The Reliability of Selected Techniques in Clinical Arthrometrics

A number of studies which have examined reliability of spinal assessment procedures in manual therapy are reviewed. The tests examined were Passive Accessory Intervertebral Movements, Passive Physiological Intervertebral Movements, Straight Leg Raise and Forward Flexion. In general, tests of pain were found to be much more reproducible than tests of compliance. Straight Leg Raise and Forward Flexion tests were consistently more reliable than the Passive Intervertebral Movement tests. Possible explanations for these findings are advanced. The role of tests of compliance based on passive intervertebral movements in clinical decision-making may need to be re-examined. An appendix on reliability theory is included for the uninitiated reader.

THOMAS A. MATYAS

Thomas Matyas, B.A.(Hons), Ph.D., is a Senior Lecturer in the School of Behavioural Sciences, Lincoln Institute of Health Sciences, Melbourne.

TIMOTHY M. BACH

Timothy Bach, M.Sc., is Lecturer in Biomechanics in the School of Biological Sciences, Lincoln Institute of Health Sciences, Melbourne.

Manual therapy employs a variety of assessment techniques such as the forward flexion (FF) test, the straight leg raise (SLR) test, passive accessory intervertebral movements (PAIVM) and passive physiological intervertebral movements (PPIVM). Collectively these tests and other similar ones may be taken to define the field of 'clinical arthrometrics'.

Clinical arthrometry provides the basis for a laudably empirical approach to treatment. Among other goals, testing is variously employed to help in the selection of a region for treatment, in the selection of appropriate manual techniques and in monitoring case progress. Clearly, then, the adequacy of the assessment procedures is a major issue in the field. However, inspection of the journal literature to 1980 revealed a remarkable dearth of systematic investigations into the reliability, validity and scaling properties of the clinical assessment procedures employed by manual therapists. Consequently, a research programme was initiated in 1980 with the intention of clarifying some of these issues.

The aim of the present paper is to review several studies whose common theme is the reliability of some techniques in clinical arthrometry. The majority of studies reviewed below are part of a continuing programme of research being carried out at the Lincoln Institute of Health Sciences in conjuction with its postgraduate curriculum. Studies were conducted by postgraduate physiotherapists working under the guidance of experienced clinicians and one or both of the authors.

The paper is organized in five sections. The first section describes a method for measuring forces applied during manual procedures. The second section reviews studies on the reliability of pain measurement with three manual techniques: the PAIVM test, the FF test and the SLR test. The third section reviews studies on assessment of spinal compliance with PAIVM and PPIVM tests. The fourth section describes our studies on the reliability of producing two grades of mobilization described by Maitland (1977). Although these are not studies of assessment techniques, the findings are relevant to those of section three. Section five conducts an integrative discussion of the studies performed to date. Each section also attempts to integrate the results of pertinent publications generated outside our programme.

I. An Indirect Method for Estimating Applied Force During Therapeutic Procedures

Studies of the reliability of therapeutic techniques have been limited by a lack of objective measures of therapist performance. While therapist perceptions may be readily obtained, measurement of the mechanical effect of therapeutic intervention is confounded by the requirement that measurement techniques should not interfere with the task. To overcome this restriction, we have developed a

method which enables the indirect measurement of forces applied by therapists during mobilization and assessment techniques.

The procedure requires that therapists perform their assessment or treatment techniques while standing on a force platform. Figure 1 illustrates the position of the therapist during application of postero-anterior pressure to the lumbar spine of a patient and indicates the three forces acting on the therapist. For this situation we can write:

$$F + G - W = ma \qquad (1)$$

where W is the weight of the therapist, F is the reaction to the force applied by the therapist to the patient, G is the ground reaction force measured by the force platform, m is the mass of the therapist, and a is the acceleration of the centre of gravity of the therapist. In order to solve this equation for the applied force, F, values of W, G and a must be known. The ground reaction force G, is readily obtained from the force platform as is the body weight W, when F and a are zero. Techniques are available which enable computation of the acceleration of the centre of gravity a, but these techniques are too tedious and time consuming for routine application. An alternative approach is to make some assumptions about the behaviour of a during mobilization and assessment techniques.

For some of the experiments reported here these assumptions present little difficulty. If a therapist palpates a point in range and holds that point for a brief period of time (0.5s-1s) while recordings are made, acceleration can be assumed to be virtually zero over this period. For the purposes of this paper, this method will be termed the static force measurement technique. Similarly, if a therapist performs oscillatory mobilizations and force platform data is sampled over a much longer period of time (20 or more oscillations), the average acceleration over the sampling period will be virtually zero (otherwise the therapist would acquire a net positive or negative velocity). The difference between body weight and the measured ground reaction force is an accurate estimate of mean applied force in both cases.

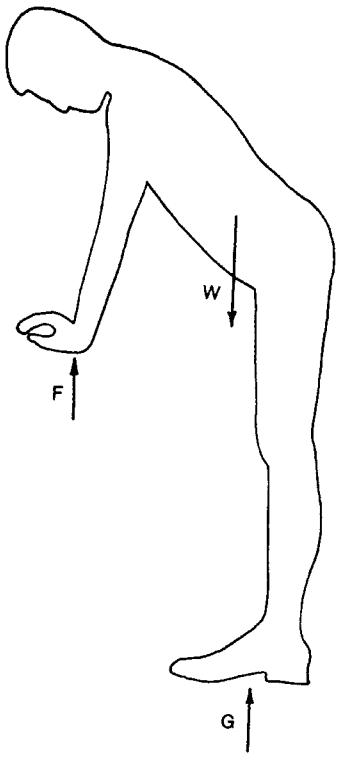

Figure 1: The forces which act on a therapist performing spinal mobilization or palpation are body weight, W; the ground reaction force, G; and the reaction to the force applied to the patient, F.

In other experiments considered here, estimates of oscillation amplitude and peak applied force were required. The method employed in these studies will be termed dynamic force measurement. Instantaneous values of applied force are much more susceptible to inertial effects than average values. Bach (1985) has adopted an empirical approach to estimate the degree of error involved in using the force platform output (estimated force) as an indirect measure of applied force under different conditions of movement amplitude and frequency. Bach (1985) found that the error associated with oscillation amplitude measurement by this technique was approximately 12%. The error of estimating peak forces by this technique was in the neighborhood of 1-3% depending on characteristics of the applied forces.

In the studies reviewed in this chapter we have measured only the vertical component of force. Many assessment and treatment techniques require that force components other than vertical be applied. However, studies described here concentrated on postero-anterior central vertebral pressures on prone patients and therefore primarily vertical forces were involved. In one experiment (Collis-Brown 1982) involving 192 measurements of applied force during a posterior-anterior PAIVM assessment the mean difference between the vertical component of the applied force and the total applied force was 1.8N. This represented 0.5% of the total range of measured forces. We have therefore chosen to neglect horizontal components of the applied force in techniques involving primarily postero-anterior movements.

An unresolved issue is that of the pressure distribution between therapist and patient. In studies of applied force reported here therapists were required to use the pisiform techniques as described by Maitland (1977, p.137). This technique involves placing the hands so that the point of contact with the spinous process is the medial border of the hand between the pisiform and the hamate. The purpose of this placement is to localize the pressure distribution as much as possible. The proportion of total applied force which acts on the vertebral body itself could differ between therapists and between patients as a result of anatomical variation in soft tissue distribution in both the hands of therapists and the backs

of patients. To our knowledge, there is no method available for obtaining precise information on these pressure distribution patterns but differences are likely to be very small. Furthermore, these errors are fixed by the experimental designs employed: the therapists' hands do not change; the individuals tested and retested are the same; the anatomical loci are the same. Thus only absolute force values will be subject to pressure-distribution error. Reliability coefficients, which are only affected by random error, will not be influenced (see Appendix).

II The Reliability of Some Movement Tests of Pain in the Lumbar Spine

Tests of pain may employ either passive movements as in PAIVM, PPIVM and SLR tests or active movements as in the FF test. These tests are employed to chart 'pain behaviour' (Maitland 1977). Although other features, such as 'quality' of pain, may also pertain, 'pain behaviour' is often conceived as a two dimensional function: pain versus range of movement (ROM). Key features of this function are: the point in ROM of pain onset (P1); the pain intensity at the limit of movement (when the limit is caused by factors other than pain), or the point in ROM where pain is of sufficient intensity to limit movement (P2); and the dynamics of pain intensity between P1 and P2, *ie* the nature of the change in pain intensity as a function of ROM. Pain assessment is an essential feature of initial diagnosis, acute pre-post evaluation of manual intervention and longer term evaluation of intersession development. Therefore intertherapist reliability, within-session test-retest reliability and between-session test-retest reliability are all relevant practical issues for evaluating P1, P2 and pain dynamics. Our studies to date have examined only some of these issues.

Results with the PAIVM test

Collis-Brown (1982) and McNeill (1982) examined in a within-session design the test-retest and intertherapist reliability of locating P1 in ROM when using PAIVM. Four physiotherapists with postgraduate qualifications in manual therapy examined two segments from each of 12 patients. Patients were included if prior examination revealed: a history of back pain or current back pain; a non-irritable condition; and discernible pain onset in at least two lumbar levels on application of PAIVM. Patients were examined in prone with the two relevant lumbar levels pre-marked. As much of the upper and lower body was covered as was possible in order to reduce body identity. No communication was permitted other than the response 'Now' to the question 'Tell me when the pain starts'. The static force measurement technique described earlier was used to measure applied forces. Therapists recorded their conclusion on a 100mm visual analogue scale (VAS). This permitted simultaneous measurement of the force at which P1 occurred and the subjective distance from ROM origin where P1 occurred according to the therapist. To control for series effects therapists examined the patients in a latin square design (Meyers and Grossen 1974), with patients randomly allocated to four groups of three. After three patients were examined by all therapists the entire procedure was repeated. The experimental design therefore provided 24 test and 24 retest measurements from each of four therapists under conditions which attempted to minimize information other than the P1 response to PAIVM.

Collis-Brown (1982) found that the average test-retest reliability coefficient for palpation conclusions was 0.73. This was only a little less than the average correlation between the test and retest forces required to produce a P1 response, which reached a value of 0.83. The difference between the two coefficients was not statistically significant. In terms of classical reliability theory this implies that 27% of the variance in P1 observed by palpation was due to random error. This error may be conceived as a composite result of at least two processes: random changes in the patients's pain condition, or in the verbal report; and random error in the therapist's ability to perceive the point in ROM where P1 was reported and record it on the VAS. An estimate of the first component may be obtained from the test-retest correlation of the forces, which do not depend on therapist perception and recording ability. This method estimates that 17% of the observed score variance was due to random error in the patient's report, although the true value will be somewhat lower because an amount should be allowed for the random error in force measurement. Nevertheless, it was apparent that random error due to therapist perception and recording was small.

From a practical point of view, however, random error destroys judgement reliability irrespective of its genesis in the patient or the therapist. To make interpretation of patient changes clearer, confidence intervals were computed for the therapist judgements. These estimated that for 95% confidence that a change does not reflect merely random error, a therapist must observe a change of at least 34% of full scale on the VAS. In clinical situations confidence as low as 80% may sometimes suffice. This was estimated to require a change of at least 22% on the VAS. It is difficult, given the lack of evidence on the size of the effect requiring measurement, to decide if the random error is sufficiently small.

McNeill (1982) examined the degree of intertherapist reliability present in the above experiment. The average intertherapist correlation was 0.62. This indicates that a substantial proportion of the variance in observed scores (38%) was attributable to intertherapist variation in performing the test. The intertherapist correlation in forces required to produce P1 was 0.75, which was not significantly lower than the

intratherapist value of 0.83. A large portion of the variability in intertherapist correlations was attributable to random error in patient report (25%) and a smaller portion to differences between therapists (13%).

The effect of conducting a broader PAIVM test, including compliance features, spasm, and a complete chart of 'pain behaviour', was investigated subsequently by Flint (1983). Four manual therapists with postgraduate qualifications independently examined one lumbar level from each of twelve patients. The patients were selected from several clinics providing that a screening physiotherapist identified, following a full examination (Maitland 1977), current back pain attributable to the lumbar region. The patients were examined in a latin square sequence as in the earlier study. The movement diagram described by Maitland (1977) was employed as a two dimensional VAS of Intensity x ROM (67x100mm). Therapists recorded P1, P2, the dynamics of pain between P1 and P2, as well as the other features typically required by a Maitland movement diagram: the limit of range (L); the point in ROM of resistance onset (R1); limiting resistance (R2); the dynamics of resistance between R1 and R2; and the behaviour of muscle spasm, if present (Maitland 1977). The screening physiotherapist premarked the level to be tested, which was the 'most symptomatic' level found in the prior examination. Therapists were required to palpate only the marked level using central PAIVM. No other patient information was given to the therapists.

Flint (1983) found that the mean intertherapist correlation for locating P1 in ROM was 0.48, somewhat lower than the 0.62 obtained by McNeill (1982). Although this difference is not statistically significant, the result indicates that additional palpation information failed to improve the reliability of P1 ratings.

Furthermore Flint's sample had a more acute status than that employed by McNeill. Thus the result also failed to support the hypothesis that P1 ratings from more acute patients would provide better reliability because acute patients are likely to have a clearer pain onset, with a distinct 'bite of pain' (Maitland 1977, Collis-Brown 1982, McNeill 1982).

As a part of the same study, Flint also examined intertherapist reliability in measuring pain intensity at P2. She found a mean intertherapist reliability coefficient of 0.75, a relatively good result and the best intertherapist reliability coefficient obtained to date in our PAIVM investigations. It is interesting to note that this feature of the movement diagram is probably more reliant on the patient's response and less reliant on the therapist's ability than any other PAIVM finding.

A final aspect of the reliability of pain assessment investigated by Flint was the degree of intertherapist agreement on whether pain, spasm or resistance was the cause of movement limitation. The mean pairwise agreement was 66.6% which proved significantly higher than the expected random agreement rate (51.8%) given the obtained base rates. Nevertheless, an intertherapist disagreement rate of 32.4% is substantial in a practical sense, since the decision about the cause of movement limitation plays a significant role in selecting treatment approach (Maitland 1977).

Results with the SLR test

The SLR is a widely used test, recommended (Cyriax 1982) for both diagnosis and progress evaluation. It is associated with a considerable body of literature discussing its underlying processes (Goddard and Reid 1965, DePalma and Rothman 1970, Murphy 1977, Breig and Troup 1979 and Cyriax 1982). Like the PAIVM test it employs passive movement, but the movement is 'physiological' rather than 'accessory'.

McFarlane (1981) examined the reliability of assessing pain onset as a point in ROM during the SLR. Twenty patients with low back pain of recent origin were selected from several Melbourne hospitals provided that they did not show an unusually high anxiety component, or failed to show a change in symptoms under 80° of SLR, or showed restricted movement or pain in the squatting test. Five SLR tests to pain onset were performed on each subject with a 90 second inter-test interval. A gravitational goniometer, was used to record the angle at P1. Medial hip rotation was manually controlled as suggested by Breig and Troup (1979).

A mean test-retest correlation of 0.96 was found between adjacent pairs of trials, indicating a very high reliability for this test. On the basis of McFarlane's data we calculated that a change of at least 13.6° should be observed in P1 if changes due to random error are to be excluded with a certainty of 95%. If typical normal ROM is estimated around 90° (DePalma and Rothman 1970, Cyriax 1982) the 95% confidence interval for test-retest change is 15% of scale, which is better than the 34% obtained with the PAIVM test (Collis-Brown 1982). Thus both metric and metric-free estimates of reliability show better values for the SLR test.

In addition, McFarlane examined the possibility that systematic trends may occur in the SLR data. She found that the range to pain onset increased between successive tests by an average of 1.2° which was a statistically significant effect. Therefore increases in range to pain onsets of 15° would probably be safer minima for error free estimates of therapeutic improvement between succeeding tests obtained within session.

A subsequent experiment performed by Puentedura (1983) to examine the effects of trunk position on the SLR test indirectly yielded confirmatory evidence of high reliability for this test. Puentedura recorded pain onset and limiting pain in seventeen young, non-symptomatic subjects who reported no history of chronic musculoskeletal illness. Electrogoniometric readings were obtained with the trunk in three posi-

tions: neutral, maximal contralateral flexion and supported lumbar lordosis. All tests were performed in supine on a flat surface. Within each posture ten pain onset and two limiting pain observations were performed. However, since repeated measures within each posture were obtained with no intervening treatment, we were able to re-examine Puentedura's raw data for test-retest reliability coefficients. Regardless of posture these proved to be uniformly high. The mean test-retest correlation between adjacent pain onset trials within a posture was 0.98. The limiting pain data yielded an average correlation of 0.96. These results confirm and extend those of McFarlane.

In the period between the studies of McFarlane and Puentedura three publications appeared (Hoehler *et al* 1982, Lankhorst *et al* 1982, Million *et al* 1982) which seem to further confirm the high reliability of the SLR test. Million *et al* (1982) found a within session retest reliability of 0.97 using nineteen patients. Lankhorst *et al* (1982) using an active SLR reported error components for both interobserver and interday aspects from a factorial design applied to 48 low backache patients. From their results we calculated an interday test-retest reliability of 0.96-0.97 and interobserver reliability of 0.93-0.96. Slightly poorer results were found by Hoehler and Tobis (1982) for interobserver reliability when measuring passive SLR ($r = 0.78$), although for active SLR the results were comparable ($r = 0.95$).

Results with the FF test

Like the SLR test, the FF test involves 'physiological' movement. However the test is one of active rather than passive movement. During active forward bending in the sagittal plane, with the knees extended, several parameters may be recorded. These include ROM to pain onset (P1) and ROM to maximum pain tolerance (P2) among others. The test is widely used as a part of various approaches to examination of the lumbar spine (Maitland 1977, Stoddard 1980, Cyriax 1982). The purpose of this subsection is to review four studies our group performed on the reliability of the FF test for measuring pain parameters.

Several methods for recording ROM during FF tests have been reported including skin distraction (Macrae and Wright 1969, Van Adrichem and Van Der Korst 1973), spondylometry (Twomey and Taylor 1979, Stoddard 1980), inclinometry (Loebl 1967), tangential hydrogoniometry (Anderson and Sweetman 1975), radiography (Hauley *et al* 1976) and photography (Troup *et al* 1967). Some of the previous literature investigating the adequacy of these measurement methods has been concerned with their relative value for assessing spinal mobility (Troup *et al* 1967, Van Adrichem and Van Der Korst 1973, Reynolds 1975, Moran *et al* 1979). Much of the evidence has been collected from normal samples (Loebl 1967, Troup *et al* 1967, Van Adrichem and Van Der Korst 1973, Reynolds 1975, Moran *et al* 1979). The purpose of the studies reported below was to examine pain measurement with a view to clinical application. Therefore simplicity was a criterion for selecting the approach to measuring ROM. This excluded radiographic and photographic methods.

The method adopted was to measure fingertip position using a measuring tape (Kapanji 1974). Apart from its simplicity this method seemed appropriate because kinesiological analysis suggests that it is influenced not only by spinal movement, but also by hip movement and a variety of associated structures including muscle and connective tissue (Farfan 1973, Van Adrichem and Van Der Korst 1973, Hart *et al* 1974). While this is a disadvantage for the assessment of specific mobility in the lumbar spine (Moll and Wright 1976), it may be an advantage in the measurement of pain, particularly pain progress, where a variety of structures may be implicated.

Kwong (1981) investigated the test-retest reliability of assessing P1 with the FF test. Twenty patients attending a physiotherapy clinic were sampled provided they had low back pain without either hip involvement, a list, or scoliosis. Patients were assessed in briefs and bare feet after the prominence of the tibial tuberosity was marked. They were required to bend forward, sliding their hands down their thighs, without deviation from the sagittal plane, until pain onset. Patients with pre-existing background pain were instructed to stop on onset of a pain change. Using the tibial tuberosity mark as origin, ROM to P1 was recorded by measuring the distance to the tip of the midfinger with a nylon tape. Three measurements with an intertrial interval of one minute were obtained from each patient. The mean test-retest correlation between trial pairs was 0.98, indicating very high reliability. Using Kwong's data we calculated that changes of 83mm or more would give 95% confidence that the observed change was not the result of random error of measurement.

Systematic error due to repeated measurement was assessed by comparing the central tendency in the three samples. No statistically significant differences were obtained, although there was a suggestion that an initial practice trial might stabilize the data.

Using the same FF measurement technique, Bruce (1981) investigated the test-retest reliability of assessing the point of limiting pain (P2). Twenty patients with low back pain were selected from a private physiotherapy clinic provided they were not restricted by bilateral hamstring tension, or had less than 60° ROM, or had an 'irritable condition'. Patients were randomly allocated to two groups of ten. Three measurements were taken from all subjects. An objective examination of the spine (Maitland 1977) was interpolated in one group between the first and second measure and in the other group between the second and the third measure. The other intertrial intervals were

three minute rests. Test-retest correlation, when only rest intervened between the two trials, was 0.98. This was consistent with Kwong's results. Test-retest correlations between trials separated by the objective spinal assessment were 0.87 and 0.99. Using the reliability coefficient of 0.98 and Bruce's raw data we calculated that changes of 33mm or more would give 95% confidence that the observed change was not the result of random error of measurement. No systematic bias due to repeated measurement was found, replicating Kwong's data.

The studies performed by Bruce and Kwong were limited to assessing within-session retest reliability. While evaluation of within-session progress is a main use of assessment in manual therapy, the results of Bruce and Kwong are not necessarily generalizable to between-session retest intervals. Therefore Patterson (1982) examined limiting pain in FF on two consecutive days. Three FF tests were conducted on Day 1 separated by one minute rest intervals. The procedure was repeated on Day 2. A sample of 12 subacute or chronic low back pain patients were selected, using similar criteria to those of Bruce and Kwong. The mean within-session retest reliability was found to be 0.98, confirming the findings obtained by Bruce. The mean between-session retest reliability was 0.97, not significantly lower than that obtained within-session. The 95% confidence interval for measuring changes within-session was 45mm, slightly higher than that obtained by Bruce. The 95% confindence interval for measuring changes between days was 52mm.

Maitland (1977, p.171) recommends that a therapeutic effect should only be assumed if an improvement of 25mm or more in limiting pain is obtained. This conclusion, based on clinical observation and in the absence of formal analysis, compares well to our experimental estimates. In terms of the random error estimate obtained by Bruce, changes in excess of 25mm afford 87% confidence. In terms of Patterson's within-session estimates 25mm changes afford 76% confidence. Between-session conclusions should be taken even more conservatively: our calculations based on Patterson's data estimate only 68% confidence for a minimum change of 25mm.

Another possibility for measurement error on reassessment is that repeated exposure to the same test may create a systematic bias. Serial effects may occur as a result of changes in the relevant anatomy/physiology caused by the initial test, placebo phenomena, or simply skill learning. In Kwong's study no statistically significant differences were obtained among the three trials. The largest mean difference was only 6mm and occured between the samples of trials 1 and 2.

The most recent study in this series was designed to determine if the high reliability found in the three previous studies was an artefact of the way the test was performed. At least two obvious hypotheses might be invoked to suggest that the high retest reproducibility resulted from factors other than pain sensation. One hypothesis is that visual and tactile feedback was available to patients in these studies since they could see their own performance and the test procedure required the hands to slide down the legs. Another hypothesis, more difficult to test, is that the high reproducibility merely represents memory for movement rather than recurrence of a given pain level at the same point in ROM.

To investigate these hypotheses Munro (1983) examined the within-session retest reliability on a modified FF test. The test was performed with a blindfold. Furthermore, instead of sliding their hands down their legs, subjects were required to bend forwards while depressing a low-friction plunger vertically with the tips of the middle fingers (Moll and Wright 1976). The plunger was part of an apparatus containing a metric scale and pointer which permitted location of movement endpoint to the nearest millimetre. A final modification to the previous procedure was that a simple motor task (manipulation of a nut and bolt) was interpolated between the test and the retest in an effort to produce some disruption in sensorimotor memory. Two groups of subjects were tested. The first group comprised 17 low back pain patients selected along criteria similar to those of the earlier studies. The second group comprised 17 asymptomatic subjects. Subjects were selected in the asymptomatic group on a matched pair basis with a low back pain subject. The matching criteria were gender and age parity (within 6 years). The asymptomatic member of each matched pair was required to perform a task yoked to the initial performace of the low back pain subject. A mechanical block, placed at the same point where the symptomatic subject showed pain onset, was used to stop forward bending of the asymptomatic subject during the test. During the retest, which followed the interpolated task, the block was not present and asymptomatic subjects were required to simply stop at the point as they recall it from the initial test. Symptomatic patients were required to stop on pain onset during both tests.

Despite the blindfold and the interpolated task, symptomatic subjects showed a test-retest correlation of 0.99. Statistical analysis revealed that this was significantly higher than the correlation shown by the asymptomatic group (0.92). The high reliability obtained confirmed the earlier FF data (Bruce 1981, Kwong 1981, Patterson 1982). More importantly, however, the superior reliability of the symptomatic group under these stringent performance requirements suggests that pain sensation was contributing, rather than visual or tactile feedback. Similarly, performance on memory alone can be rejected, although a more convincing demonstration could probably have been obtained by employing a longer intertest interval and an interpolated task using the same joints as the FF test, but which does not aggravate the pain.

In conclusion, our studies of the FF test for pain have consistently produced high reliability estimates and suggest that pain is indeed being accessed. This finding is in contrast to the deprecatory conclusions of some other authors (Hart *et al* 1974, Reynolds 1975, Moll and Wright 1976, Moran *et al* 1979). The FF test is said not to be a good measure of spinal mobility (Hart *et al* 1974, Reynolds 1975, Moll and Wright 1976). This may be so but the point is irrelevant to the measurement of pain and its progress. The FF test is said to be influenced by structures other than those of the lumbar spine (Reynolds 1975, Moran *et al* 1979). We have already addressed this issue indicating that from the point of view of monitoring pain progress this may be an advantage. In general therefore, results indicate that the FF test should not be overlooked as a simple and reliable clinical test for assessing pain changes, particularly if other aspects of the assessment have established the nature of the underlying pain process. Finally, it is interesting to note that the reliability coefficients of the SLR and FF tests, both of which involve 'physiological movements', were comparable and consistently higher than those obtained for PAIVM tests of pain.

III The Reliability of Some Clinical Procedures for Assessing Compliance

Manual tests of spinal compliance probably form the most characteristically unique contribution of manual therapy to the diagnostic armamentarium. Their objective is to employ the therapist's perception of displacement and 'resistance' to obtain a subjective model of spinal compliance, which can be used for a variety of decisions (Maitland 1977). That this involves a perceptual model of spinal compliance, including dynamic parameters, can be seen most clearly in the development of the two-dimensional movement diagram (Maitland 1977). Manual assessment of compliance contains, in counterpart to pain assessment, some key parameters: the point in ROM of resistance onset (R1); the point in ROM where resistance limits passive movement (R2); and the compliance function which links R1 and R2. Compliance tests have a role in: initial diagnosis, including selection of the level to be treated and type of mobilization to be utilized; the evaluation of progress within-session following treatment; and progress between sessions (Maitland 1977).

Consequently test-retest and intertherapist reliability are relevant issues. The majority of our studies to date have been concerned with PAIVM (Baker 1981, Millman 1981, Wong 1981, Weeks 1982, Allen 1983, Flint 1983) although one study involving PPIVM (Clarkson 1982) is also reported below.

Studies evaluating the reliability of R1 and R2 assessment with PAIVM

Despite the relatively widespread use of the PAIVM assessment procedures described by Maitland (1977), a review of the literature prior to 1981, the time of our group's initial study (Baker 1981, Wong 1981), revealed a remarkable dearth of systematic attempts to evaluate the reliability of these procedures.

An initial study designed to estimate intertherapist reliability for locating R1 and R2 in ROM was conducted by Baker (1981) and Wong (1981). Three therapists independently examined six spinal levels from each of eighteen subjects. The subjects had an age range of 18 to 54 and no history of recent spinal pain. The six levels examined were C2, C6, T2, T10, L2 and L4. The three cephalad processes were examined with thumbs in apposition. The three caudad processes were examined with pisiform technique. Each therapist was required to mark R1, R2 and the compliance function linking them on a 45x60mm movement diagram (Maitland 1977). The ROM to R1 and to R2 was then obtained to the nearest millimetre. Using these measures, intertherapist correlation coefficients for each joint were then obtained for each pairwise combination of therapists.

Intertherapist correlations were lower than those obtained in PAIVM tests of pain. The mean coefficient for R1 across all spinal levels was 0.30. The best mean correlation for a single level was 0.64, obtained from L4. This was significantly superior to the other coefficients obtained. The mean correlation for R2 across all spinal levels was 0.28 and the best mean correlation for a single level was 0.58, obtained from L2. The L2 value was significantly superior to that of C6, T2 and T10. Other differences between the reliabilities given by the six levels were not statistically significant. Although the mean reliability coefficients of 0.30 and 0.28 were statistically significant, they were disappointingly low.

In a subsequent study Weeks (1982) examined the within-session and interweek test-retest reliability for locating R1. Four therapists independently examined three joints from each of twelve subjects. None of the subjects had a history of recent spinal pain. The age range was 20-50 years. Each therapist palpated C2, T4 and L5 on two occasions one week apart. Within each session the joints were assessed twice on a rotational basis across the twelve subjects, *ie* the examination of eleven subjects intervened between the first and second assessments within the session. Therapists were required to mark the location of R1 on an 80mm VAS marked in quarters. Apart from the areas to be examined, subjects' bodies were draped.

Distances to R1 were then used to compute, for each segment, the within-session and interweek reliability coefficients for each therapist. The within-session correlation was 0.46 when averaged across all four therapists and all three joints. The interweek reliability coefficient averaged an extremely poor 0.09, which was significantly worse

than even the disappointingly low within-session correlation.

Since the four therapists examined the same subjects it was also possible to replicate the estimate of intertherapist reliability obtained by Wong (1981). Over both days and across all joints the mean pairwise intertherapist correlation was 0.25, confirming the low estimate obtained by Wong.

In general, therefore, the two studies indicated that PAIVM assessment of compliance parameters has poor reliability. However these estimates should be interpreted in the light of two methodological issues which weaken the generalizability of the estimates. The first issue is that in both studies the sample comprised trainee therapists in the second half of the postgraduate diploma specializing in manual therapy. It is possible to argue that such a sample may not have been representative of the ability which a sample of fully trained and more experienced practitioners would demonstrate.

The second issue of generalizability refers to the sample of subjects used by the two studies, which in both cases had no recent history of spinal pain, unlike the subjects typically seen in clinical practice. The quantification of reliability is affected to a degree by the range of individual differences among the joints examined. The mathematical theory of reliability clearly indicates that restricting the range of variation will tend to reduce the reliability coefficient (see Appendix). The reliability coefficient is the ratio of the true-score variance to total (true-score plus error) variance. The size of the error variance may be assumed to remain constant over the sample as a whole when the same method of measurement is employed. However, if the true-score variance is reduced because true individual differences between the measured entities has been reduced, then the random error component will be a larger *proportion* of the total variation and the overall correlation coefficient will be reduced. In other words, if the range of variation in compliance parameters which results from individual differences in a non-clinical sample is substantially different from the range obtained in clinical samples, the reliability estimates obtained will tend to be biased. The issue being an empirical one, the logical approach is to examine a clinical sample. Flint (1983), whose results have been reported in part above, chose that approach.

The study carried out by Flint in contrast to those of Weeks and Wong, employed a clinical sample; gave therapists an 'ecologically valid' assessment task, since they were required to do a full pain and passive movement diagram on a clinical subject; and used four fully qualified therapists with post-qualification experience ranging from nine months to three years. The intertherapist reliability coefficient for locating R1 in ROM was found to be 0.38 on the average, which is not significantly higher, in either the statistical or the practical sense, than that of 0.30 reported by Wong.

The reliability of differentiating spinal levels on the basis of compliance perception following PAIVM's

In clinical assessment PAIVM tests may be used in the attempt to locate compliance parameters on a perceptual ratio scale so that they may be used to guide diagnosis, assess progress and assist in the selection of grades of therapeutic movement typical of the approach described by Maitland (1977). This purpose guided the orientation of the studies reported in the previous subsection. An alternative purpose for PAIVM tests is to assess the presence of compliance abnormalities by palpation, on a comparative basis, across several spinal levels. Relevant parameters include 'end feel', soft tissue resistance and postero-anterior amplitude of joint movement (Maitland 1977).

Millman (1981) examined test-retest and intertherapist reliability for blind discrimination of the stiffest spinal level. Therapists were blindfolded and required to select, only by performing PAIVM with pisiform, which of the six unidentified levels presented in random sequence was stiffest. The levels included were L4 to T11.

Therapists were permitted to repalpate any levels they were uncertain about until they came to a firm decision. Each of three therapists examined the same thirteen nonclinical subjects on two occasions within one session of testing. The results indicated that preconceptions about anatomical variation in stiffness were adequately controlled by this procedure because therapists' ability to identify which anatomical levels they were on was not significantly better than that attainable by chance. Furthermore, therapists were unable to guess at better than chance rates when they were performing a retest.

Under these conditions, which imposed a strict dependence on palpatory information, the mean test-retest agreement rate was 31%. Statistically, this was significantly better than the agreement rate of 16.7% predicted by a model which assumed that therapists were randomly selecting one level among six. The analysis also showed that 31% was significantly worse than the agreement rate of 50% predicted by a model which assumed that therapists were able to reject four levels with certainty, but were guessing which of the remaining two levels was stiffest. The best model was that which assumed therapists were able to reject three of the levels but guessed among the remaining three. These models are, of course, imaginary. They should not be taken to imply that therapists decide literally following the processes assumed by these models. The models do however provide a valuable frame of reference for interpretation.

The analysis of intertherapist agreement showed that the average pairwise agreement was 25.7%. This was significantly better than the 16.7% predicted by a model assuming complete guessing. It was also significantly worse than the 33% predicted by a model

which assumed that therapists were able to reject three levels with certainty, but had to guess among the remaining three spinal levels.

Millman's results therefore suggested that by palpation alone therapists can discriminate better than chance those differences in stiffness derived from anatomical variation. Unfortunately, the degree of agreement, though better than chance, was nevertheless low from the point of view of practical diagnostics. For example, it seems likely that a therapist would be able to narrow the range of clinically relevant levels down to three, or perhaps even two, by using the case history, the other test data and epidemiological knowledge.

However, the generalization of Millman's data faces some problems. First, the source of variation between spinal levels was that due to natural anatomical differences in non-symptomatic spines. In clinical decision making the stated objective is to identify the presence of an abnormality. The frame of reference for the therapist presumably is some cumulated memory model of what is normal (Maitland 1977). Whether the difference between the immediate perceptual trace from an abnormal joint and the cognitive template of normality is an easier discrimination to perform than the discrimination between recently experienced perceptions of stiffness which differ according to anatomical variation between spinal levels, seems to be a moot point in the light of the complexity of the issue and the lack of evidence.

A second problem stems from the choice of 'stiffest level' (Millman 1981) as the object of discrimination. In clinical theory the finding of abnormality may involve a broader base of compliance features. These include 'end feel' and soft tissue resistance as well as postero-anterior ROM (Maitland 1977).

A third problem arises from the nature of the therapist sample. The three therapists all had a minimum of four years clinical experience. However, although they had satisfactorily completed more than half of the postgraduate diploma specializing in manual therapy, including the spinal assessment and treatment portion of the course, it may be that the lack of full qualification and post-specialization experience was a factor in their performance.

Allen (1983) conducted a study which attempted to resolve some of the issues raised by Millman's study. Five lumbar levels from each of twelve patients recruited from several clinics were examined. All patients had a history of back pain. Seven patients had symptoms which had persisted over six months. Three physiotherapists with specialist prostgraduate qualifications in manual therapy and a minimum of eighteen months of post-specialization experience performed the assessments. Millman's procedure was replicated, but therapists were asked to select which level had the greatest soft tissue resistance, which had the most abnormal 'end-feel', and which had the smallest postero-anterior amplitude of movement. In addition, therapists were required to indicate which of the five levels should be selected for treatment and which of the three indicators of abnormality mentioned above had most influenced their selection.

Allen's data revealed a very high degree of coherence between the specific indicators of abnormality. On over 97% of occasions two or three of these indicators identified the same level as that selected to be 'most abnormal'. Therefore reliability estimates were prepared only for the decision of which level should be selected for treatment. The test-retest agreement rate averaged 47.2%, somewhat higher than Millman's 31%. However, our analysis of the results obtained by these studies did not indicate the improvement to be statistically significant. The inter-therapist agreement rate averaged 26.4% on a pairwise basis in Allen's study. This is very similar to Millman's result and not significantly better than the 20% agreement rate which would be expected from a random guess model. The high coherence between specific indicators seems to imply either that abnormal compliance tends to manifest simultaneously through the several parameters, or that therapists tend to be biased towards 'false alarms' of abnormality having found a single abnormal sign from the level in question. The low degree of reliability suggests that the latter explanation should be preferred. Furthermore, since the test-retest reliability indicates that some degree of consistent information was transmitted even though intertherapist agreement was very low, it seems reasonable to hypothesize that therapists make global judgements of abnormality, on perceptual dimensions which are probably not consistent and which may be difficult to verbalize.

The reliability of compliance ratings following PPIVM tests

Passive movement of a 'physiological' type provides another testing approach which may be used for diagnosis or progress evaluation (Maitland 1977, Cyriax 1982).

Kaltenborn and Lindahl (1969) examined the intertherapist reliability of ten therapists during assessment of intervertebral joint mobility. A four-point rating scale consisting of no movement, hypomobility, normal movement and hypermobility was used. Kaltenborn's ratings were used as a criterion for agreement. Each of the therapists independently gave 13 assessments. Their conclusion of 'remarkably good' agreement was not accompanied by a formal analysis. However, the following results were reported: complete agreement from three therapists; 2 disagreements from two therapists; 3 disagreements from one therapist; and 4 or 5 disagreements from the remaining three therapists. This represents an average agreement rate of about 84%.

Gonnella *et al* (1982) examined the intertherapist and retest reliability of

five therapists employing PPIVM tests on lumbar segments. On each of two days, which were separated by a 13 day interval, each therapist independently evaluated the six segments of five young, nonsymptomatic subjects. Two evaluations, one under 'normal' and one under blindfold conditions were performed within each session. Forward bending, side bending (left and right) and rotation (left and right) were performed. A seven point rating scale was used, with 'ankylosed' and 'unstable' as the end values. In addition 'plus' and 'minus' qualifiers were permitted, which produced a potential 13-point scale. In practice the scale values employed by the observers were limited to the range 1-4, producing an effective seven-point scale biased towards hypomobility. In fact, the distribution was probably even more restricted because the extreme scale values (1.0 and 4.0) seem to have occurred very infrequently (*eg* 2% for forward bending, the only test for which sufficient data was available to extract a result). Gonnella *et al* concluded that 'results on intertherapist reliability were disappointing' (p.442). Although this conclusion is not immediately apparent from their analysis of the data, our reanalysis of the evidence Gonnella *et al* presented (p.440) confirmed their conclusion. For example, with the forward bending manoeuvre we calculate that intertherapist agreement reached 78% when agreement is defined (Gonnella *et al* 1982) as ratings differing by less than one full scale unit. However, the agreement rate expected from the chance agreement model is 71%. The high degree of chance agreement is the combined effect of a restricted distribution of mobility together with a definition of agreement which accepts a variation of half a scale value (see Appendix).

Thus the PPIVM research literature presented until 1982 a somewhat equivocal overview. One study claimed good results for intertherapist reliability (Kaltenborn and Lindahl 1969), while another found poor results (Gonnella *et al* 1982). A further problem was that of non-generalizability of findings, either because the evaluation samples included few subjects (Kaltenborn and Lindahl 1969) or nonsymptomatic subjects (Gonnella *et al* 1982).

Therefore, Clarkson (1982) investigated the intertherapist reliability of four experienced physiotherapists specialized in manual therapy. The test sample comprised ten subjects aged 20-55, all of whom had a history of low back pain. One subject had a radiographically confirmed sacralization of L5. Others included a retired dancer, a footballer and a champion runner. That is, there was an effort to obtain a wide cross-section of test joints. Each therapist independently assessed each vertebral segment from S1 to T12 using the PPIVM technique for forward flexion described by Maitland (1977). Therapists used a five-point scale with the end values being 'ankylosed' and 'hypermobile'. On the average, the pairwise intertherapist agreement rate was 45%. Statistically this was significantly better than the 37% expected to occur from chance agreement. However, from a clinical point of view it does not seem a very encouraging result. When the 'stiff' and 'very stiff' ratings were amalgamated to produce a four-point scale like that of Kaltenborn and Lindahl (1969) the agreement rate became 57%. This seems substantially lower than the 82% obtained by Kaltenborn and Lindahl. The results are also poorer than the 78% agreement rate obtained by Gonnella *et al* (1982), although the comparison is complicated by differences in the rating scales used.

Further evidence about the reliability of PPIVM is available outside the research literature of physiotherapy. Rotational manoeuvres similar to the techniques employed by physiotherapists are encountered in osteopathy (Johnston 1982). Recently, Johnston *et al* (1982a) reported on the intertherapist reliability obtained by one osteopathic physician and two student physicians. The tests employed were cervical rotation, cervical sidebending and several trunk motions. The experimental sample comprised 161 volunteers which included 84 students and 71 patients. However the report does not clarify the particular characteristics of the subsamples used to assess the reliability of the different motions. Therapists were required to indicate if resistance to passive motion was symmetrical or asymmetrical for left and right manoeuvres. For cervical rotation the three therapists agreed on 42% of the 43 subjects tested this way. For cervical sidebending they agreed on 33% of 36 subjects. Although these agreement rates appear rather low they were significantly higher than those expected to occur by chance (19% and 14%, respectively). Furthermore, these are three-way agreements rather than pairwise agreement rates as in the other studies reviewed by this section. Unfortunately the report by Johnston *et al* 1982a makes extraction of mean pairwise agreement difficult, thereby precluding direct comparisons. In addition the ratings required were somewhat different. Nevertheless, in terms of clinical significance, the results seem rather disappointing, a conclusion shared by Johnston *et al* (1982a).

In a subsequent study on cervical rotation, Johnston *et al* (1982b) evaluated intertherapist reliability when only subjects with strong indications of asymmetry were included in the sample. Preselection of subjects was based on agreed examination findings by two faculty osteopaths. Three student therapists then independently examined the subjects. The pairwise agreements for each student with the faculty examiners were 71%, 62% and 57%. While the agreement rate from the first student was significantly higher than expected to occur by chance, this was not the case for the other two sets of ratings. Given the preselected sample of subjects and the restriction of ratings to symmetry or left and right asymmetry, the 63% average agreement rate is disappointing in terms of clinical significance, particularly since

cervical rotation seemed the most promising test in the prior study (Johnston *et al* 1982a).

It may be tempting to dismiss Johnston's low reliability as resulting from therapist inexperience, but Kaltenborn and Lindahl's (1969) 84% agreement rate was based on a group which included a variety of experience. In any case, poor results were also found in studies with experienced therapists (Clarkson 1982, Gonnella *et al* 1982).

Interpretation of the studies examining PPIVM test reliability is complicated further by the variety of rating scales, subjects and spinal levels used. Furthermore, agreement rates are difficult to compare directly because they are influenced by distributional properties including response base rates, which may vary across studies.

To facilitate comparisons of agreement rates we therefore expressed the results of the above studies in terms of Cohen's kappa (see Appendix). Since kappa expresses the proportion of obtained agreements relative to that expect to occur by chance, it facilitates comparisons across studies which employ different rating scales, test joints, or other methodological features which might alter the statistical properties of the therapists' responses. A second advantage is that it is a correlation-like index, which varies between zero and one (unless observed agreements are less than expected by chance). Using data presented in the published reports, we found kappas of 0.64 for Kaltenborn and Lindahl (1969), 0.37 for Johnston *et al* (1982b), 0.24 for Gonnella *et al* (1982) and 0.15 for Clarkson (1982). In general therefore the studies of PPIVM tests do not seem to yield a very good degree of intertherapist reliability, particularly within the framework of clinical requirements. In view of the variety of therapist backgrounds, subjects used (including symptomatic and nonsymptomatic) and other variables, this conclusion probably has good generalizability and concurs with the more recent of previous interpretations (Gonnella *et al* 1982; Johnston *et al* 1982).

Reliability of spinal mobility assessment using combined PAIVM and PPIVM tests

In addition to the investigations cited in the previous three subsections which have involved either PPIVM or PAIVM assessment, a number of studies reported in the literature have used combined assessment techniques to rate spinal mobility. Because of the combined nature of the assessment task utilized in these studies, it is not possible to separate the individual reliability of any one of the tests involved. However the studies outlined below provide some insights into therapist performance.

Jull (1978) reported a study which examined the intertherapist reliability of rating the mobility of the upper three cervical joints following PAIVM and PPIVM tests. Each therapist performed 81 tests ranking each joint on a five point scale with the extremes of 'hypermobile' and 'no movement'. A total agreement rate of 88% was claimed, which is highly encouraging. However, a number of methodological issues suggest that this agreement rate should be interpreted with caution. Given the relative infrequency of 'hypermobility' and 'no movement' ratings likely to occur in the population, the effective range of variability may have been somewhat reduced. Unfortunately, no data on the relative frequency of findings in each category were reported. Furthermore, several decisions came from a given spinal segment. This could have introduced further restrictions in the (*a priori*) subjective range of potential variation. Finally the generalizability of the data is limited by the fact that the smallest sample viable for an intertherapist reliability study was used: two therapists.

In a later report, Jull (1982) provided further evidence of intertherapist reliability for combined PPIVM and PAIVM tests of lumbar segments. Two therapists examined one subject on three successive occasions. The intersession interval was one day. The intertherapist reliability coefficient was 0.35, which has been interpreted to mean that 'examiners correlated highly' (Jull 1982, p.75). Although the result was significantly different from no correlation, in the statistical confidence sense, a reliability coefficient of 0.35 is not high. In fact, the majority of the variance in the observed scores is attributable to error when the coefficient is so low. A similar argument applies to the intersession reliability coefficient reported to be only 0.10.

In a further study, Jull and Lane (1983) published findings related to assessment of lumbar spinal mobility. A subsample of 20 normal subjects from a population of 100 males and 100 females with no history of back pain were examined. Postero-anterior accessory glide and all passive physiological movements were assessed in six intersegmental levels from T12/L1 to L5/S1. Each level was classified on a five point rating scale from 'hypermobile' to 'very stiff'. The retest agreement rate for the single participating therapist was 87.3%. Intertherapist agreement on a subsample of five subjects was reported to be 82.2% between the therapist and an independent observer. Once again, these high agreement rates should be interpreted with caution because limited sample variability will increase the agreement attainable by chance. On the basis of the averaged data published by Jull and Lane (1983) for their full population, we estimate that an agreement rate of 38% could have been typically expected to occur by chance. If the test subsample had consisted of only the younger subjects the chance agreement estimate would have been 61%. Using the chance agreement rate for the whole population we computed a Cohen's kappa for the retest agreements of 0.79 and for the intertherapist agreements of 0.71. Using the 61% estimate, kappa values would have been 0.67 and 0.54 respectively.

Grant (1980) examined lumbar spinal mobility in groups of dancers and non-dancer controls using a number of techniques including passive movement tests. Within the study, two observers performed twenty tests on five subjects rating lumbar levels on a four point scale from 'hypermobile' to 'very stiff' and an interobserver agreement rate of 90% was obtained. The actual frequency distribution of test findings was not included in the report nor was it indicated from which of the experimental groups the subjects were drawn. Therefore we did not proceed to estimate kappa, the more appropriate coefficient.

It should be pointed out that it was not the primary intention of Jull (1978, 1982), Jull and Lane (1983) or Grant (1980) to measure reliability of assessment *per se*, but only to determine the reliability of the therapists who performed assessments for the various studies. Consequently, the generalizability of these results is in all cases limited by the fact that absolute minimum numbers of therapists were involved in both retest and intertherapist trials. Furthermore it is not clear whether the judgements resulting from the several segments sampled from a given subject were statistically independent. Lack of independence could have artificially raised the estimate of reliability.

Jull and Bogduk (1985) examined the reliability of diagnosis of zygapophyseal joint disorders in a group of twenty patients attending a pain clinic because of cervical pain. A trained therapist stipulated the abnormal cervical level after a full subjective and objective examination, including passive physiological and accessory movements. To provide an objective criterion, medial branch blocks (Bogduk 1985) were used to selectively anaesthetize nerves supplying cervical joints. Perfect agreement between the diagnosis of the therapist and the medial branch block was obtained. A subsample of four subjects was independently examined by another manipulative therapist, with perfect agreement on the abnormal joint. The results of Jull and Bogduk (1985) might suggest that palpatory tests can perfectly diagnose the level to be treated. It should be noted however that the patient sample had severe pain, which was often irritable (Jull and Bogduk 1985, p.163) and that the manual assessment not only included pain reproduction, but also was conducted in the context of other information produced by a full objective and subjective examination. Although the authors claim that the pathological joints had such abnormal compliance features as 'limited range of motion', 'abnormal quality of resistance' and 'abnormal limitation to the movement' (Jull and Bogduk 1985, p.164), they also report that 'reproduction of pain was invariably associated with these abnormal qualities of movement' (p.164). On the basis of our experience with assessment of compliance features (low reliability) and pain (high reliability) an alternative hypothesis is indicated: that provocation and reproduction of pain was the key factor in reliable identification of the injured level. This interpretation seems preferable because it is more parsimonious, being consistent with both our group's results and those of Jull and Bogduk (1985).

IV Reliability in the Production of Therapeutic Passive Movement

The reliability with which therapeutic movement is produced has received no systematic investigation according to our reviews of the journal literature. The degree of intratherapist or intertherapist variation in production of passive movement is presumably an important factor, at least theoretically, since some descriptions of mobilization techniques do identify various grades and do recommend selective use according to various conditions, *eg* Maitland (1977). Until systematic empirical studies are conducted to assess the differences in therapeutic outcome due to different grades of mobilization, the actual importance of using selected grades of mobilization, or of the reliability with which they are produced, must remain a problem which is justified only theoretically or through clinical anecdote. Nevertheless, given the broad influence on clinical and educational practice which description of grades of mobilization have attained, the issue seems to require far greater attention than it has received to date.

However, the primary purpose for reporting here two pioneering studies (Banting 1982, Mitchell 1983) conducted in our laboratories on this issue is that the reliability with which selected grades of movement are produced is indirectly related to the reliability with which compliance is assessed. That grades of mobilization are related to assessment of compliance is clear from descriptions of clinical procedures (Maitland 1977). The link was even more explicit in the definitions used by Banting (1982) and Mitchell (1983) when instructing the therapists in their studies. Grade II mobilizations were defined as 'large amplitude movements to the point where R1 is just perceived, at a rate of two to three oscillation per second' (Banting 1982, Mitchell 1983). Grade IV mobilizations were defined as 'a small amplitude movement just up to and touching the end of available joint range' (Mitchell 1983). Again two to three oscillations per second was the recommended oscillation frequency.

To investigate reliability, both studies adopted the strategy of presenting several spinal levels from several individuals thus ensuring a variety of ranges and joint mobilities. The reproducibility of peak force of mobilization can then be examined within the frame of reference provided by the variations due to anatomical and individual differences. When the same levels, from the same subjects are examined, the intertherapist and retest correlations for peak force are then akin to the reliability coefficients for locating

R1 and R2 in range presented by other studies (Baker 1981, Wong 1981, Weeks 1982, Flint 1983), particularly given the explicit definitions used by Banting and Mitchell.

In both studies the force platform technique already described was used to assess the forces of mobilization while therapists performed central PAIVM. The output of the force platform was monitored by computer. This permitted calculation of peak force of mobilization for each oscillation, as well as of oscillation amplitude and frequency by means of the dynamic force measurement technique described earlier. The data on the latter two parameters is important to the wider issue of reproducibility of technique but is less directly relevant to the present theme. It is considered in detail elsewhere (Banting *et al* 1985).

Banting (1982) examined intertherapist reliability in seven physiotherapists with specialist postgraduate qualifications in manual therapy. The least experienced therapist had more than nine months of clinical practice since completion of the specialist qualification. The sample comprised graduates from schools in three different Australian States. Each therapist mobilized four premarked spinal levels (T11, T9, T7, T5) from each of four subjects using central PAIVM delivered with the pisiform technique. Each level was mobilised for 20 seconds. Among other parameters, the peak forces during a cycle were calculated and averaged for all the cycles of a trial. Scores from the 16 levels mobilized by each therapist were then used to compute pairwise intertherapist correlations. The mean intertherapist correlation was a very poor 0.22. In addition systematic biases were found between the seven therapists when the peak forces were averaged across the 16 spinal levels (Banting 1982). Two therapists showed a 'light touch' (7.6N and 9.8N), three were two to three times more forceful (14.5N, 16.3N, 20.6N) and two showed nine or more times that force (50.2N, 87.1N). An analysis of variance confirmed these differences to be statistically significant (Banting 1982).

Mitchell (1983) replicated and extended Banting's study. Subjects were eight experienced physiotherapists with specialist postgraduate qualifications in manual therapy. Each mobilized twenty spinal levels comprising T9, T11, L1, L3 and L5 from one female and three male volunteers with no history of back pain. The same twenty segments were mobilized again one week later. Thus the design assessed both intertherapist and test-retest reliabilities for Grade II and Grade IV movements. In order to maintain comparability all joints were pre-mobilized by the experimenter. Thus all therapists, including the starter, were dealing with previously mobilized spines.

Among other parameters, Mitchell (1983) calculated the peak force for each oscillation. Following the earlier study (Banting 1982), trial averages were computed, from which intertherapist and retest correlations were obtained. Mitchell confirmed that intertherapist reliability for Grade II movements was low (r = 0.25) and showed that this was also the case for Grade IV (r = 0.16). In addition he found poor test-retest reliability for both Grade II (r = 0.22) and Grade IV (r = 0.42).

Systematic biases were also evident in the data. The peak forces for Grade II when averaged over the twenty segments showed an intertherapist range from 2.2N to 46.7N on Day 1. Even the trimmed range, excluding the extreme therapists, was 13.0N to 30.2N. On Day 2 the range was 3.9N to 26.4N. Analysis of variance confirmed that there were significant differences between therapists and between days (Mitchell 1983). Similarly, for Grade IV, the intertherapist range was 150.9N to 329.3N on Day 1 and 89.2N to 222.4N on Day 2. Again analyses of variance confirmed that there were statistically significant differences between therapists and between days (Mitchell 1983).

The studies of Banting and Mitchell relate to those for R1 assessment in the case of Grade II movement peak forces and to those of R2 in the case of Grade IV peak forces. The findings show very good consistency. Thus for intertherapist reliability in locating R1 the comparison figures are 0.30 (Wong 1981), 0.25 (Weeks 1982) and 0.38 (Flint 1983). These confirm the Grade II results (r = 0.22, r = 0.25). The comparison figures for intertherapist reliability in locating R2 are 0.28 (Baker 1981) and 0.24 (Flint 1983), which seem to support Mitchell's Grade IV result (r = 0.16). The poor test-retest correlation obtained by Mitchell for Grade II (r = 0.22), is if anything, better than the low value obtained by Weeks over the same interval (r = 0.09). Therefore the results obtained by Mitchell and Banting reinforce the conclusion of poor reliability for estimation of spinal compliance during PAIVM.

V Discussion

An overview of the studies presented above suggests several patterns in the findings (*cf* also Table 1). In general, pain tests were more reliable than tests assessing features of compliance. This effect was obtained even when very similar testing techniques were used such as when PAIVM was used for both P1 and R1 assessment. A second feature of the results is the excellent reliability obtained with the SLR and FF tests for pain. The correlation coefficients (0.96-0.98) were superior to those obtained by P1 assessment with PAIVM (0.73). These differences are statistically significant. A third aspect is the consistent finding of superior test-retest reliability over intertherapist reliability. This is a common result in most fields of measurement. Before the clinical implications of these findings are considered it is appropriate to discuss some factors which may account for the obtained results.

Table 1:
Summary of reliability coefficients

Measure	*Test movement*	*Retest r(κ)*	*Inter-observer r(κ)*	*CI (95%)*	*Source*	*Comments*
ROM	PAIVM	.88			Collis-Brown	after correction for error in patient report.
	PAIVM	.86			Grisold	
	PAIVM		.78		McNeill	after correction for error in patient report.
R1	PAIVM		.30		Wong	
	PAIVM		.25		Weeks	
	PAIVM		.38		Flint	
	PAIVM	.46		17%	Weeks	intrasession
	PAIVM	.09				intersession
	PAIVM		.22		Banting	peak applied force during Grade II
	PAIVM	.22	.25		Mitchell	peak applied force during Grade II
R2	PAIVM		.28		Baker	
	PAIVM	.42	.16		Mitchell	peak applied force during Grade IV
P1	PAIVM	.73		34%	Collis-Brown	therapist location on VAS
	PAIVM	.83			Collis-Brown	measured force at patient report of P1
	PAIVM		.62		McNeill	therapist location on VAS
	PAIVM		.75		McNeill	measured force at patient report of P2
	SLR	.96		13.60°	McFarlane	
	SLR	.98			Puentedura	
	SLR	.97			Million *et al*	
	SLR	.96-.97	.93-.96		Lankhorst *et al*	intersession
	SLR		.78		Hoehler	passive test
	SLR	.95			Hoehler	active test
	FF	.98		83mm	Kwong	
	FF	.99			Munroe	
	FF	.91			Million *et al*	skin distraction
	FF	.95	.97		Lankhorst *et al*	intersession,skin distraction
	FF		.50		Hoehler	skin distraction

Factors which may account for the superior reliability of pain assessment

Although pain tests showed better reliability than tests of compliance features, there are procedural differences between the pain tests investigated. Therefore the comparison between PAIVM assessment of pain and compliance features is probably the most appropriate for discussion.

In order to locate P1 by PAIVM a physical stimulus is applied. The patient must sense and report pain onset, and the therapist must then relate that event to a point in ROM. In order to locate R1 a similar physical stimulus is applied, the therapist must sense the occurrence of the 'onset of resistance', then relate that event to a point in ROM. For both tests some of the total error will be due to stimulus application and some to the ability to locate a point in ROM. Thus the essential difference between the two judgemental processes is that tests of pain involve only one judgement, that of ROM, while tests of compliance require the judgement of both the compliance feature and ROM. It may appear therefore that the issue is simply a question of which of these contrasting perceptual processes contains more error. However, the quantitative theory of reliability shows clearly that reliability is a function of both error and true score variation (see Appendix). The same amount of error (in metric terms) means poorer reliability if the true score variation is small rather than large.

It is important to note that the low correlations obtained for R1 and R2 are at least in part due to the restricted range of true score variability. R1 tends to be restricted to the lower third of range while R2 tends to be restricted

Table 1:
Summary of reliability coefficients

Measure	*Test movement*	*Retest r(K)*	*Inter-observer r(K)*	*CI (95%)*	*Source*	*Comments*
P2	SLR	.96			Puentedura	
	FF	.98		33mm	Bruce	
	FF	.97		52mm	Patterson	intersession
	FF	.98		45mm	Patterson	intrasession
Compliance	PPIVM		(.64)		Kaltenborn	
	PPIVM		(.37)		Johnston (1982b)	
	PPIVM		(.24)		Gonnella *et al*	
	PPIVM		(.15)		Clarkson	
	Mixed	.35	.10		Jull (1982)	combined PPIVM and PAIVM
	Mixed	(.67-.79)	(.54-.71)		Jull and Lane	combined PPIVM and PAIVM
Level Selection	PAIVM	(.16)	(.11)		Millman	stiffest level
	PAIVM	(.34)	(.08) (1.00)		Allen Jull and Bogduk	level to be treated pathological level full objective and subjective examination

to the upper third. We re-examined the data of Wong (1981), Weeks (1982) and Flint (1983) to confirm this tendency. The standard deviation of R1 in ROM occupied respectively 8%, 8.3% and 7.9% of scale in the three studies, confirming the restricted variability of R1 in both normal and clinical populations and indicating very good consistency between the three independent studies. In contrast, the standard deviation of P1 was 23.2% of ROM in the Collis-Brown (1982) study. We have applied equation *A.21* (see Appendix) to compute what the obtained correlations would have been had the true score variability been the same as that observed by Collis-Brown for P1. The intrasession test-retest correlation of Weeks (0.46) becomes 0.83; the intersession correlation (0.09) becomes 0.25 and the intertherapist correlation obtained by Flint (0.38) becomes 0.76. It seems possible therefore to account for the poorer reliability of compliance feature assessment without suggesting that therapists perceive R1 or R2 more poorly than patients perceive P1 or P2. It seems rather that therapists face a more difficult discrimination problem when attempting to locate R1.

Factors which may account for the inferior reliability of passive intervertebral movement tests of pain

Passive intervertebral tests, whether 'accessory' or 'physiological', invariably yielded poorer reliability coefficients than those of the gross movement tests such as SLR and FF. To avoid the confounding contribution of pain versus compliance assessment, an appropriate comparison available for discussion is between FF or SLR tests of pain versus PAIVM assessment of pain.

In FF or SLR tests a gross 'physiological' movement provides the stimulus for pain elicitation, the patient must then perceive and report pain onset (or similar parameters) and ROM can be recorded via goniometry or measures of relatively large linear displacements. In PAIVM tests a more localized movement is the stimulus for pain elicitation, the patient must perceive and report pain onset (or similar parameters), then the therapist must through subjective evaluation of ROM, record where the pain occurred.

The issues for discussion therefore seem to be: the reliability of subjective ROM assessment by the therapist versus goniometric or similar methods for ROM assessment; and the reliability of pain elicitation by gross physiological movement versus localized PAIVM.

As might be expected, goniometric assessment is typically reported to show high reliability (Leighton 1955, Myers 1961, Boone *et al* 1978, Ekstrand *et al* 1982). However, the reliability of assessing ROM by palpation does not seem to have been previously investigated. Initial evidence that therapists do not introduce a very large amount of error at the stage of locating the P1 report in ROM was obtained by Collis-Brown (1982). His test-retest correlation when based upon force platform data, which does not involve therapist judgement of ROM, was 0.83. When based upon therapist determined data it was 0.73. Thus adding subjective

ROM assessment to the total process did not reduce reliability substantially. The degree of error due to patient report and force measurement technique is represented in the force test-retest correlation (0.83). It is possible to calculate what the test-retest reliability would have been if no error had arisen from these processes (see Appendix). This indirect estimate of test-retest reliability for locating a point in ROM was 0.88.

Additional evidence for high intra-therapist reliability of ROM assesment was obtained by Grisold (1983). Therapists were asked to palpate end of range of a single lumbar level using the pisiform technique. They were then asked to palpate one, two, three, four, five, six and seven eighths of range in a random order prescribed by the experimenter. This procedure was repeated eight times, varying the order of presentation of point in range each time.

The static force platform technique (see Section 1) was used to measure applied force for each of the 64 trials. Average test-retest correlations were 0.86, almost identical to the 0.88 computed from the data of Collis-Brown after correction for variation in patient report.

Nevertheless, although the ability of therapists to locate a point in ROM seems relatively high, particularly in consideration of the difficulty of the task, it is lower than that of goniometric and related techniques (0.96-0.98), thus accounting in part for the lower reliability of P1 assessment through PAIVM. That the assessment of ROM cannot be the full explanation for the superior reliability of the SLR and FF tests is clear from Collis-Brown's (0.83) retest correlations for applied force at P1. This coefficient is analogous to those derived from goniometric measurement during SLR test, or length measurement during FF tests. Our statistical analysis revealed that 0.83 was significantly lower than either 0.98 or 0.96. Thus some of the superiority in reliability exhibited by SLR and FF tests appears attributable to the second factor, *ie* the way pain is elicited.

Manual application of accessory movement seems to be more susceptible to random error than the application of physiological movement. Our evidence suggests that production of PAIVM is likely to contain significant error in comparison to the limited distribution of R1 and R2 over the ROM (Baker 1981, Wong 1981, Weeks 1982, Flint 1983). Biomechanical studies confirm the difficulty facing the therapists. Punjabe *et al* (1977) have measured 4mm displacement between lumbar vertebral bodies when forces of about 160N were applied in the anterior direction to the cephalad vertebra *in vitro*. Collis-Brown (1982) and McNeill (1982) measured maximum forces applied during PAIVM tests of about 350N. It is reasonable to assume that this load is equally distributed between the intervertebral joints on either side of the assessed level. This implies that similar loads (350/2 = 175N) were applied by the therapist to lumbar intervertebral joints during PAIVM as were applied in the *in vitro* studies of Punjabe *et al* (1977). Similar intervertebral displacement would therefore be expected in the two cases. The *in vitro* observations of Punjabe *et al* have been tentatively confirmed *in vivo* by Thompson (1983) who developed an apparatus for measuring applied load and relative intervertebral displacement simultaneously. The apparatus consisted of a proof-ring strain gauge through which force was applied centrally to a lumbar vertebra (L3). Two parallel linear-displacement transducers attached to the strain-gauge and adjusted to contact the spinous processes of vertebrae immediately above and below the loaded processes were used to measure relative displacement between L2 and L3 and between L3 and L4. Results for three subjects indicated that the caudad joint exhibited more displacement (3-5mm) than the cephalad joint (1-3mm) with applied loads of 250N. Again, if the assumption is made that this load is distributed equally between the intervertebral joints above and below, this represented a force of 125N at each joint. These data suggest that therapists are required to produce very small variations in displacement, sometimes by the application of large forces. Both factors seem conducive to poor performance.

In contrast to the difficulties presented to reliable stimulus production during PAIVM, the procedures of FF and SLR tests seem to be taking advantage of a naturally available system for amplification of joint movements. Anatomical evidence indicates that relatively gross physiological movements will produce very small intervertebral movements. For example, during forward flexion of the trunk, approximately the first 60° is accomplished by spinal structures alone. Farfan (1973) and Allbrook (1957) have shown that approximately 12° of this total is contributed by the L5-S1 joint and a further 12° by the L4-L5 joint. The remaining lumbar joints contribute about 7° each with the remainder distributed over the relatively immobile thoracic vertebrae. It is a commonly held view that for trunk flexion angles less than 60°, the lumbar joints contribute to the total in an amount proportional to their contribution to maximal flexion (although, we have been unable to find quantitative evidence which relates to this point). According to this model, as the trunk moves through 5°, the lower lumbar vertebral joints move through an angle of 1° and the higher joints through about 0.5°. At the same time, the shoulders, a distance of 0.5m away from the lumbar vertebral joints move through an arc length of about 4cm by comparison with the fractions of millimeters displacement at the joints themselves. The amplification effect is quite clear. Similar arguments pertain to structures affected by the SLR.

The implication is that effects well within the control of the therapist's motor skill (or in the case of active movement tests within the patient's motor skill) would produce quite small changes at the spine, thus improving

the signal to noise ratio of the manoeuvre.

Further possible problems with passive intervertebral movement tests

The argument has already been put that although therapists' ability to locate a point in ROM is reasonably good (r = 0.86-0.88), the narrow range of variation in compliance parameters places particularly high reliability requirements on the therapist, if the measures are to distinguish phenomena of interest. To appreciate the difficulty further, consider the results obtained by Weeks (1982) who demonstrated that an intrasession change of at least 17% of scale would have to occur in R1 for therapists to detect it with 95% confidence. Since R1 in the population probably varies over about a third of the scale according to the data of Baker (1981), Weeks (1982) and Flint (1983), intrasession changes exceeding half of the total range of individual differences in R1 would have to occur for reliable detection by the therapist. It is equivalent to requiring a joint which is in the lower quartile of R1 in the population to change to the upper quartile. This seems a very unlikely proposition. The detection of intersession change, or of absolute location in range for R1 or R2, provides an even bleaker picture.

Another aspect of the judgement task presented to therapists is identification of a specific point within ROM. In assessment of P1, this simply involves judgement of the current point in ROM at the time of patient report of pain. In assessments of compliance features this requires identification of the feature and subsequent estimation of the point in ROM at which this feature occurs. Identification of a feature seems to require that the feature exists in mechanical terms in order to provide a stimulus. It also seems to require that the feature be definable uniquely in terms of the therapists perceptions of 'joint feel'. The experiments of Banting (1982) and Mitchell (1983), in which therapists were required to perform mobilizations to a particular point in ROM, indicated wide variations in therapists' 'connotations' of R1 and R2, since vastly different forces were utilized to reach the same point in range on the same subject. In the textbook (Maitland, 1977) which established the nomenclature and theory in this field, we have been unable to find a precise *operational* definition of 'resistance'. Therapists with whom we have discussed this issue have not been able to reach consensus on a definition. A discussion of the distinctions between these definitions and their implications for the construction of the movement diagram are, however, beyond the scope of this review.

In the studies reported here specific features of pain or compliance were recorded on the two dimensional movement diagram. This two dimensional VAS helps clarify the therapist's assessment task and is recommended for summarizing and communicating clinical descriptions (Maitland 1977). It is of interest to examine the demands it makes upon the therapist. For example, the horizonal axis, which scales ROM, is defined by Maitland to represent 'any range of movement from the starting position at A to the limit of normal range at B. It makes no difference whether the movement depicted is small or large . . . Point B is always constant and always at the extreme of normal average range of passive movement' (Maitland 1977, p.317). This definition shows clearly that the therapist is not merely required to respond on a psychophysical scale according to current sensory input, a difficult enough task under the circumstances, but also has to make that scale relative to 'normal average range of movement'.

Several problems may be seen to arise from defining the scale relative to normal average range. First, the therapist is required to alter the scale in relation to past experience. This is likely to introduce a variety of biases (Kahneman *et al* 1982, Slovic *et al* 1977). Second, the therapist is apparently required to store many models of normality, since a different model will be required for different joints, different movements and perhaps other subsets as well, such as those generated by gender or age. This requirement places an even larger burden on memory. Third, the parameter for mental modelling is 'average normal range'. This seems rather vague, particularly since it requires statistical interpretation from the observer. Human intuitive perception of the statistical parameters of data samples suffers from several biases (Slovic *et al* 1977, Kahneman *et al* 1982). All of these factors are likely to increase the error of scaling. Nowhere in the clinical literature have we been able to discover evidence that therapists can in fact cope with such complexity of judgement. Our data, which consistently returned very poor intertherapist correlations, suggest that the task is too difficult.

A final problem, at least for the central PAIVM data reported above, may be seen to arise from the sensory information afforded by the technique. An essential value of passive movements to clinical theory seems to lie in the highly localized nature of their probing. As such the ROM of interest would appear to be only that which is relative to adjacent structures, rather than the overall movement through space described by the segment tested. However, consider the following statement from Maitland (1977, p.34): 'If the pressure is applied as a single slow pressure, the vertebral movement will not be appreciated at all; if it is applied too quickly it can only be interpreted as shaking. However, if the pressure is then relaxed and reapplied and repeated two or three times a second, the amount of movement which can take place will be readily appreciated'. As this statement indicates, the perception of relative movement relies not on direct sensation of displacement but on perception of phenomena which are not uniquely determined by relative displacement. As such, the movement diagram seems to place a burden of complex and undefined biomechanical

interpretation on the therapist, which will be conducive to the introduction of error. In fact when direct manual sensation of displacement relative to adjacent segments is not concomitantly undertaken, the situation we have usually observed to be the case during PAIVM tests, the movement diagram borders on being a biomechanical non-sequitur to PAIVM. An alternative VAS more directly defined in terms of the sensory experience of the performing therapist may be preferable.

Clinical implications

As neither of us is trained with a clinical background in manual therapy we wish to confine our comments to a series of questions which the psychometric and biomechanical evidence presented above seem to raise.

Tests of pain have generally been considered most important in the assessment procedure (Maitland 1977). The reviewed results suggest good to excellent reliability for this aspect of assessment.

However, the poor reliability shown by tests of vertebral compliance during passive movement raises several questions about their role in clinical practice. Presumably one of the major virtues of passive movement tests of compliance is that they help localize the pathology. However, are there no satisfactory substitutes for achieving this goal? It is not yet clear that these tests would be required, even if reliable, given the plethora of other case information, together with epidemiological knowledge. Jull and Bogduk's (1985) results interpreted in the context of those reported here, suggest that pain reproduction will very reliably select the level to be treated. How often do joint conditions present in the absence of pain? Furthermore, is precision in selection of level to be treated necessary? If there is no adverse effect associated with intervention at inappropriate levels, the additional resource cost involved would appear to be marginal, thus permitting a 'fail-safe' strategy to locality of intervention.

Another role for passive movement tests of compliance seems to be to aid in the selection of a direction and grade of movement. Again the question arises, could this decision be made on the basis of the other information? Furthermore, the research literature has yet to demonstrate that the grade of movement (in the respects defined by compliance features) selected is critical to clinical outcome. In any case, both inter and intratherapist reliability in application of movement grades was demonstrably unreliable. Could the mobilization procedure be made to be more reliant on patient comfort and particularly patient feedback rather than on manual reassessment following treatment? If so, there is ample literature in the experimental psychology of motor skills which suggest that performance with feedback tends to be superior (Sage 1977). Perhaps feedback-based treatment, utilizing pain report as feedback, is the *de facto modus operandi* and the intertherapist unreliability in the absence of pain merely confirms this.

A third role which might be attributed to passive movement tests of compliance is to evaluate progress. If reliability is the criterion for selecting tests of progress, then the evidence presented indicates clearly superior alternatives. The objection may be raised that localized compliance changes must be uniquely traced. However, the case that compliance changes *per se* are pathological or uniquely related to pathology has yet to be definitively outlined in the research literature.

A final role which might be attributed to passive movement tests of compliance in clinical decision strategy is that of confirmatory tests. A confirmatory test is undertaken to reassure that a decision taken on another test is adequate. This is a common, but often misused clinical strategy. If test A correctly predicts a criterion variable (*eg* pathology of a given type) on 80% of occasions and if test B does likewise, then the final probability of a 'confirmed' decision which is also a correct decision is actually 64%! This arises because 'confirmation' implies that both tests yield the same prediction, thereby invoking the multiplicative law of contingent probability. On 4% of occasions the tests will confirm each other, but be simultaneously wrong (0.20 x 0.20 = 0.04). On 16% of occasons test A will be correct, but test B will disagree (0.80 x 0.20 = 0.16) and on another 16% vice-versa, making a total of 32% of occasions containing difficult disagreements. These figures deteriorate if one of these two tests should have a lower percentage of valid predictions.

In conclusion therefore, the obtained results suggest that the assessment role of passive movement tests of compliance be seriously reconsidered, particularly PAIVM in its present form. If a case can be made that unique, essential information is provided by the passive assessment of compliance, and we reiterate that such a case has not yet been made in accordance with the rigors of empirical science, then it would seem that new methods of testing must be developed which achieve that purpose.

Present limitations and future directions

The conclusions drawn in this review must be understood in the light of the limitations imposed by the methodology of the studies providing the evidence for these conclusions. In particular, since we are reporting on an incomplete series of small studies, which of necessity must be limited in their sampling, several issues require discussion.

The number of therapists investigated in any one study was typically small. However most issues were addressed by more than one study and consistent results were obtained. Some of the studies reviewed here involved student manual therapists who had varying degrees of clinical experience in physiotherapy practice but who had not yet completed their specialist programme in manual therapy. They had

however completed and passed the unit relevant to the particular procedures assessed. Several points can be put forward to argue the case that poor results were not the effect of therapist inadequacy or inexperience. Firstly, it might be argued that student therapists had recently completed a period of very intensive clinical training and were in fact likely to perform better than practising therapists who used some of these techniques less frequently. Secondly, when studies utilizing student therapists were replicated with experienced therapists, no significant differences in results were obtained. Thirdly, motor learning research suggests that when learning occurs in the absence of exteroceptive feedback, variability about some mean performance is reduced but the average performance remains unchanged (Gibson 1969). It is conceivable that upon completion of a period of formal training the therapists no longer receive information about the correctness of skills employed in practice from a common source and are therefore continuing to learn in the absence of shared feedback. We might therefore expect improvements in test-retest reliability in more experienced therapists. However, because their experience may have been individualized, it is possible that there will be no change, or even a deterioration, in intertherapist reliability. Finally, we could argue that the therapists selected represent a cross section of practising therapists and therefore represent the general level of therapeutic skills. We are unaware of factors which could have biased the samples toward the 'poor' therapists; in fact, in some cases, efforts were made to involve the more respected and established members of the therapeutic community.

In addition to 'type of therapist' other variables were sampled. These include anatomical location and assessment technique. Although cervical and thoracic segments were sampled in some studies the lumbar segments were observed much more frequently. The data from non-lumbar segments collected so far does not suggest that significantly better results for PAIVM tests of compliance will be obtained in these segments. Finally, it should be clear that since all studies reported are about spinal joints, no statement can be made about reliability in the assessment of peripheral joints. Clearly this aspect requires further investigation, particularly because substantial differences exist between spinal and peripheral joint assessment. For example, in peripheral joints goniometry is more readily applicable with current techniques. Furthermore, a contralateral joint is available for simple comparison in peripheral joints. Contralateral comparisons in spinal joints, when they are appropriate, seem rather more complex because both joints belong to the affected level.

In the assessment of spinal-joint compliance a number of interesting reliability comparisons remain to be conducted. The PAIVM data collected to date is limited to central PAIVM. The reliability of unilateral PAIVM seems deserving of investigation since among other differences to central PAIVM, a contralateral comparison of sorts is available. In addition most of the evidence collected so far relates to PAIVM technique. The reliability of PPIVM tests, particularly in the cervical spine also seems to deserve further investigation. In PPIVM the stimulus movement at the spine may be more controllable than in PAIVM because of the mechanical advantage argument invoked above in the discussion of the superior reliability of SLR and FF tests of pain. This could be particularly so for cervical movement where the therapist has a more manageable structure than the trunk. Furthermore, unlike PAIVM tests, during PPIVM tests the therapist is required to *directly* palpate the *relative* movement of adjoining segments in addition to sensing the force required to produce that movement.

A number of lines of research are also suggested by the results obtained. For example, we have already mentioned that the poor reliability of passive assessment of compliance features indicates that their contribution to the overall clinical decision process should be carefully assessed. Our group has taken some initial steps in that direction (Cunningham 1982, Walker 1984). If compliance assessment proves in the future to be essential to clinical decision-making more reliable tests will need to be developed. It may be necessary to develop instrumented approaches to this problem. Thompson's study (1983) is a first step in that direction in our laboratories. In any case, such instrumentation will be required if adequate surveys of spinal joint compliance are to be completed in order to provide the normative data currently missing from the scientific literature of manual therapy. The poor reliability found for production of selected grades of movement further strengthens the requirement to investigate the dependence of clinical outcome on particular grades of mobilization, a requirement initially posed by the apparent absence of formal study on this central issue in some approaches to mobilization.

Manual therapy, at this point in its development, is in the position of having developed to the stage of a complex clinical theory well in advance of a sound base of verifiable, empirical data. It should be clear from the foregoing that even within the narrow aims selected by the respective investigators a great deal remains to be done. It is our hope that the above will prove to be a seminal contribution in the field of clinical arthrometrics.

Acknowledgement

We are indebted to our clinician colleagues who have been a constant source of inspiration and feedback; to the many therapists who have so generously given their time, skills and other resources; and above all to our students, without whose work this would not have been possible. We hope they will continue to confront the unknown with curiosity, rationality and constructive hard work.

References

Allbrook D (1957), Movements of the lumbar spinal column, *The Journal of Bone and Joint Surgery*, **39B**, 339-345.

Allen D (1983), The reliability of determining the most abnormal lumbar joint using passive accessory intervertebral movements. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Anderson JAD and Sweetman BJ (1975), A combined flexi-rule/hydrogoniometer for measurement of lumbar spine and its sagittal movement, *Rheumatology and Rehabilitation*, **14**, 173-179.

Bach TM (1985), An indirect method for measuring forces applied during therapeutic intervention and assessment techniques. Unpublished manuscript.

Baker M (1981), Interobserver reliability of range impairing stiffness ratings obtained from passive accessory intervertebral movements. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Banting J (1982), Intertherapist reliability in the performance of a grade II mobilization movement. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Banting JB, Mitchell WN, Bach TM and Matyas TA (1985), Reliability in the execution of selected grades of mobilization in manual therapy. Unpublished manuscript.

Boone DC, Azen SP, Lin CM, Spence C, Baron C and Lee L (1978), Reliability of goniometric measurement, *Physical Therapy*, **58**, 1355-1360.

Breig A and Troup JDG (1979), Biomechanical considerations in the straight-leg-raising test, *Spine*, **4**, 242-250

Bogduk N (1985), A scientific approach to cervical diagnosis, *Proceedings of the Australian Physiotherapy Association Conference*, Brisbane.

Bruce P (1981), The test-retest reliability of physiological movement as a method of assessing range of movement to the level of pain tolerance. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Clarkson M (1982), Intertherapist reliability in assessing stiffness ratings in the lumbar spine obtained from passive physiological intervertebral movements. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Cohen J (1960), A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, **20**, 37-46.

Collis-Brown GL (1982), Test retest reliability of pain onset ratings obtained from passive accessory intervertebral movements. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Cunningham G (1982), Clinical decision making in manipulative therapy: the effect of antecedent information on palpation findings. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Cyriax J (1982), *Textbook of Orthopaedic Medicine*, Vol. 1. (9th ed.), Baillière Tindall, London.

DePalma AF and Rothman RH (1970), *The Intervertebral Disc*, WB Saunders, Philadelphia.

Edwards AL (1964), *Statistics for the Behavioural Sciences*, Holt, Rinehart and Winston, New York

Ekstrand J, Wiktorsson M, Oberg B and Gillquist J (1982), Lower extremity goniometric measurements: a study to determine their reliability, *Archives of Physical Medicine and Rehabilitation*, **63**, 171-175.

Farfan HF (1973), *Mechanical Disorders of the Low Back*, Lea and Febiger, Philadelphia

Fleiss JL, Cohen J and Everitt BS (1969), Large sample standard errors of kappa and weighted kappa, *Psychological Bulletin*, **72**, 323-327.

Flint R (1983), Intertherapist reliability for the assessment of joint measurement behavior by means of passive accessory intervertebral movements (PAIVMs). Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Gibson E (1969), *Principles of Perceptual Learning and Development*, Appleton-Century-Crofts, New York.

Goddard MD and Reid JD (1965), Movements induced by straight leg raising in the lumbosacral roots, nerves and plexus, and the intrapelvic section of the sciatic nerve, *Journal of Neurology, Neurosurgery and Psychiatry*, **28**, 12-18

Gonnella C, Paris SV and Kutner M (1982), Reliability in evaluating passive intervertebral motion, *Physical Therapy*, **62**, 436-444.

Grant R (1980), Lumbar sagittal mobility in hypermobile individuals. *Proceedings of the Manipulative Therapy Association of Australia*, Adelaide.

Grisold PM (1983), Estimating range of movement from passive accessory intervertebral movements; the nature of the scale and the reliability of performance. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Guilford JP (1954), *Psychometric Methods*, McGraw-Hill, New York, chs 13, 14.

Hanley EN, Matter RE and Frymoyer JW (1976), Accurate roentgenographic determination of lumbar flexion-extension, *Clinical Orthopaedics and Related Research*, **115**, 145-148.

Hart FD, Strickland D and Cliffe P (1974), Measurement of spinal mobility, *Annals of Rheumatic Diseases*, **33**, 136-139.

Hartman DP (1977), Considerations in the choice of inter-observer reliability estimates, *Journal of Applied Behavior Analysis*, **10**, 103-116.

Hoehler FK and Tobis JS (1982), Low back pain and its treatment by spinal manipulation: measures of flexibility and asymmetry, *Rheumatology and Rehabilitation*, **21**, 21-26.

Hollenbeck AR (1978), Problems of reliability in observational research, in GP Sackett (Ed.), *Observing behaviour, Vol 2: Data collection and Analysis Methods*, University Park Press, Baltimore.

Hubert L (1977), Kappa revisited. *Psychological Bulletin*, **84**, 289-297.

Johnston WL (1982), Passive gross motion testing: Part I. Its role in physical examination, *Journal of the American Osteopathic Association*, **81**, 298-303.

Johnston WL, Elkiss ML, Marino RV and Blum GA (1982a), Passive gross motion testing: Part II. A study of interexaminer agreement, *Journal of the American Osteopathic Association*, **81**, 304-308.

Johnston WL, Beal MC, Blum GA, Hendra JL, Neff DR and Rosen ME (1982b), Passive gross motion testing Part III Examiner agreement on selected subjects, *Journal of the American Osteopathic Association*, **81**, 309-313

Jull G (1978), Clinical observations of upper cervical mobility, *Proceedings of the Inaugural Congress of the Manipulative Therapy Association of Australia*, Sydney

Jull G (1982), Passive intervertebral movements of the lumbar spine, in Toward a better understanding of spinal pain, *Proceedings of the Manipulative Therapy Association of Australia Annual Conference*, Brisbane

Jull GA and Lane MB (1983), Aspects of lumbar spine mobility in a normal population, in KD Bower (Ed.) *International Conference on Manipulative Therapy Proceedings*, Perth

Jull GA and Bogduk N (1985), Manual examination: An objective test of cervical joint dysfunction, *Proceedings of the Australian Physiotherapy Associaton Conference*, Brisbane

Kahneman D, Slovic P and Tversky A (Eds) (1982), *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge

Kaltenborn F and Lindahl O (1969), Reproducibility of the results of manual mobility testing of specific intervertebal segments, *Lakartidningen* (Swedish Medical Journal), **66**, 962-965

Kapandji IA (1974), *The Physiology of the Joints, Vol 3*, (2nd ed.) Livingstone, Edinburgh.

Kwong HF (1981), Test-retest reliability of pain onset assessed by active 'physiological' movements. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Lankhorst GJ, Van de Stadt RJ, Vogelaar TW, Van der Korst JK and Prevo AJH (1982), Objectivity and repeatability of measurements in low back pain, *Scandinavian Journal of Rehabilitative Medicine*, **14**, 21-26.

Leighton JR (1955), Instrument and technic for measurement of range of joint motion. *Archives of Physical Medicine and Rehabilitation*, **36**, 571-578.

Loebl WY (1967), Measurement of spinal posture and range of spinal movement, *Annals of Physical Medicine*, **9**, 103-110.

Macfarlane A (1981), Test-retest reliability of straight leg raise as determined by pain onset. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences. School of Physiotherapy, Melbourne.

Macrae IF and Wright V (1969), Measurement of back movement,*Annals of Rheumatic Diseases*, **28**, 584-589.

Maitland GD (1977), *Vertebral Manipulation*, (4th ed.), Butterworth, London.

McNeill KE (1982), Intertherapist reliability of pain onset from passive accessory intervertebral movements. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences. Schoolof Physiotherapy, Melbourne.

Meyers LS and Grossen NE (1974), *Behavioral Research: Theory, Procedure, Design*. WH Freeman and Co., San Francisco, 164-166.

Milhon R, Hall W, Haavik Nilsen K, Baker RD and Jayson MIV (1982), Assessment of the progress of the back pain patient, *Spine*, **7**, 204-212.

Millman AJ (1981), Test-retest reliability of range impairing stiffness ratings obtained from passive accessory intervertebral movements. Unpub-

lished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne

Mitchell WN (1983), Reliability in the performance of Grade II and Grade IV mobilizations. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Moll J and Wright V (1976), Measurement of spinal movement, in M Jayson (Ed.), *The Lumbar Spine and Back Pain*, Sector, London

Moran HM, Hall MA, Barr A and Ansel BM (1979), Spinal mobility in the adolescent, *Rheumatology and Rehabilitation*, **18**, 181-185.

Munro R (1983), The contribution of pain versus memory for position and exteroceptive feedback in the forward flexion test. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Murphy RW (1977), Nerve roots and spinal nerves in degenerative disc disease, *Clinical Orthopaedics and Related Research*, **129**, 46-57

Myers H (1961), Range of motion: Part I — introductory review of literature, *Physical Therapy Reviews*, **29**, 195-205.

Nunally JC (1978), *Psychometric theory*, (2nd ed.), McGraw-Hill, New York.

O'Keefe PJ (1981), The spinal compliance tester. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Patterson S (1982), The test-retest reliability of lumbar flexion when limited by pain. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiothrapy, Melbourne.

Puentedura L (1983), The effects of trunk position on straight leg raise in normal subjects. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences. School of Physiotherapy, Melbourne.

Punjabe MM, Krag MH, White AA and Southwick WO (1977), Effect of preload on load displacement curves of the lumbar spine, *Orthopaedic Clinics of North America*, **8**, 181-192.

Reynolds PM (1975), Measurement of spinal mobility: a comparison of three methods, *Rheumatology and Rehabilitation*, **14**, 180-185.

Sage GH (1977), *Introduction to Motor Behavior: A Neuro-psychological Approach*, (2nd ed.), Addison Wesley, Reading, Massaechusetts, ch 20.

Slovic P, Fischhoff B and Lichtenstein S (1977), Behavioral decision theory, *Annual Review of Psychology*, **28**, 1-39.

Stoddard A (1980), *Manual of Osteopathic Technique*, (3rd ed.), Hutchinson, London.

Thompson R (1983), Measurement of relative intervertebral displacement in the lumbar spine during application of a PAIVM. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Troup JGD, Hodd CA and Chapman AE (1967), Measurements of the sagittal mobility of the lumbar spine and hips, *Annals of Physical Medicine*, **9**, 308.

Twomey LT and Taylor JF (1979), A description of two new instruments for measuring the ranges of sagittal and horizontal plane motions in the lumbar region, *Australian Journal of Physiotherapy*, **25**, 201-203.

Van Adrichem JA and Van Der Korst JK (1973), Assessment of the flexibility of the lumbar spine, *Scandinavian Journal of Rheumatology*, **2**, 87-91.

Walker D (1984), A survey of treatment selection and subjective certainty at different stages of clinical assessment. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Weeks PM (1982), Test-retest reliability of stiffness onset using passive accessory intervertebral movements. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Wong M (1981), Interobserver reliability of stiffness onset ratings obtained from passive accessory intervertebral movements. Unpublished Postgraduate Diploma Dissertation, Lincoln Institute of Health Sciences, School of Physiotherapy, Melbourne.

Appendix

Reliability theory is a highly developed field with ample presentation of its concepts (Guilford 1954, Edwards 1964, Nunally 1978). This appendix will only review selected issues of interest to a number of the studies reported in this review. A knowledge of basic statistical theory (mean, variance, correlation, statistical inference) is assumed in the following discussion.

The reliability of a measurement process refers to the dependability, or reproducibility of observed scores when these are obtained from measurements of the same events. Realiability classically relates the extent to which observed scores represent the true values of the events measured. Equation *(A.1)*, where $X_o =$ observed score $X_t =$ true score and $E =$ error component, shows that the observed value can be represented as being partly composed of true quantity and partly error.

$$X_o = X_t + E \qquad (A.1)$$

If X_t is known the discrepancy of X_o readily quantifies the error. The larger E is the more unreliable is the observation.

Two patterns of error can occur: systematic error, such that E is constant; and random error, such that E is unpredictably variable from measurement to measurement. If the error is constant, then observed scores will be the same across several measurements of a given true value. The instrument is therefore not considered unreliable under these circumstances. Constant error does affect the truth of the absolute value, but the difference between two observed scores will be equal to the true score difference. However, if the error is random, measurements will vary unpredictably even when the same true value is under observation. The quantitative theory of reliability is concerned therefore with random error.

Since E may vary from one occasion of measurement to the next, a consequent problem is how to summarize the 'typical' size of E. Furthermore, the interest usually lies in describing how reliable an observation process is for a variety of objects which lie on a common dimension, rather than in describing the reliability for measuring only one object. This also requires the definition of a method for indexing the 'typical' value of error. Thus in estimating error, a sample of values is usually generated. Hence the issue of 'typical' error is a problem in sampling theory and the associated descriptive statistics.

If a sample consisting of one measurement of several objects is taken, then each score could be expressed as a deviation from the sample mean rather than in raw score units. Equation *(A.2)* then follows from *(A.1)*:

$$x_o = x_t + e \qquad (A.2)$$

where $x_o = X_o - \bar{X}_o$, the deviation of the observed raw score from the mean of the observed scores; $x_t = X_t - \bar{X}_t$, the deviation of the true score from the mean of the true scores; and $e = E_o - \bar{E}$ the deviation

of the error component from the mean of the error components.

Since the problem is to obtain a measure of 'typical' amount of random error, the deviation scores could be averaged over the sample. However, if error is random, there will be just as much positive deviation as negative deviation, yielding a misleading average of zero. To overcome this, statisticians deal with squared deviation, which has the effect of removing the algebraic sign. The mean squared deviation score will not average to zero. In deviation score units the average may be obtained as follows:

$$x_o^2 = (x_t + e)^2 \qquad (A.3)$$

Hence, $$x_o^2 = x_t^2 + e^2 + 2x_t e) \qquad (A.4)$$

Summing over the sample and dividing by the number of cases yields the averages:

$$\frac{\Sigma x_o^2}{n} = \frac{\Sigma(x_t^2 + e^2 + 2x_t e)}{n} \qquad (A.5)$$

That is, $$\frac{\Sigma x_o^2}{n} = \frac{\Sigma_t^2}{n} + \frac{\Sigma e^2}{n} + \frac{\Sigma 2x_t e}{n} \qquad (A.6)$$

Since the error is randomly positive and negative in equal quantity, over the total sample $\frac{\Sigma 2x_t e}{n}$ will tend to be zero, as in the earlier argument.

Therefore, $$\frac{\Sigma x_o^2}{n} = \frac{\Sigma x_t^2}{n} + \frac{\Sigma e^2}{n} \qquad (A.7)$$

The average squared deviation from the mean is known as the variance, or s^2. Therefore, the variance of observed scores is composed of true variance plus error variance:

$$s_o^2 = s_t^2 + s_e^2 \qquad (A.8)$$

Since s_e^2 is the amount of squared error (in deviation units) per case it seems to be an adequate measure of 'typical' error.

However, there are several drawbacks to using s_e^2 as the sole index of reliability. One is that the units of error are squared, which makes interpretation awkward. This is easily resolved by defining the squared root of s_e^2 to be the 'standard error of measurement':

$$s_e = \sqrt{s_e^2} \qquad (A.9)$$

This index is more readily interpreted and is commonly cited.

Frequently the interest lies in measuring change from one occasion to another. In these situations each of the two measurements will introduce some error. If d_o is the observed difference score in deviation units, x_o is the observed deviation score on the second occasion and e' is the random error in deviation units, then it follows from *(A.2)* that:

$$d_o = x_o - x_o$$
$$= (x_t - x_t) + (e - e')$$

If the true difference in deviation units is

$$d_t = (x_t - x_t)$$

then:

$$d_o^2 = [d_t + (e - e')]^2$$
$$= d_t^2 + (e - e')^2 + 2d_t(e - e')$$
$$\therefore \Sigma d_t^2 = \Sigma d_t^2 + \Sigma(e - e')^2 + \Sigma 2d_t(e - e')$$

Note that $\Sigma 2d_t(e - e') = 2d_t \Sigma(e - e')$ and $\Sigma(e - e') = \Sigma e - \Sigma e'$. Since both e and e' are random within (and between) the measurement samples, then $\Sigma e = \Sigma e' = 0$ and $\Sigma(e - e') = 0$ (within the limits of sampling error) following the earlier argument. Thus $\Sigma 2d_t(e - e') = 0$ and:

$$\Sigma d_o^2 = \Sigma d_t^2 + \Sigma(e - e')^2$$
$$= \Sigma d_t^2 + (\Sigma e^2 + \Sigma e'^2 + 2ee'$$

Again, since e and e' are random, the positive and negative components will be equal (within the limits of sampling error). Thus $\Sigma 2ee' = 0$ and:

$$\Sigma d_o^2 = \Sigma d_t^2 + (\Sigma e^2 + \Sigma e'^2)$$

If n is the number of pairs of observed scores, then:

$$\frac{\Sigma d_o^2}{n} = \frac{\Sigma_t^2}{n} + \frac{(\Sigma e^2}{n} + \frac{\Sigma e'^2)}{n}$$

$$\therefore s_{do}^2 = s_{dt}^2 + (s_e^2 + s_{e'}^2)$$

Consequently the error of measuring change will be larger than the error for measuring on either occasion. The standard error of measuring changes $(s_{e\ diff})$ will be:

$$s_{e\ diff} = \sqrt{s_e^2 + s_{e'}^2} \qquad (A.10)$$

The standard error of measurement however is a measure of 'typical' error. The error will sometimes be less, sometimes more. Most often, it is assumed that error is variable in both direction and magnitude, with small errors more probably than large errors. Although situations may arise where other assumptions are better, it is unusual to imagine that the errors around a true value are normally distributed. Thus the mean of a sample of observed values of the same event will be the best estimate of the event's true score. If the errors around this true value follow the assumed normal distribution it is possible to calculate over what range some specified proportion of observed values will fall. This statistic is known as the confidence interval *(CI):*

$$CI_{(1-\alpha)} = \bar{X}_o \pm Z_\alpha s_e \qquad (A.11)$$

where *(1 - a)* is the confidence level and Z_α is the appropriate value from the normal distribution. An analogous equation can be written for difference scores by substituting $s_{e\ diff}$ for s_e. The virtue of transforming a standard error into confidence intervals is that it acknowledges the error to be variable and permits calculation of the proportions of observations which will occur within some given error range, or vice-versa. It thus more completely models the error of measurement.

Another drawback to both s_e^2 and s_e is that they are metric bound indexes. That is, standard errors of various measures are not readily comparable: different approaches to measurement must often be compared; measurement units are sometimes arbitrary; the comparative reliability of measurement in different fields is an issue at times. In these cases a unit-free index of reliability is preferable. Percentages or proportions are often used to resolve such a problem. From *(A.8)* it follows that:

$$\frac{s_t^2}{s_o^2} + \frac{s_e^2}{s_o^2} = 1 \qquad (A.12)$$

Thus s_t^2/s_o^2 is the proportion of observed score variance due to true score variance and s_e^2/s_o^2 is the proportion due to error score variance. The former may be defined as a coefficient of reliability. For a perfectly errorless measurement method $s_e^2 = 0$. Thus $s_t^2 = s_e^2$ and the reliability coefficient s_t^2/s_o^2 will be 1. As s_e^2/s_o^2 increases so the reliability diminishes. For a measurement method which is maximally errorful all the observed variance is error variance, *ie* $s_e^2 = s_o^2$. In this case $s_t^2 = 0$ and the coefficient of reliability will be zero.

How then to estimate s_e and its associated statistics in practice? Clearly one way might be to measure a set of events whose values are known, then calculate s_o^2 and s_t^2 from observed and known values. From this, s_o^2 and its derivatives $s_{e'}$, $s_{e\ diff'}$, the reliability coefficient and various confindence intervals could be obtained.

Unfortunately, in practice, particularly in new fields of measurement, this is often impossible, since the true values are not known. However, although the true values are not known, it can be safely assumed that if a variety of events are measured, some variation in true scores should occur. If these events are measured again the initial values should be exactly reproduced provided there is no error. To the extent that there is random error the relative position among the initial observations will not be reproduced.

It should be noted that failure to reproduce scores can result from several processes. Instruments or observers may be unstable over time (test-retest unreliability). Measures taken by two observers may differ (interobserver unreliability). Measures taken by two versions of the same instrument or test may differ (parallel form unreliability). In measurements composed of several observations a part-score from a subset of the observations may be compared to a part-score based on another subset (internal consistency). These are all different practical methods for obtaining two estimates of the same underlying true value. Although the error introduced in attempting reobservation by different methods are likely to be different, all these practical approaches to establishing reliability have in common the need to quantify the degree to which one set of observations predicts another set of observations of the same events.

The correlation coefficient *r* (Edwards 1964) is a measure of the degree to which one data set predicts another. If the sample of events is remeasured (*eg* on another occasion, or by another observer), then equation *(A.13)* relates the observed scores Y_o the true scores Y_t and the error E:

$$Y_o = Y_t + E \qquad (A.13)$$

All of equations *(A.2)* — *(A.12)* can be rewritten for these second measurements. Since reliability can be defined as the extent to which measurements predict remeasurements of the same events, the correlation between X and Y will be an index of reliability. The correlation coefficient is defined as the average cross-product of the standardized score on X and Y:

$$r = \frac{\Sigma Z_x Z_y}{n} \qquad (A.14)$$

We will assume that the reader is already familiar with the theory of correlation, which indicates how this index relates to scattergrams; and how it varies between 0 (when X and Y are randomly related) and 1.0 (when X, Y coordinates plot perfectly on a straight line).

An algebraically equivalent equation for r can be written in deviation scores since $Z_x = (X - \bar{X})/S_X$ and $Z_y = (Y - \bar{Y})/S_y$:

$$r = \frac{\Sigma x_o\ y_o}{\sqrt{\Sigma x_o^2 \Sigma y_o^2}} \qquad (A.15)$$

If e_x and e_y are the error deviation scores for X and Y respectively, then:

$$\begin{aligned}\Sigma xy &= \Sigma(x_t + e_x)\ (y_t + e_y) \\ &= \Sigma x_t y_t + \Sigma x_t e_y + \Sigma y_t e_x + \Sigma e_x e_y\end{aligned}$$

Since e_x and e_y are random (with positive and negative values equivalent and randomly paired to particular x_t's or y_t's) it follows that $\Sigma x_t e_y = 0$, $\Sigma y_t e_x = 0$, and also that $\Sigma e_x e_y = 0$. Thus $\Sigma s_o y_o = \Sigma x_t y_t$. Since the same events are being remeasured $x_t = y_t$ and therefore

$$\Sigma x_o y_o = \Sigma x_t^2 = \Sigma y_t^2 \qquad (A.16)$$

Furthermore, since the same events are being measured twice, within the limits of sampling error the two samples X and Y should have the same variance, $s_x^2 = s_y^2$. The

variance being the average squared deviation it follows that $\Sigma x_o^2 / N = \Sigma y_o / n$. That is:

$$\Sigma x_o^2 = \Sigma y_o^2 \qquad \text{(A.17)}$$

From equations *(A.16)* and *(A.17)* equation *(A.18)* may be rewritten:

$$r = \frac{\Sigma x_t^2}{\Sigma x_o^2} = \frac{\Sigma y_t^2}{\Sigma y_o^2}$$

Dividing both nominators and denominators by n defines r in terms of variances:

$$r = \frac{\Sigma x_t^2/n}{\Sigma x_o^2/n} = \frac{\Sigma y_t^2/n}{\Sigma y_o^2/n}$$

$$r = \frac{s_t^2}{s_o^2} \qquad \text{(A.18)}$$

where s_t^2 = true score variance and s_t^2 = observed score variance. Equation *(A.18)* allows the very important conclusion that the correlation between two measures of the same sample of events is in fact the reliability coefficient defined from equation *(A.12)*.

This conclusion not only enhances the interpretation of reliability and its evaluation in practice, but also permits evaluation of the other useful index of reliability s and its derivative the confidence interval. From equation *(A.12)* it follows that $1 - r = s_e^2 / s_o^2$. Thus $s_e^2 = s_o^2 (1 - r)$ and therefore:

$$s_e = s_o \sqrt{1 - r} \qquad \text{(A.19)}$$

Note that both r and s_o are readily calculated from experimental data.

The preceding theory outlines the rationale and interpretational basis of the major classical indexes for quantifying reliability: the standard error of measurement *(s_e)*; the reliability coefficient *(r)*; and the confidence interval *(CI)* around the true score (or around the change score). A number of further conclusions are derivable from the foregoing. An entire exposition of these is beyond the scope of this contribution. However two aspects are important to arguments presented in the main text.

One aspect is that the reliability coefficient is sensitive to the amount of true score variance. If the true score variance is for some reson restricted the reliability coefficient will be reduced provided the error of measurement remains constant. This conclusion follows from equations *(8)* and *(18)*. Since $r = s_t^2/s_o^2$ and $s_o^2 = s_t^2 + s_e^2$ then:

$$r = \frac{s_t^2}{s_t^2 + s_e^2} \qquad \text{(A.20)}$$

Imagine an experiment where a blindfolded human subject is required to palpate 10 cubes, the sides of which vary in 5mm steps from 10mm sides to 55mm. The cubes are then repalpated. The subject is required to judge their size on both occasions and a reliability coefficient is calculated. Conversely imagine the same experiment with 10 cubes varying in 1mm steps from 20mm to 29mm. Since the same palpatory technique is employed on similar events the random error of measurement in metric terms e should remain comparable (within the limits of sampling variation). Thus s_e is assumed constant across the two experiments. However the true score variance will be larger in the first experiment with cubes ranging from 10mm to 55mm. It follows from equation *(A.20)* that if s_t^2 diminishes when s_e^2 remains constant, then the ratio r will diminish also. It is therefore important that reliability studies use stimuli with a range of variability which is representative of the events to which the instrument or observational procedure will be ultimately applied. If a range restriction does occur a correction is available:

$$R = \frac{r(S/s)}{\sqrt{1 - r^2 + y^2(S/s)}} \qquad \text{(A.21)}$$

In equation *(21)*R = correlation for uncurtailed distribution, S = standard deviation of uncurtailed distribution, r = correlation of the curtailed distribution, s = standard deviation of the curtailed distribution.

Another conclusion derived from reliability theory which is relevant to the main text is that the observed correlation between two variables will be less than the theoretically possible correlation between their true scores. This occurs because both variables are measured with some random error. If the reliability coefficients for both variables are known, the theoretically possible relationship between the two variables when measured without error can be calculated (2). If X and Y are the two variables, $r_{x_t y_t}$ = the correlation between the true X and Y scores, $r_{x_o y_o}$ = observed correlation between X and Y, r_{xx} = the reliability coefficient for measuring X and r_{yy} = the reliability coefficient for measuring Y, then:

$$r_{x_t y_t} = \frac{r_{x_o y_o}}{\sqrt{r_{xx} r_{yy}}} \qquad \text{(A.22)}$$

The preceding discussion is concerned with the theory of reliability as it applies to variables measured on interval or ratio scales such as might occur in goniometry. Often clinical measurement is categorical in nature, such as when rating abnormality, or when rating the stiffness of a joint along a five point scale. A reliability theory needs to be defined for these situations also.

A frequently employed measure for describing test-retest or interobserver reliability of categorical data is the per-

centage of agreement between the two sets of observations. The rationale is similar to that of the reliability coefficient: the presence of error will reduce agreement. Although simple and widely used, percent agreement has some deficiencies.

Even if measurement is totally unreliable, that is if the observed categories arose at random, there will be some degree of agreement. Furthermore, that degree will be influenced by the distribution of measurements which arise from random processes. These distributions are different in different circumstances.

For example, the number of categories in the scale will influence randomly obtained agreement levels. On a two point scale, if both responses are equiprobable the expected agreement rate is 0.50. On a four point scale, if all responses are equiprobable, the expected agreement rate is 0.25.

In addition, the assumption of equiprobability may be inappropriate. If the incidence of the two middle categories in the four category example was 0.4 for each and if the incidence for the two extreme categories was 0.1 for each, then basic probability theory indicates the percentage of expected agreement (P_e) would be P_e = (.1 x .1) + (.1 x .1) + (.4 x .4) + (.4 x .4), that is 0.34. The two distributions of marginal probability, furthermore, need not be identical as in this example. Nevertheless, basic probability theory can readily yield expected proportions of agreement under the random model.

Another factor which can influence the proportion of agreement under a random model is the definition of agreement. Let *A, B, C, D* be the four categories of the above rating. Let *a, b, c, d* be the proportion of ratings in the respective categories obtained on the first round of measurements and let *a', b', c', d'* be the corresponding proportions on the second round. If agreement is defined as not only the conjuction of identical ratings, but also of adjacent ratings, then elementary probability theory concludes that $P_e = aa' + ab' + ba' + bb' + bc' + cc' + cd' + dc' + dd'$. In the above example $a = a' = 0.1$, $b = b' = 0.4$, $c = c' = 0.4$, $d = d' = 0.1$. Therefore $P_e = 0.82$, which is substantially larger than 0.34, the result obtained with the stricter agreement rule.

To overcome such disadvantages Cohen (1960) defined the statistic kappa:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \qquad \text{(A.23)}$$

Kappa expresses observed agreement P_o relative to expected agreement. It also expresses that difference as a proportion of the distance between random and perfect agreement. Thus kappa is very similar to the reliability coefficient. If P_o is 100%, then $\kappa = 1$; if $P_o = P_e$, then $\kappa = 0$. As P_o exceeds P_e so kappa grows. Although the analogy between k and r is limited, a number of problems are practically resolved by this statistic. The probability distribution of kappa has been investigated (Fleiss and Cohen 1969, Hubert 1977) and it is a method with relevance to a wide variety of reliability problems when categorical data is encountered (Hartman 1977, Hollenbeck 1978). Other correlation-like statistics, such as Φ, are applicable to problems of association in categorical data, but a full discussion of their relative values is beyond the scope of this appendix.