nerve: a new diagnostic method for carpal tunnel syndrome. Muscle Nerve 1991;14:1173–1176.
7. Chaudhry V, Cornblath D, Mellits E, Avila O, Freimer M, Glass J, Rein J, Ronnett G, Quaskey S, Kuncl R. Inter- and intra-examiner reliability of nerve conduction measurements in normal subjects. Ann Neurol 1991;30:841–843.
8. Chaudhry V, Corse A, Freimer M, Glass J, Mellits E, Kuncl R, Quaskey S, Cornblath D. Inter- and intraexaminer reliability of nerve conduction measurements in patients with diabetic neuropathy. Neurology 1994;44:1459–1462.
9. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
10. Dawson G, Scott J. The recording of nerve action potentials through skin in man. J Neurol Neurosurg 1949;12:259–267.
11. De Jesus PV, Hausmanova-Petrusewicz I, Barchi RL. The effect of cold on nerve conduction of slow and fast nerve fibers. Neurology 1973;23:1182–1189.
12. Denys EH. AAEM Minimonograph #14. The influence of temperature in clinical neurophysiology. Muscle Nerve 1991;14: 795–811.
13. Dioszeghy P, Stalberg E. Changes in motor and sensory nerve conduction parameters with temperature in normal and diseased nerve. Electroencephalogr Clin Neurophysiol 1992;85: 229–235.
14. Dyck PJ, Kratz KM, Lehman KA, Karnes JL, Melton III LJ, O'Brien PC, Litchy WJ, Windebank AJ, Smith BE, Low PA, Service FJ, Rizza RA, Zimmerman BR. The Rochester Diabetic Neuropathy Study: design, criteria for types of neuropathy, selection bias, and reproducibility of neuropathic tests. Neurology 1991;41:7999–8807.
15. Fleiss JL. Statistical methods for rates and proportions, 2nd edition. John Wiley & Sons; 1981. 321 p.
16. Franzblau A, Werner R, Valle J, Johnston E. Workplace surveillance for carpal tunnel syndrome: a comparison of methods. J Occ Rehab 1993;3:1–14.
17. Franzblau A, Werner RA, Albers JW, Grant CL, Olinski D, Johnston E. Workplace surveillance for carpal tunnel syndrome using hand diagrams. J Occ Rehab 1994;4:185–198.
18. Gerr F, Letz R, Landrigan PJ. Upper-extremity musculoskeletal disorders of occupational origin. Ann Rev Publ Health 1991;12:543–566.
19. Grundberg AB. Carpal tunnel decompression in spite of normal electromyography. J Hand Surg 1983;8:348–349.
20. Johnsen B, Fuglsang-Frederiksen A, Vingtoft S, Fawcett P, Liguori R, Nix W, Otte G, Schofield I, Veloso M, Vila A. Inter- and intraobserver variation in the interpretation of electromyographic tests. Electroencephalogr Clin Neurophysiol 1995;97:432–443.
21. Kimura J. Electrodiagnosis in diseases of nerve and muscle: principles and practice, 2nd edition. Philadelphia: FA Davis; 1989. 709 p.
22. Kimura J. Facts, fallacies, and fancies of nerve conduction studies. Muscle Nerve 1997;20:777–787.
23. Kimura J. Principles and pitfalls of nerve conduction studies. Ann Neurol 1984;16:415–429.
24. Kramer MS, Feinstein AR. Clinical biostatistics LIV. The biostatistics of concordance. Clin Pharmacol Ther 1981; 29: 111–123.
25. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Stat Med 1994;13:2465–2476.
26. Olney RK. Neurophysiologic evaluation and clinical trials for neuromuscular diseases. Muscle Nerve 1998;21:1365–1367.

27. Redmond MD, Rivner MH. False positive electrodiagnostic tests in carpal tunnel syndrome. Muscle Nerve 1988;11: 511–518.
28. Rempel D, Evanoff B, Amadio PC, de Krom M, Franklin G, Franzblau A, Gray R, Gerr F, Hagberg M, Hales T, Katz JN, Pransky G. Consensus criteria for the classificatin of carpal tunnel syndrome in epidemiologic studies. Am J Pub Health 1998;10:1447–1451.
29. Salerno DF, Franzblau A, Armstrong TJ, Werner RA, Albers JW, Bromberg MB. Nerve conduction studies among workers: normative values. Muscle Nerve 1998;21:999–1005.
30. Stata Statistical Software: Release 5.0 College Station, TX: Stata Corporation; 1997.
31. Valensi P, Attali J-R, Gagant S, the French Group for Research and Study of Diabetic Neuropathy. Reproducibility of parameters for assessment of diabetic neuropathy. Diab Med 1993; 10:933–939.
32. Werner RA, Franzblau A, Albers JW, Buchele H, Armstrong TJ. Use of screening nerve conduction studies for predicting future carpal tunnel syndrome. Occup Environ Med 1997;54: 96–100.