INTERNATIONAL JOURNAL OF LEPROSY

Volume 64, Number 3
Printed in the U.S.A.
(ISSN 0148-916X)

Intra- and Inter-Tester Reliability of Sensibility Testing in Leprosy[1]

Wim H. van Brakel, Ishwar B. Khawas, Khadga Singh Gurung, C. Marleen Kets, Monique E. van Leerdam, and William Drever[2]

Regular assessment of peripheral nerve function is essential to prevent or to minimize impairment, disability and deformity in leprosy ([2, 11, 31, 34, 38]). Following a "neuritis workshop" at Karigiri, India, in 1980, seven different tests of sensory and motor function were recommended as "mandatory" ([32]). These included tests such as nerve trunk palpation and a stretch and compression test to elicit tenderness or nerve pain, direct evaluation of axonal function through nerve conduction velocity measurement, as well as assessment of end organ function, like touch, (moving) 2-point discrimination, pain sensation and muscle power. In the 13 years or so since this workshop, only voluntary muscle testing and sensory testing with graded nylon monofilaments or ballpoint pen have gradually been introduced in the clinical management of leprosy patients [Bell-Krotoski, J.A. Hand screen for early detection and monitoring of peripheral neuropathy, part II. The Star **51** (1992) 3–7 and [9, 18, 24, 26, 31, 38]].

The concept that any test or measurement should itself be tested for reliability and validity before use in clinical medicine or research is relatively new ([19, 29, 35]). It is, therefore, not surprising that very few studies have looked at this issue in relation to clinical testing in leprosy. Almost all of the available data on reliability of sensibility testing with Semmes-Weinstein monofilaments (SWM) and moving 2-point discrimination (M2PD) refer to testing of non-leprosy patients.

The reliability or consistency of a test or measure is the subject of this paper. Validity—whether a test measures what it is intending to measure—is related to reliability in that a test can only be valid if it is also reliable. Validity will not be discussed here. Recent studies at our hospital of concurrent and criterion validity of the tests under consideration have been published elsewhere ([36]).

We conducted two studies to determine the intra- and inter-tester reliability of sensibility testing with SWM, M2PD, and pin prick. The results of these studies, which were carried out at Green Pastures Hospital in Pokhara, Nepal, are presented in this paper.

MATERIALS AND METHODS

Reliability. The "reliability of a test" refers to whether the instrument is measuring something in a reproducible and consistent fashion ([35]). Based on this, we defined reliability of sensory testing instruments as "the ability of a test to measure sensibility in a reproducible and consistent fashion." As a concept, reliability " . . . may be thought of as the ratio of 'signal' to 'noise' in a measure" ([21]).

The first component of reliability assessment is test-retest reliability, also called repeatability. This is defined as " . . . the degree to which a measure is consistent or reproduced when readministered by trained staff in maximally similar circumstances" ([21]). The second component is internal consistency, which refers to the homogeneity of the different items on a measurement scale ([29, 35]). This type of reliability is only relevant when composite scores are used, as in questionnaire surveys or psychological assessments. The third component is inter-

[1] Received for publication on 23 May 1995; accepted for publication in revised form on 6 March 1996.

[2] W. H. van Brakel, Ph.D.; I. B. Khawas, B.A.; K. S. Gurung, B.A., Green Pastures Hospital, P. O. Box 28, Pokara, Nepal. C. M. Kets; M. E. van Leerdam, Haarlemmermeerstraat 155[1], 1058 JZ Amsterdam, The Netherlands. W. Drever, M.B.B.S., c/o Leeds Medical School, Leeds University, U.K.

Reprint requests to Dr. van Brakel, c/o INF, P. O. Box 5, Polhara, Nepal. FAX = 977-61-20430; email = Brakel@npl.healthnet.org

rater reliability, which is the degree of agreement of measurement between two or more raters when they rate the same subject or phenomenon ([21]). To evaluate reliability of a measurement both the repeatability and the inter-rater agreement must be assessed.

The choice of the terms "rater," "observer" or "tester" depends largely on the circumstances, i.e., the test under review. Because "tester" is the most appropriate term in our situation, this is the term used hereafter. In some situations it may be important to examine how well a given pair or team of testers can reproduce test results between them. This may reveal random or nondifferential variability, or differential variability (tester bias), if one tester rates consistently higher (or lower) than the other(s). We were more interested in evaluating the reliability of the tests themselves and, therefore, we chose a design with many different tester pairs, each only testing 1–4 patients ([23, 37]).

Selection of patients. No randomization was used since the objective was test comparison within the same patient. Only patients with a stable nerve function—no changes in motor or sensory scores during the previous 6 months, or acute leprosy reactions or neuritis—were asked to take part in the study. Most patients were in- and outpatients of Green Pastures Hospital (GPH) in Pokhara. For the inter-tester reliability study, a number of similar patients from The Leprosy Mission Anandaban Hospital in Kathmandu, Nepal, were included. All patients had an established diagnosis of leprosy. A few patients had only one hand or foot tested if missing digits or severe deformities made testing at the prespecified sites on the other hand/foot impossible.

Testers and study design. The intra-tester reliability study involved two trained physiotechnicians with long experience in nerve function assessment at GPH, each testing 15 patients (test A). Most patients were re-tested the next day and some patients after 2 days (test B). At the second examination the testers were blinded for the result of the previous test.

Forty-one different pairs of testers performed the tests for the inter-tester reliability study. Included were 5 trained physiotechnicians, 1 expatriate occupational therapist, 4 nurses/paramedical workers (PMW) who had been familiarized with the testing techniques, and 2 expatriate medical students who were also well acquainted with the testing methods. Twenty-one pairs examined only 1 patient between them, 15 pairs examined 2, 4 pairs examined 3, and 1 pair examined 4 patients. Where possible, the two testers examined the patient on the same day. Sometimes the second examination was only possible the next day. The second tester was unaware of the results of the first examination.

SWM. Patients were tested using the standard set of 5 "colored Semmes-Weinstein monofilaments" as described by Bell-Krotoski ([4]). This test evaluates touch sensibility thresholds. The score per site varies from 0–5. A score of 5 was given when the thinnest monofilament in the test series was felt; a score of 0, if even the thickest filament was not felt. These filaments (Semmes-Weinstein log numbers 2.83, 3.61, 4.31, 4.56 and 6.65) are equal to approximate application forces of 70 mg, 200 mg, 2 g, 4 g and 300 g when applied in such a way that the filament bends slightly ([6]). The sites tested were: for the median nerve, the volar surface of the distal phalanx of the thumb and index finger and the palmar skin over the second metacarpophalangeal joint; for the ulnar nerve, the volar surface of the distal phalanx of the little finger, the palmar skin over the fifth metacarpophalangeal joint and on the proximal end of the hypothenar eminence; on the sole of the foot (posterior tibial nerve), the big toe, the plantar skin over the first and fifth metatarsal heads, and the heel.

M2PD. Moving touch sensibility of the median and ulnar nerves was tested with the M2PD test as described by Dellon ([15]). This test, which evaluates density of (quickly adapting) touch receptors, can be done with a simple paper clip. For this study we used a Disk-Criminnator™ (available through P. O. Box 13692, Baltimore, Maryland 21210, U.S.A.), a plastic disk on which pairs of metal prongs are mounted with different inter-prong distances. The test was first explained to the patient. Randomly, one or two prongs were moved from proximal to distal over the test site, giving as little pressure as possible, and the patient was asked whether he felt one or two prongs. The smallest inter-prong distance for which

TABLE 1. *Inter-tester agreement between 125 paired SWM tests on the little finger of leprosy patients in Nepal, using a 6-point scale.*[a]

	tester B						
tester A	0	1	2	3	4	5	Total
0	10	6					16
1	5	12	2	1			20
2	1	4	1	1			7
3	1	2	2	3	3	3	14
4				2	3	3	8
5			2	4	8	46	60
Total	17	24	7	11	14	52	125

[a] $\kappa = 0.45$ (0.34–0.57), $\kappa_w = 0.89$ (0.85–0.93).

TABLE 2. *Interpretation of values of* κ *and* κ_ω.[a]

Value of κ or κ_w	Strength of agreement
≤ 0.40	poor
0.41-0.60	moderate
0.61-0.80	good
0.81-1.00	very good

[a] Modified from Altman([1]).

the patient gave two correct responses out of three trials was recorded in millimeters for that site. The smallest distance tested was 2 mm. If the testing device was not felt at all, the score was recorded as 0. If the device was felt, but only as one moving point, a score of 1 was given. The recorded scores were transformed to a scale of 0–13 by subtracting the scores (except 0 and 1) from 15. Thus, 0 was the worst and 13 the best result possible. Sites were the same as for the SWM, but only the thumb, index and little finger, big toe and heel were tested.

Pin prick. Pain sensation was tested using commercially available, standard-type wooden toothpicks. The toothpick was applied randomly with the sharp end or the blunt end, and the patient was asked to indicate whether he felt "sharp" or "blunt." The sites tested were the same as for the M2PD test plus the plantar skin over the first metatarsal head. The score per site was the number of correct responses out of five trials. This test was only added at the time of the inter-tester study, so no intra-tester reliability results were available.

Statistical methods. Agreement between two consecutive tests or between two testers was measured with weighted kappa (κ_w) statistics ([12, 13]). A typical agreement table is shown in Table 1. Quadratic disagreement weights were used, i.e., weights for off-diagonal cells were the square of the deviation of the pair of observations from exact agreement ([13, 27]). Thus, a difference of 1 would be weighted as 1, a difference of 2 as 4, of 3 as 9, and so on. Perfect agreement was assigned a weight of "0." Reference values for κ_w were adapted from Altman ([1]) and are given in Table 2. Paired differences between tests and testers were evaluated using the Wilcoxon matched-pairs signed-rank test. The magnitude of the differences between tests or testers relative to the maximum score of each test was calculated using a new indicator, the mean percent difference (mpd).

$$\text{Mean percent difference} = \frac{\Sigma|\text{score A} - \text{score B}|}{\text{n (maximum score)}} \times 100$$

This gives an easy-to-interpret figure representing the average variability between tests or testers as a percentage of the maximum score of the test, thus making direct comparison between different tests possible. For example, an mpd of 6.3% for inter-tester comparison of SWM testing on the thumb shows that, on average, the test scores differed by only 6.3% of the maximum score of 5. Two independent mpds can be compared using standard formulas for standard error and difference between proportions (SND test). The mpd should not be used to test a null hypothesis of no

TABLE 3. *Intra- and inter-tester reliability of sensibility testing of the hand with Semmes-Weinstein monofilaments in leprosy patients in Nepal.*

Site	type of test	number of pairs	Weighted kappa (95%CI)[a]	mpd[b] (%)	Wilcoxon p value[c]
thumb	intra	59	0.83(0.74-0.93)	6.4	0.35
	inter	127	0.87(0.82-0.92)	6.3	0.44
index finger	intra	59	0.87(0.79-0.96)	4.4	0.25
	inter	125	0.87(0.80-0.94)	7.0	0.43
median (3 sites)[d]	intra	59	0.84(0.58-1.0)	5.2	0.31
	inter	125	0.90(0.71-1.0)	6.6	0.043
little finger	intra	59	0.92(0.85-0.99)	6.8	0.42
	inter	125	0.89(0.85-0.93)	11	0.064
hypothenar	intra	59	0.92(0.88-0.95)	7.1	0.19
	inter	128	0.84(0.79-0.89)	11	0.049
ulnar (3 sites)[e]	intra	59	0.92(0.67-1.0)	7.1	0.49
	inter	125	0.91(0.73-1.0)	9.6	0.011
big toe	intra	59	0.92(0.88-0.96)	7.1	0.21
	inter	125	0.79(0.70-0.87)	14	0.63
mtp1[f]	intra	59	0.89(0.83-0.96)	8.1	0.6
	inter	130	0.76(0.67-0.86)	14	0.66
heel	intra	59	0.83(0.71-0.94)	8.1	0.52
	inter	130	0.79(0.71-0.87)	12	0.55
posterior tibial	intra	59	0.88(0.63-1.0)	7.3	0.67
(3 sites)[g]	inter	125	0.83(0.66-1.0)	12	0.23

[a] 95% confidence interval.
[b] Mean percent difference.
[c] Wilcoxon's matched-pairs signed-rank test.
[d] Thumb + 2nd metacarpal phalangeal joint + index finger.
[e] Little finger + 5th metacarpal phalangeal joint + hypothenar eminence.
[f] First metatarsal phalangeal joint.
[g] Big toe + 1st metatarsal phalangeal joint + heel.

difference since negative differences have been eliminated by taking the absolute value of the difference.

A p value of $< 5\%$ was used as the level of statistical significance. Of the kappa and weighted kappa statistics, the 95% confidence interval (CI) is given. Analysis was done using Epi Info software, version 5.01 ([14]) and SPSS for Windows, version 6. Matrixes for calculating weighted kappa were constructed in a spreadsheet (Quattro Pro for Windows, version 1).

TABLE 4. *Intra- and inter-tester reliability of sensibility testing of the hand with moving 2-point discrimination in leprosy patients in Nepal.*

Site	type of test	number of pairs	Weighted kappa (95%CI)[a]	mpd[b] (%)	Wilcoxon's p value[c]
thumb	intra	59	0.75(0.52-0.98)	7.7	0.83
	inter	123	0.70(0.52-0.88)	10	0.26
index finger	intra	59	0.80(0.56-1)	5.6	0.13
	inter	121	0.73(0.56-0.90)	9.0	0.89
little finger	intra	59	0.82(0.68-0.96)	10	0.39
	inter	122	0.82(0.73-0.91)	13	0.51
big toe	intra	59	0.82(0.66-0.97)	11	0.044
	inter	122	0.74(0.63-0.84)	17	0.48
heel	inter	126	0.54(0.40-0.67)	23	0.88

[a] 95% confidence interval.
[b] Mean percent difference.
[c] Wilcoxon's matched-pairs signed-rank test.

RESULTS

Patients. Thirty patients were selected for the intra-tester and 67 for the inter-tester agreement studies. They represented different immunological classification subgroups. We assumed an equal and independent chance of (dis-)agreement for the left and right hand and foot of each patient. They were, therefore, pooled, giving one sample of 60 hands and feet for the intra-tester study and 134 hands and feet for the inter-tester study. For various reasons not every site could be tested on all patients; this is the reason for the slight variation in the number of pairs between different sites (e.g., Table 3, column 3).

Intra-tester agreement. Table 3 shows the statistics for the SWM test. Weighted kappas (κ_w) ranged from 0.83 (thumb) to 0.92 (little finger, hypothenar and big toe). There were no apparent differences between κ_w values for the hands and feet. None of the differences between test A and test B were statistically significant according to Wilcoxon's matched-pairs signed-rank test. Mean percent differences (mpd) did not exceed 8.1% (minimum 4.4%).

The M2PD results are given in Table 4. Weighted kappas ranged from 0.75 (thumb) to 0.82 (little finger and big toe). Only the difference in scores of the big toe was significant at the 5% level ($p = 0.044$, Wilcoxon). The maximum mpd was 11%. Although all SWM weighted kappas were higher than those of the M2PD, these differences were not statistically significant.

Inter-tester agreement. The SWM results (Table 3) showed a minimum weighted kappa of 0.76 (first metatarsal head) and a maximum of 0.89 (little finger). The mpd ranged from 6.3%–14%. The paired differences between tester A and tester B were close to or < 5% only for the median nerve combined ($p = 0.043$), little finger ($p = 0.064$), hypothenar ($p = 0.049$) and ulnar combined ($p = 0.011$). Inter-tester agreement for the M2PD (Table 4) showed weighted kappa values ranging from 0.54 (heel) to 0.82 (little finger). None of the inter-tester differences were significant. Values of κ_w tended to be higher for the hand than for the foot, but this difference was only significant for the two highest values (little finger vs big toe, $z = 2.07$, $p = 0.019$).

TABLE 5. *Intra- and inter-tester reliability of sensibility testing of the hand with a pin prick in leprosy patients in Nepal.*

Site	number of pairs	Weighted kappa (95%CI)[a]	mpd[b] (%)	Wilcoxon's p-value[c]
thumb	123	0.57(0.39-0.74)	15	0.018
index finger	121	0.66(0.52-0.81)	12	0.56
little finger	122	0.85(0.78-0.91)	11	0.28
big toe	122	0.64(0.53-0.76)	20	0.039
mtp1[d]	126	0.51(0.37-0.66)	23	0.19
heel	126	0.45(0.30-0.60)	24	0.51

[a] 95% confidence interval.
[b] Mean percent difference.
[c] Wilcoxon's matched-pairs signed-rank test.
[d] First metatarsal phalangeal joint.

Weighted kappas were significantly different for the thumb and the heel ($z = 1.76$, $p = 0.039$ and $z = 3.15$, $p = 0.0008$, respectively).

Table 5 shows the inter-tester agreement for the pin prick test. Values of weighted kappa ranged from 0.45 (heel) to 0.85 (little finger). Inter-tester differences were significant for the thumb and the big toe (p values 0.018 and 0.039, respectively). All weighted kappas, except of the little finger, were significantly lower than those of the SWM test ($p = 0.022$–0.00003, z test).

Intra- versus inter-tester agreement. The monofilament weighted kappas were significantly higher for intra- than for inter-tester agreement for the hypothenar, big toe and first metatarsal phalangeal joint; M2PD weighted kappas only for thumb and heel.

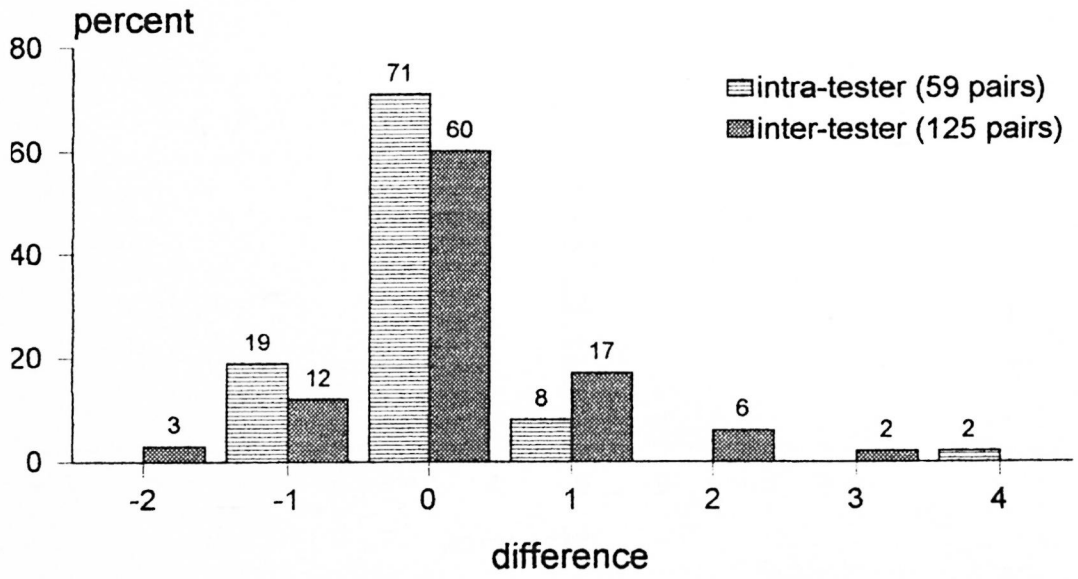

FIG. 1. Distribution of within- and between-tester differences in SWM test scores (test site = little finger).

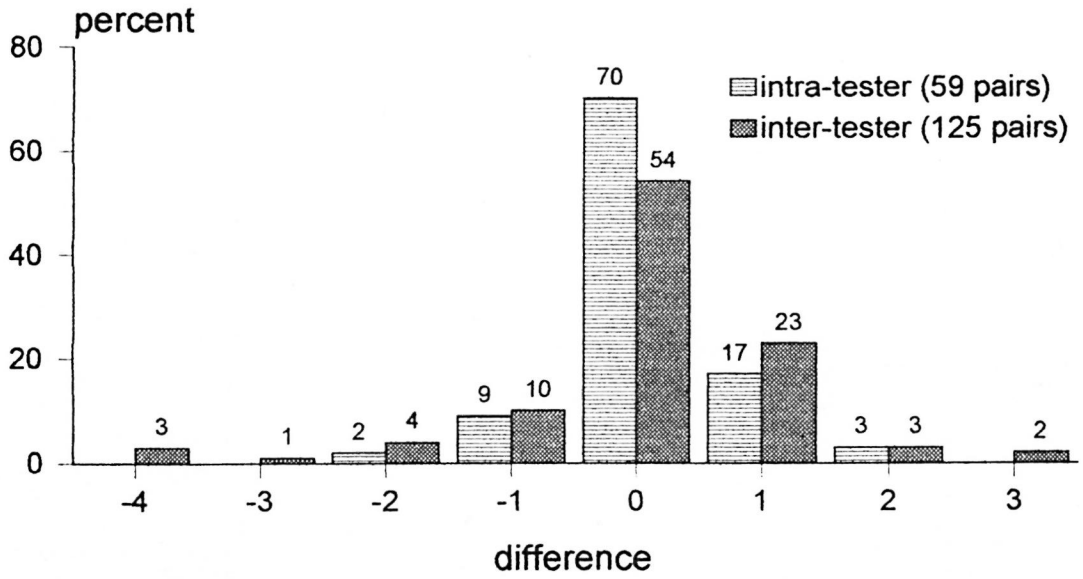

FIG. 2. Distribution of within- and between-tester differences in SWM test scores (test site = big toe).

The distribution of between-test and between-tester differences of the SWM is illustrated in Figure 1 for the little finger and in Figure 2 for the big toe. Figure 3 illustrates the same for the M2PD on the little finger.

Tester subgroups. We analyzed subgroups according to skill and experience (data not shown). Although the number of paired observations in each subgroup was small, there was a striking and significant difference between the physiotechnician-only group (N = 28), the various mixed subgroups and the PMW-only group (N = 13). The weighted kappa values in the physio group were significantly higher than in the whole group (e.g., SWM little finger 0.98 vs 0.89, $p < 0.0001$; M2PD little finger 0.99 vs 0.82, $p = 0.00027$; pin prick little finger 0.97 vs 0.85, $p = 0.027$).

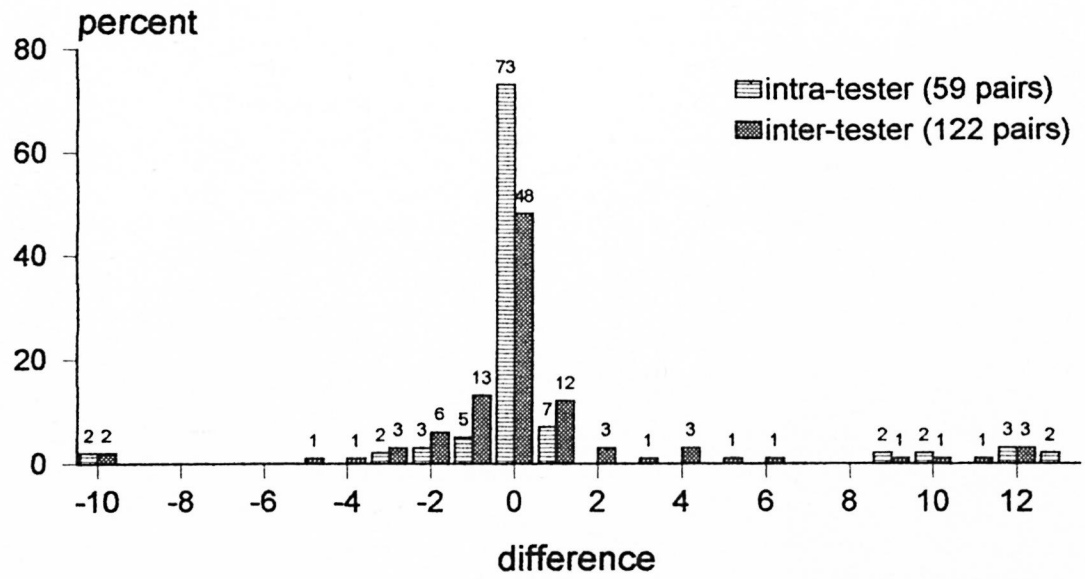

FIG. 3. Distribution of within- and between-tester differences in M2PD test scores (test site = little finger).

DISCUSSION

The objective of reliability testing is twofold. First to evaluate how reliable or consistent a given test or measurement is under the present circumstances and, second, to identify sources of measurement variability that may be amenable to improvement.

Intra-tester reliability depends on within-test variability (e.g., random measurement error, calibration in different temperatures, relative humidities, variation in application force, etc.), within subject (patient) variability (cooperation, concentration, etc.); variability in what is being measured (e.g., changes in blood pressure); within-tester variability (accuracy, skill, experience) and other factors that may vary in the environment between the first and the second test. Good intra-rater reliability is perhaps the most basic quality that a test should have.

Inter-tester reliability depends on the variability between raters, such as differences in motivation, level of training and testing technique, in addition to the sources of intra-tester variability (error). Considering these sources of variability, it follows that reliability is context specific. For instance, if reliability of a test of visual acuity has been established in North America, the reliability of the same test will need to be re-evaluated if it is to be used in a different context, say, India.

SWM. Intra-tester agreement measured with weighted kappa was "very good" by the (arbitrary) standards in Table 2. Inter-tester agreement was very good for the hand and good for the foot. It is interesting to note that agreement of testing on the sole of the foot was as good as on the palm of the hand. The thicker skin with callus on the sole has a higher touch perception threshold, but this apparently does not affect the reliability of testing. Normal for our patients is 2 g on the foot and 200 mg on the hand ([22]). A mean percent difference (mpd) range of 4.4%–8.1% for intra- and 6.3%–14% for inter-tester comparison seems acceptable. An average difference of 0.5 filament would give an mpd of 10%.

Birke, *et al.* ([8]) reported an intra- and inter-tester intra-class correlation coefficients (ICC) of 0.88–0.93 for all sites tested with SWM, corresponding quite closely to our range of κ_w (0.76–0.92). Bell and Tomancik studied application of the five standard SWM to four small strain gauges ([6]). They do not report any reliability coefficients, but a measure of variability, the coefficient of variation (CV; standard deviation divided by the mean). In their case this was an appropriate indicator of variability since they measured directly in milligrams on a continuous scale. The CV ranged from 6%–9% for intra- (1 tester, 105 applications/filament) and 5%–8% for inter-tester (5 testers, 450 applications/filament) comparison. They also measured application force variability of ballpoint and pin prick testing; CVs were 11% and 14%, respectively.

Diamond, *et al.* tested the feet of 31 diabetic subjects with three different SWM ([17]). There were two testers for the inter-tester comparison. Kappa values for intra- and inter-tester reliability on this 4-point scale ranged from 0.72 to 0.83.

M2PD. Values of weighted kappa were all good to very good, except for inter-tester agreement on the heel, which was only moderate. This was not surprising to us since we had found much variability in normal M2PD values on the heel in a different study (to be published elsewhere). The mpds were acceptable (5.6%–17%) except for the heel (23%). A mean between-test(er) variability of 1 mm would give an mpd of 7.7%, while 2 mm would give 15%. Most values were between these two.

Dellon, *et al.* reported inter-tester agreement between two testers examining 30 nerve-injured patients ([16]). Pearson's r was 0.92 for both index and little fingers. Paired differences were not examined. Assuming that a weighted kappa would have been of similar magnitude, their reliability was considerably higher than ours. This could be attributed to the fact that the one tester-pair were highly trained (a hand surgeon and an occupational therapist) and experienced in using the M2PD test. Our 41 tester pairs were less well trained and (mostly) far less experienced in using M2PD.

Pin prick. Except for the little finger, all weighted kappas were significantly lower than for the SWM. Mpd values were almost twice as high on the foot as on the hand. This may be due to the fact that the thicker skin and callus on the sole makes distinguishing sharp from blunt more diffi-

cult. There was an indication that the test was performed more reliably by physiotechnicians than by less-trained staff. However, reliability was clearly inferior to the SWM and the M2PD, making this test less suitable for serial comparisons.

Sources of variability and ways to improve agreement. Our intra-tester reliability was probably as good as can be reasonably expected under the circumstances. Both testers were experienced and well-motivated physiotechnicians. It is likely that reliability would be somewhat less under field conditions where staff motivation tends to be less good and time pressure more acute. In addition, it is often impossible to find the "quiet room" that is said to be "mandatory" for accurate assessment ([3]). A total lack of privacy is common. Therefore, distraction and lack of concentration will likely be sources of variability in the measurement of sensibility thresholds under field circumstances. Green Pastures Hospital, Anandaban Hospital, or the McKean Rehabilitation Centre, Chiangmai, Thailand, are not "a field setting" ([8]), as is often believed. It will, therefore, be necessary to re-assess reliability of sensibility testing under real field conditions.

Another source of potential variability is the state of hydration of the skin areas to be tested. McAuley, *et al.* demonstrated a small but significant decrease in the average monofilament sensibility threshold after soaking of the hands for 30 min ([28]). They conclude that attention should be given to testing the patient under similar hydration circumstances if comparison between measures is to be made.

The findings of Bell and Tomancik referred to above indicate that the amount of measurement error due to the SWM instruments themselves is small ([6]). But calibration and standardization in terms of diameter and length of the filaments is important for comparability of results ([3]). Incorrect application technique may also result in spurious results ([25]). A further refinement of the instrument, the Weinstein Enhanced Sensory Test, was recently introduced by Weinstein himself ([39]). Bell and Buford showed that in M2PD testing with a handheld instrument variability in application force was greater than in testing with the SWM ([5]). Our indicator of variability (the mean percent difference) was slightly higher for the M2PD than for the SWM, but these differences were only statistically significant for the heel. It seems, therefore, that in a clinical setting M2PD has an acceptable instrument error.

It was encouraging to see the standard of inter-tester reliability that was achieved by our very mixed group of testers, which included some experienced but also several very inexperienced examiners. We attempted an analysis of subgroups according to skill and experience to see what influence these factors have on the observed variability. The weighted kappa values in the physio-only group were significantly higher than in the whole group, and differences between the physio group and the PMW-only group were even greater. In the physio-only group there were no significant differences between intra- and inter-tester reliability. This indicates that skill and experience were the main source of inter-tester variability in our data and that, therefore, training and practice play a vital role in the reliability of these sensory tests. Whenever leprosy PMW or multipurpose health workers are trained in nerve function assessment as much practice as possible should be given.

Reliability statistics. Reliability is often expressed in Pearson correlation coefficients (r) ([16, 30]). Several authors have argued that this is not an appropriate statistic ([10, 30, 33, 35]). The main reason is that the Pearson r measures the association and co-variance between two variables, rather than agreement ([10, 30]). Ottenbacher and Tomchek ([30]) and Sheikh ([33]) have convincingly shown that high correlation coefficients may, in fact, hide substantial or even total disagreement. A second reason why Pearson's r is inappropriate is that our data scales are ordinal rather than interval or even continuous ([1]). Pearson's r is appropriate as reliability coefficient for continuous data, provided that paired between-test(er) differences are examined for departure from zero ([21, 27]).

Weighted kappa is recommended as the measure of scaled agreement between categorical scales of more than two categories ([1, 13, 20, 27]). Quadratic disagreement weights are recommended to ensure comparability of weighted kappa values between studies ([27]). The highest direct agreement is ob-

tained if scales are collapsed to binary variables. In practical terms, this means that classifying sensory impairment as either "absent" or "present" is the most reliable way of testing. For screening purposes this could be done with just one monofilament, provided the appropriate cut-off (normal) values are known. However, the strength of SWM and M2PD is that they provide a graded measurement of sensibility with which (gradual) deterioration or improvement can be monitored.

"The interpretation of the magnitude of weighted kappa is like that of unweighted kappa . . . ," according to Fleiss ([20]). When using the (arbitrary) criteria given by Altman ([1]) a weighted kappa between 0.21 and 0.40 would still be called fair. We prefer stricter criteria and, therefore, classified a value of ±0.40 as "poor agreement" after Fleiss (Table 2).

Intra-class correlation coefficients (ICC) have been recommended as reliability coefficients ([7, 38]). The ICC is computed from analysis of variance components and is widely used in reliability statistics ([30]). But it has been observed that when quadratic disagreement weights are used, the ICC is identical to the κ_w ([27, 35]). Our weighted kappa results can, therefore, be interpreted as if they were ICC results. Cohen also noted that weighted kappa with quadratic disagreement weights is numerically very close to Pearson's r, provided that the marginal distributions of the agreement table are not very different ([13]).

The reliability results found in our and the above-mentioned centers are encouraging, and indicate that perhaps even under real field conditions acceptable standards of reliability can be achieved.

CONCLUSIONS

1. Reliability of the SWM test was very good, closely followed by the M2PD test.

2. There was evidence that the main source of variability between testers was testing skill and experience. Among the experienced physiotechnicians there was no significant difference between intra- and inter-tester reliability.

3. Reliability of the pin prick test was clearly less good than that of the SWM and M2PD.

4. The mean percent difference is an easy-to-interpret statistic expressing between test(er) variability as a percentage of the maximum test score.

5. The reliability of these instruments should be re-assessed under field conditions.

SUMMARY

We conducted an intra- and inter-tester agreement study of three sensory screening tests used in nerve function assessment of leprosy patients: the Semmes-Weinstein monofilament (SWM) test, moving 2-point discrimination (M2PD), and the pin prick test. The weighted kappa (κ_w) statistic was used as the reliability coefficient. The SWM had intra-observer κ_ws ranging from 0.83 to 0.92 and inter-observer κ_ws ranging from 0.76 to 0.89. The M2PD had intra- and inter-tester κ_ws ranging from 0.75 to 0.82 and 0.54 to 0.82, respectively. Inter-tester agreement for the pin prick test ranged from 0.45 to 0.85. There was evidence that the main source of variability between testers was testing skill and experience. Among the experienced physiotechnicians there was no significant difference between intra- and inter-tester reliability. We conclude that reliability of the SWM test was very good, closely followed by the M2PD test. Reliability of the pin prick test was less good than that of the SWM and M2PD, making it less suitable for serial testing.

RESUMEN

Se hizo un estudio sobre la concordancia de los resultados de 3 pruebas sensoriales dentro (variabilidad interna) y entre (variabilidad externa) el personal encargado de aplicarlas. Las pruebas de función nerviosa fueron la prueba del microfilamento de Semmes-Weinstein, la discriminación entre 2 puntos móviles, y la prueba del piquete de alfiler. El coeficiente de confiabilidad de los resultados se calculó de acuerdo a la estadística de K_w. La prueba del microfilamento tuvo una K_w interna de 0.83 a 0.92, y una K_w externa de 0.76 a 0.82. La prueba de disciminación entre 2 puntos tuvo una K_w interna de 0.75 a 0.82, y una K_w externa de 0.54 a 0.82. La concordancia interna para la prueba del alfiler varió de 0.45 a 0.85. La variabilidad en los resultados estuvo en razón directa de la habilidad y la experiencia de quienes aplican las pruebas. No hubieron discrepancias significativas cuando se trató de personal experimentado. Concluimos que la prueba de los microfilamentos fue la más confiable, seguida por la prueba de discriminación entre 2 puntos móviles. La

prueba del alfiler fue menos confiable que las anteriores y no es muy recomendable como una pruebas de rutina.

RÉSUMÉ

Nous avons réalisé une étude de concordance intra- et inter-observateurs pour trois tests de dépistage sensoriel utilisés pour l'évaluation de la fonction nerveuse de malades de la lèpre : le test au monofilament de Semmes-Weinstein, la discrimination de deux points mobiles et la piqure avec une épingle. Le test kappa pondéré (κ_w) a été utilisé comme cefficient de fiabilité. Les monofilaments avaient un κ_w intra-observateur allant de 0.83 à 0.92, et un κ_w inter-observateurs allant de 0.76 à 0.89. Le test de discrimination de deux points mobiles avait des κ_w intra- et inter-observateurs allant respectivement de 0.75 à 0.82 et 0.54 et 0.82. La concordance inter-observateurs pour le test à l'épingle allait de 0.45 à 0.85. Il y avait des signes que la source principale de variabilité entre les observateurs était l'habileté au test et l'expérience. Parmi les physiotechniciens expérimentés, il n'y avait pas de différence significative entre les concordances intra- et inter-observateurs. Nous concluons que la fiabilité du test au monofilament était très bonne, suivie de près par le test de discrimination de deux points mobiles. La fiabilité du test de la piqure à l'épingle était moins bonne que celle des monofilaments et de la discrimination de deux points mobiles, rendant ce test moins approprié pour des tests en séries.

Acknowledgment. We are indebted to the staff of our physiotherapy department at Green Pastures Hospital who spent much of their time assisting in this study. We are grateful to Dr. C. Ruth Butlin for allowing us to involve the Anandaban Hospital physio department in this study and to the staff of this department for their cooperation. We wish to thank Dr. Michael Hills for helpful suggestions for the statistical analysis and Prof. F. G. I. Jennekens for helpful comments on the manuscript. The work at Green Pastures Hospital is dedicated to the service and glory of God.

REFERENCES

1. Altman, D. G. *Statistics for Medical Research.* London: Chapman and Hall, 1991.
2. Becx-Bleumink, M., Berhe, D. and T'Mannetje, W. The management of nerve damage in the leprosy control services. Lepr. Rev. **61** (1990) 1–11.
3. Bell-Krotoski, J. A. Light touch-deep pressure testing using Semmes Weinstein monofilaments. In: *Rehabilitation of the Hand.* 3rd edn. Hunter, *et al.*, eds. St. Louis: C. V. Mosby Co., 1989, pp. 585–593.
4. Bell-Krotoski, J. A. "Pocket" filaments and specifications for the Semmes-Weinstein monofilaments. J. Hand. Ther. **3** (1990) 26–31.
5. Bell-Krotoski, J. A. and Burford, W. L., Jr. The force/time relationship of clinically used sensory testing instruments. J. Hand Ther. **1** (1988) 76–85.
6. Bell-Krotoski, J. A. and Tomancik, E. The repeatability of testing with Semmes-Weinstein monofilaments. J. Hand Surg. **12A** (1987) 155–161.
7. Berk, R. A. Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. Am. J. Ment. Defic. **83** (1979) 460–472.
8. Birke, J. A., Brandsma, J. W., Schreuders, T. and Piefer, A. Nylon filament measurements in leprosy patients and normal subjects in Thailand. (Abstract) Int. J. Lepr. **61** (1993) 146A–147A.
9. Birke, J. A. and Sims, D. S. Plantar sensory threshold in the ulcerative foot. Lepr. Rev. **57** (1986) 261–267.
10. Bland, J. M. and Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet **1** (1986) 307–310.
11. Brandsma, W. Basic nerve function assessment in leprosy patients. Lepr. Rev. **52** (1981) 161–170.
12. Cohen, J. A coefficient of agreement for nominal scales. Educ. Psych. Meas. **20** (1960) 37–46.
13. Cohen, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psych. Bull. **70** (1968) 213–220.
14. Dean, A. G., Dean, J. A. and Dicker, R. C. Epi Info, Version 5: a word processing, database, and statistics program for epidemiology on microcomputers. Stone Mountain, Georgia: USD, Inc., 1990.
15. Dellon, A. L. The moving two-point discrimination test: clinical evaluation of the quickly adapting fiber-receptor system. J. Hand Surg. **3** (1978) 474–481.
16. Dellon, A. L., Mackinnon, S. E. and Crosby, P. M. Reliability of two-point discrimination measurements. J. Hand Surg. **12A** (1987) 693–696.
17. Diamond, J. E., Mueller, M. J., Delitto, A. and Sinacore, D. R. Reliability of diabetic foot evaluation. Phys. Ther. **69** (1989) 797–802.
18. Dorairaj, A., Reddy, R. and Jesudasan, K. An evaluation of the Semmes-Weinstein 6.10 monofilament compared with 6 nylon in leprosy patients. Indian J. Lepr. **60** (1988) 413–417.
19. Ewing Fess, E. The need for reliability and validity testing in hand assessment instruments. J. Hand Surg. **11A** (1986) 621–623.
20. Fleiss, J. L. *Statistical Methods for Rates and Proportions.* 2nd edn. New York: Wiley, 1981.
21. Johnston, M. V. and Keith, R. A. Measurement standards for medical rehabilitation and clinical applications. In: Physical medicine and rehabilitation. Clin. N. Am. **4** (1993) 425–429.
22. Kets, C. M., van Leerdam, M. E., van Brakel, W. H., Deville, W. and Bertelsmann. Reference values for touch sensibility thresholds in healthy Nepalese volunteers. Lepr. Rev. **67** (1996) 28–38.

23. KLEYWEG, R. P. *Treatment of Guillain-Barre syndrome,* Ph.D. thesis, Rotterdam, 1990.
24. LEHMAN, L. F., ORSINI, B. and NICHOLL, A. The development and adaptation of the Semmes-Weinstein light touch-deep pressure test in Brazil. (Abstract) Int. J. Lepr. **61** (1993) 144A–145A.
25. LEVIN, S., PEARSHALL, G. and RUDERMAN, R. Von Frey's method of measuring pressure sensibility in the hand: an engineering analysis of the Semmes-Weinstein pressure aesthesiometer. J. Hand Surg. **3** (1978) 211–216.
26. LIENHARDT, C., PACQUIER, R., LE MAITRE, C. and WHEELER, J. Comparability of ball pen and nylon filaments in testing sensory function of patients with leprosy in Nepal and Ethiopia. (Abstract) Int. J. Lepr. **61** (1993) 157A.
27. MACLURE, M. and WILLETT, W. C. Misinterpretation and misuse of the Kappa statistic. Am. J. Epidemiol. **126** (1987) 161–169.
28. MCAULEY, D. M., EWING, P. A. and DEVASUNDARAM, J. K. Effect of hand soaking on sensory testing. Int. J. Lepr. **61** (1993) 16–19.
29. MCDOWELL, I. and NEWELL, C. *Measuring Health: a Guide Rating Scales and Questionnaires.* New York: Oxford University Press, 1987.
30. OTTENBACHER, K. J. and TOMCHEK, S. D. Measurement in rehabilitation research. In: Physical medicine and rehabilitation. Clin. N. Am. **4** (1993) 463–473.
31. PALANDE, D. D. and BOWDEN, R. E. M. Early detection of damage to nerves in leprosy. Lepr. Rev. **63** (1992) 60–72.
32. PEARSON, J. M. H. The evaluation of nerve damage in leprosy. Lepr. Rev. **53** (1982) 1119–130.
33. SHEIKH, K. Disability scales: assessment of reliability. Arch. Phys. Med. Rehabil. **67** (1986) 245–249.
34. SRINIVASAN, H. Prevention of disabilities in patients with leprosy; a practical guide. Geneva: World Health Organization, 1993.
35. STREINER, D. L. and NORMAN, G. R. *Health Measurement Scales: A Practical Guide to Their Development and Use.* Oxford: Oxford University Press, 1989.
36. VAN BRAKEL, W. H., SHUTE, J., DIXON, J. A. and ARZET, H. Evaluation of sensibility in leprosy—a comparison of various clinical methods. Lepr. Rev. **65** (1994)
37. VAN SWIETEN, J. C., KOUDSTAAL, P. J., VISSER, M. C., SCHOUTEN, H. J. A. and VAN GIJN, J. Interobserver agreement for the assessment of handicap in stroke patients. Stroke **19** (1988) 604–607.
38. WATSON, J. M. Disability control in a leprosy control programme. Lepr. Rev. **60** (1989) 169–177.
39. WEINSTEIN, S. Fifty years of somatosensory research: from the Semmes-Weinstein monofilaments to the Weinstein Enhanced Sensory Test. J. Hand Ther. **6** (1993) 11–22.