0008-3194/88/133–138/$2.00/©JCCA 1988

Several strategies for evaluating the objectivity of measurements in clinical research and practice

Joseph C Keating, Jr, PhD*

The objectivity (interexaminer reliability) of measurement in chiropractic provides a basis for judging the quality of information in clinical research and practice. Objectivity may be determined by formal measurement evaluation studies and by sampling within clinical trials. Interpretation of inter-examiner reliability requires descriptive and inferential statistics selected on the basis of the mathematical properties of data, appreciation of the clinical meaning of a particular measure, and recognition of the role of chance. Methods of data analysis include scatter-plots, contingency tables, time-series graphs, and correlational and concordance coefficients. Many kinds of objectivity evaluations are well within the capacities of private practitioners and student clinicians. (JCCA 1988; 32(3): 133–138)

KEY WORDS: measurement, objectivity, reliability, chiropractic.

L'objectivité (fiabilité entre examinateurs) de mesure en chiropractie fournit une base pour juger de la qualité de l'information en recherche et pratique cliniques. L'objectivé peut être déterminée par des études sérieuses d'évaluation de mesure et par un échantillonnage lors d'essais cliniques. L'interprétation de la fiabilité entre examinateurs requiert des statistiques descriptives et obtenues par déduction choisies sur la base des propriétés mathématiques des données, l'appréciation de la signification clinique d'une mesure particulière et la reconnaissance du rôle de la chance. Les méthodes d'analyse des données incluent les mosaïques, les tables de contingence, des graphiques de séries chronologiques et des coéficients de corrélation et de concordance. De nombreux types d'évaluations d'objectivé peuvent très bien être effectuées par des praticiens privés et des étudiants cliniciens. (JCCA 1988; 32(3) 133–138)

MOTS CLÉS: mesure, objectivité, fiabilité, chiropractie.

Although the art of chiropractic is more than 90 years old, only recently have chiropractors of any number begun to critically evaluate their methods of healing. Clinical trials of manipulative and conservative therapeutics are a priority in chiropractic, and investigators must be concerned with the quality of clinical measurements in these experiments. The credibility (internal validity) of clinical trials depends importantly upon the accuracy of clinical data.

One indicator of the accuracy of clinical measurements is objectivity or inter-examiner reliability. Objectivity refers to the degree of association or correlation between observations obtained by 2 or more examiners (Es.). By comparing the findings of independent (blinded) Es, the credibility of the measurements can be increased. Low inter-observer reliability indicates that at least one of the Es has not accurately observed the target phenomenon, while high reliability suggests that observers *may be* accurately describing a clinical variable. Discrepancies between paired observations may suggest the degree and potential sources of measurement error, and provide a basis for evaluating change within clinical trials. Concern for the objectivity of measurement is reflected in such familiar (non-research) strategies as the second opinion and the independent medical examination.

Some inter-examiner reliability data are available to support and refute the objectivity of some of the clinical measurements of special interest to the chiropractor. These include motion palpation of the spine[1–4], measures of the leg differences[5,6], x-ray marking systems[7,8], inclinometer and goniometric ranges of motion[9–13], thermographic measures[6,14], and measures of bilateral symmetry[6,15], among others.

Objectivity data in chiropractic are far from thorough, however. Many kinds of chiropractic measurements and assessments have rarely or never been evaluated for between-doctor reliability, for example, motion palpation of the thoracic spine, cervical goniometry, manual muscle testing, and paraspinal electromyography. Moreover, the available objectivity literature often suffers interpretative limitations due to small samples of Es, incorrect or incomplete statistical analyses, and insufficient replications. Additionally, most objectivity data have been generated by formal measurement evaluation studies, and not within the context of clinical trials.

Measurement evaluation studies

Clinical researchers have employed several strategies to assess the inter-observer reliability of measurements in clinical reports. The most common is the formal evaluation of reliability between observers. DeBoer et al.[3] provide a good example of this tactic in their study of palpatory evaluation of the cervical spine. In this project three chiropractors performed palpatory examinations of 40 subjects (Ss). Each observer rated abnormality (presence or absence of fixation, muscle tension, and pain) at each cervical segment. (Subjects were also examined a second time by the same three examiners in order to assess test-retest as well as inter-observer reliability). The abnormality scores for cervical regions from each doctor was then compared with scores from each doctor by means of the Kappa statistic[16,17], a procedure for determining the degree of concordance (agreement) beyond chance among pairs of Es.

Results[3] suggested that, at least among these patients and doctors, palpatory examination of the lower cervical spine is fairly objective, agreement on findings in the upper cervical spine is less reliable, and that little significant agreement is

* Associate Professor and Director of Research, Northwestern College of Chiropractic, Bloomington, Minnesota 55431
(612) 888-4777
Research Consultant, Stockton Foundation for Chiropractic Research

found on palpatory examination of the mid-cervical spine. Mior et al.'s[4] subsequent study of inter- and intra-examiner agreement for palpatory findings in the upper cervical spine provides a partial replication of DeBoer et al.'s[3] initial efforts. Unfortunately, Mior et al.[4] did not determine the probability associated with their inter-examiner Kappa coefficient (K = 0.15, $p > 0.05$/not significant), and inappropriately concluded that cervical motion palpation in their study was objective. However, if further evaluations reconfirm the reliability of palpatory findings in the lower cervical spine[3], future investigators may develop confidence in these measures as methods of identifying and monitoring spinal dysfunctions, and will cite these studies as the sources of their confidence.

Another example of a formal measurement evaluation is provided by Keeley et al.[9], who studied the test-retest and inter-examiner reliability of inclinometer measurement of lumbar ranges of motion (gross motion, hip motion, right and left straight leg raising) in chronic low back pain patients and non-patients. Of 16 inter-rater correlation coefficients (Pearson's "r"), only two were less than $r = 0.90$ (hip motion: $r = 0.74$ and $r = 0.82$), which suggests that inclinometer goniometry may be very accurate for *some* measurements. Keeley's data follows a pattern found in goniometric studies of lumbar and other joints[11,13,18-20]. However, Keeley's group employed only two physical therapists to make measurements, and so the generalizability (external validity) of their findings is minimal. Their work requires (and deserves) replication with other goniometric evaluators and Ss.

Measurement evaluation studies such as these are important because they take a focused and critical look at assessment strategies which may be useful in clinical trials, or which may already be employed by practitioners. However, there is no certainty that a measurement procedure shown to be objective in one or more evaluation studies will always be so. For example, if unrepresentative Ss (e.g. normals) are inadvertantly included in a patient sample, the measurement may not be objective when employed in the field, or in a sample recruited for a clinical trial. Practitioners may then inappropriately generalize from published measurement evaluations to the clinical situation. Similarly, different doctors may subtly vary in their measurement procedures, and so vary in objectivity. And, in an ultimate sense, complete confidence is never justified when extrapolating from one sample of Ss to another, since generalization of results (inference) always involves some degree of wishful thinking. While this dilemma is true for all sciences, it is particularly germane in the clinical disciplines, where controls are often much weaker than in laboratory research, and variability among Ss is usually much greater.

Objectivity sampling

Clinical trials have as their goal the investigation of the usefulness of rehabilitative, therapeutic and preventive interventions. To accomplish this goal investigators must be convinced of the meaningfulness and accuracy of their observations. In addition to measurement evaluations, some researchers have sought to sample the objectivity of data collected within the context of the clinical trial. Typically, this strategy involves concurrent measurement by at least two Es during some or all (100% sample) of the observations made[21]. For example, Schnelle and co-workers[2], in a trial of behavioral strategies designed to reduce urinary incontinence among nursing home patients, employed two observers during 17% of baseline (pretreatment) and 10% of treatment phase recordings. They computed a %-agreement index for dual observer recordings, and assumed that it was representative of inter-observer agreement throughout the trial. While this assumption is not strictly justified (since an observer may subtly alter the observation or measurement procedure when she/he knows a reliability check is being made), we may usually have more confidence in generalizing from sampled to unsampled observations within a study vs. between studies. If Schnelle et al.[22] had relied upon prior evaluation studies of observation systems similar to theirs, confidence in the validity of their dramatic treatment effects (increase in appropriate toileting) would be weaker.

Paired observations do not guarantee the accuracy of measurements in clinical trials, since several threats to measurement reliability may not be readily detectable. For example, if "observer drift"[23,24] or "instrument decay"[25] are relatively constant between doctors across a series of observations, no change in reliability indices may be noticed, although the accuracy of measurement has changed. Nonetheless, reliability sampling within clinical trials provides a practical, if imperfect, tactic for evaluating the objectivity of clinical measurements and the "internal validity"[25-27] of clinical experiments.

Evaluating reliability data

Four statistical procedures for evaluating the objectivity of chiropractic clinical measurements are noted here: 1) linear correlational methods, 2) confidence intervals, 3) concordance coefficients, and 4) time-series analyses.

1 *Linear correlation*

Correlational statistics (e.g., Pearson's "r", intraclass correlation, Spearman's "rho") provide an index of the linear relationship between 2 or more variables (e.g., cervical ranges of motion in degrees). Several conditions must be met in order for linear correlational techniques to be appropriate for evaluation of objectivity: a) the data must be rank-ordered, and preferably, interval or ratio data, b) the distributions of scores should be approximately normal, and c) each set of paired observations must be independent of every other set, so as to avoid serial dependency. Linear correlational statistics provide an index of the extent to which the scores of one observer can predict the scores of a second observer, and permit calculation of regression equations and confidence intervals.

Correlational data may also be graphically displayed in scatter-plots to permit visual estimation of the linear association between paired scores. Scatter-plots have been infrequently employed in published reports, perhaps owing to journal space limitations. However, visual inspection of a scatter-plot can aid in interpreting the correlation coefficient (see Figure 1). Moreover, linear correlational methods such as Pearson's "r" can mask a significant non-linear relationship, but the scatter-plot could reveal the need for further analysis. Four hypothetical scatter-plots of paired goniometric measurements are presented in Figure 1.

Pearson's linear correlational coefficient (r) provides an index of the strength of linear covariance between paired scores. The coefficient varies from $r = -1.0$ (a perfect inverse relationship; not depicted) to $r = 1.0$ (a perfect positive linear relationship; see Figure 1a). As the value of Pearson's r approaches 0, the linear relationship weakens. The significance of a given coefficient may be determined by reference to a table of critical values of r[28].

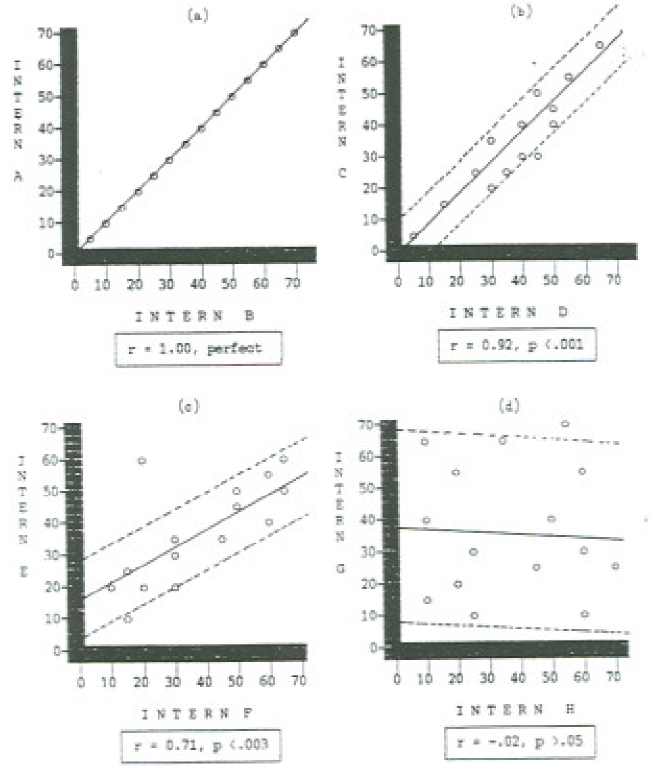


Figure 1 Scatter-plots of paired goniometric measurements of cervical flexion in 4 samples of patients measured by 4 pairs of interns. Solid lines indicate regression of Y (vertical axis) on X (horizontal axis); broken lines indicate limits of 95% confidence interval. These hypothetical data illustrate: a) a perfect linear relationship between scores from Interns A and B, b) a strong, but imperfect relationship between scores from Interns C and D, c) a weaker relationship between scores from Interns E and F, and d) a near-zero relationship between scores from Interns G and H. The greater the difference between each score and a regression line (line of best fit), the greater the error between intern pairs.

2 *Confidence intervals*

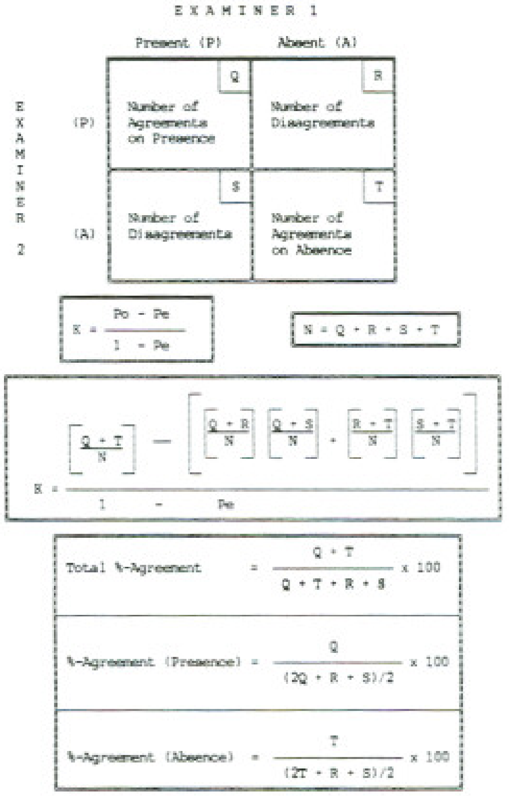
In contrast to correlation coefficients, which indicate the strength of linear association between 2 variables, computation of standard errors of estimate (SEoE) and the construction of confidence intervals provide indices of inter-observer error which are expressed in the original units of measurement[29]. The SEoE[28] provides a measure of the dispersion of error around a regression line, and is analagous to the standard deviation around a mean. A 95% confidence interval may be constructed by defining an area which is 4 standard errors wide (± 2 SEoE) around the regression line. This interval defines the extent of error which must be tolerated in order to be accurate 19 out of 20 predictions from one observer to another. Confidence intervals are depicted in Figure 1.

An alternative means of constructing confidence intervals involves predetermining a maximum allowable error between examiners on the basis of clinical judgement[29] or prior data[7], and determining the percentage of paired observations within the error limit. The advantage of this method is its ease of computation, so long as the minimum acceptable %-agreement for parametric data is high (e.g., the traditional 95%). Lesser %-agreements are suspect, since chance agreements are usually unknown. Percent-agreement intervals may provide a useful complement to methods based on the normal distribution (e.g., SEoE), but are usually inappropriate if not accompanied by some estimate of chance agreement.

3 *Concordance statistics*

The linear correlational coefficient is not appropriate for some types of data in chiropractic. For example, clinical observations in the healing arts frequently involve dichotomous choices such as presence vs. absence of fixations, bedwets, or organic pathology. In these instances the correlation coefficient does not apply, since neither rank-ordered nor interval data are involved. Rather, the objectivity evaluation should seek to determine whether pairs of observers can agree (beyond chance) on the presence vs. absence of some phenomenon.

Unfortunately, many objectivity studies have inadequately analyzed concordance for categorical data. Bergstrom and Courtis[30], for example, determined the %-agreement for presence and %-agreement for absence of palpable fixations at each of five lumbar joints in 100 chiropractic students examined by two interns. They concluded that their findings "demonstrated a high inter-examiner reliability", but made no effort to determine chance levels of agreement. In fact, they reported modest agreement on presence (60-88%) and weak agreement on absence (12-38%). Since they did not provide a contingency table for paired dichotomous choices (see Figure 2), it is not possible to estimate whether their results could be due to chance. Since the percentages they obtained are determined in part by the actual frequency of occurrence of the target phenomenon (see Figure 3), they are flawed as indicators of inter-examiner reliability. No firm conclusions about the objec-



		EXAMINER 1 Present (P)	EXAMINER 1 Absent (A)
EXAMINER 2	(P)	Q Number of Agreements on Presence	R Number of Disagreements
	(A)	S Number of Disagreements	T Number of Agreements on Absence

$$K = \frac{Po - Pe}{1 - Pe}$$

$$N = Q + R + S + T$$

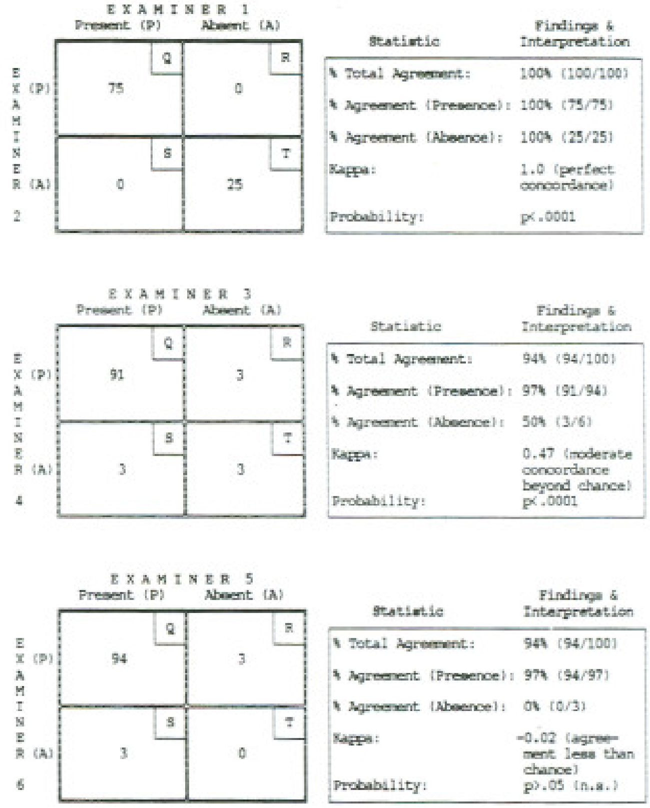
$$K = \frac{\left[\frac{Q + T}{N}\right] - \left[\left[\frac{Q + R}{N}\right]\left[\frac{Q + S}{N}\right] + \left[\frac{R + T}{N}\right]\left[\frac{S + T}{N}\right]\right]}{1 - Pe}$$

$$\text{Total \%-Agreement} = \frac{Q + T}{Q + T + R + S} \times 100$$

$$\text{\%-Agreement (Presence)} = \frac{Q}{(2Q + R + S)/2} \times 100$$

$$\text{\%-Agreement (Absence)} = \frac{T}{(2T + R + S)/2} \times 100$$

Figure 2 Contingency table illustrating 4 possible outcomes resulting from 2 doctors' (observers', examiners') dichotomous choices: Agreement on Presence, Disagreement, Agreement on Absence. Two versions of the Kappa formula are given below the contingency table; Q, R, S and T refer to the frequencies in the cells. Kappa (K) is equal to [the proportion observed (Po) minus the proportion expected (Pe)] divided by [one minus the proportion expected]. Lastly, computation is shown of %-agreement statistics derived from the contingency table.



		EXAMINER 1 Present (P)	EXAMINER 1 Absent (A)
EXAMINER 2	(P)	Q 75	R 0
	(A)	S 0	T 25

Statistic	Findings & Interpretation
% Total Agreement:	100% (100/100)
% Agreement (Presence):	100% (75/75)
% Agreement (Absence):	100% (25/25)
Kappa:	1.0 (perfect concordance)
Probability:	p<.0001

		EXAMINER 3 Present (P)	EXAMINER 3 Absent (A)
EXAMINER 4	(P)	Q 91	R 3
	(A)	S 3	T 3

Statistic	Findings & Interpretation
% Total Agreement:	94% (94/100)
% Agreement (Presence):	97% (91/94)
% Agreement (Absence):	50% (3/6)
Kappa:	0.47 (moderate concordance beyond chance)
Probability:	p<.0001

		EXAMINER 5 Present (P)	EXAMINER 5 Absent (A)
EXAMINER 6	(P)	Q 94	R 3
	(A)	S 3	T 0

Statistic	Findings & Interpretation
% Total Agreement:	94% (94/100)
% Agreement (Presence):	97% (94/97)
% Agreement (Absence):	0% (0/3)
Kappa:	-0.02 (agreement less than chance)
Probability:	p>.05 (n.s.)

Figure 3 ·Three hypothetical distributions of paired dichotomous choices. In the first contingency table perfect concordance between examiners is illustrated. In the second contingency table high agreement on the presence, and moderate agreement on the absence of the target phenomenon yield moderate and significant concordance beyond chance. In the third example the examiners' inability to agree on absence of the target phenomenon, despite high agreement on presence, yields less concordance overall than would be expected by chance.

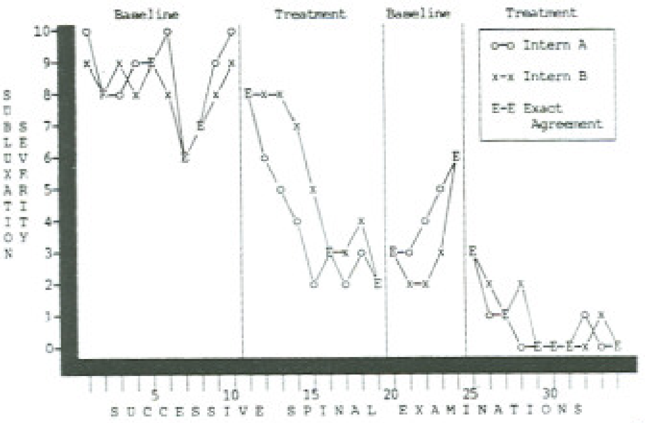
tivity of passive motion palpation of the lumbar spine can be drawn from their report.

The Kappa statistic[16,17] provides a partial solution to the problem of interpreting concordance between paired dichotomous choices. The Kappa coefficient provides an index of the extent of agreement between observers' dichotomous choices beyond the agreement expected by chance[31]. Agreement greater than chance is noted when Kappa exceeds zero ($K > 0$); perfect concordance is indicated by $K = 1$. Rosner[17] details a method of calculating the standard error of Kappa, and thereby, the probability associated with a given Kappa coefficient (i.e., its significance). Boline et al.[1] provide an example of the combined use of Kappa and %-agreement statistics in evaluating three methods of lumbar palpation. DeBoer et al.[3] provide a chiropractic example of a modified (weighted) Kappa statistic used to evaluate three-option concordance between pairs of doctors. Maclure and Willett[32] review a number of limitations in the use of Kappa.

4 *Time-series data*

An infrequently used but practical method of judging and reporting objectivity data also involves graphic display of paired observations. This strategy is especially suitable in time-series research[24]. An example is provided in Figure 4.

Graphic display permits the research consumer to judge the agreement of primary and calibrating observers at key points throughout the clinical trial. This information is obscured by correlational and concordance statistics, which yield an index based on a summary of many observations. Moreover, graphic display of paired observations requires little statistical sophistication. By encouraging paired observations by chiropractors, this tactic may also increase the validity of observation, since a second, calibrating observer may detect a target abnormality or event overlooked by a sole observer. Additionally, when ratio or interval data are employed, visual inspection permits judgement of the absolute differences between observers' findings at examination, and thereby provides a basis for judging the magnitude of therapeutic effects. For example, in the hypothetical data presented in Figure 4, the maximum discrepancy between examiners was only half the amount of change from the initial baseline to the final treatment phase. The credibility of clinical improvement over the course of the trial is enhanced by the relatively parallel observations by this pair of examiners.



Difference:	1 0 1 1 0 2 0 0 1 1	0 2 3 3 3 0 1 1 0	0 1 2 2 0	0 1 0 2 0 0 0 1 1 0	All Phases
n (per phase):	10	9	5	10	34
Mean (per phase):	0.7	1.4	1.0	0.5	.88
Range (per phase):	0-2	0-3	0-2	0-2	0-3
%-Exact Agreement:	40	33	40	60	44
%-Agreement (± 1):	90	56	60	90	76
%-Agreement (± 2):	100	66	100	100	91

Figure 4 Hypothetical example of paired observations of 2 interns who are blinded to one another's findings, blinded to the type of treatment, phase of study, treating doctor, patient history and records. In this imaginary case the doctor's treatment seems to produce a reliable change in level and trend of "subluxation severity", as recorded by the 2 spinal examiners. Visual inspection of these data suggests acceptable levels of objectivity. The absolute differences between examiners, and %-agreement statistics are shown below.

Combining descriptive and inferential statistics

Although some chiropractic objectivity studies may be criticized for over-reliance on descriptive statistics (e.g., %-agreement) and underutilization of inferential significance tests, the appropriate use of descriptive methods can be very helpful, both for the researcher and the research consumer. Scatter-plots of ratio and interval data can serve as preliminary screens of linear vs. non-linear trends between examiners[33], and can illustrate the meaning of reliability coefficients and error estimates for readers. Similarly, concordance coefficients such as Kappa should be accompanied by contingency tables. Moreover, in some cases descriptive statistics may provide the only evaluative strategy. For example, although one is tempted to apply Pearson's correlational technique to the paired observations reported in Figure 4, this would violate the requirement for independence of scores. The resulting Pearson's coefficient would probably overestimate the relationship between paired observations. Journal space limitations may often discourage publication of some kinds of graphic data presentations. It is recommended, however, that more frequent and judicious use of descriptive methods to be encouraged among chiropractic reseachers and journals.

A buddy system for more objective chiropractic data

The pace and quality of chiropractic clinical research could be increased if chiropractors, assistants, interns and pre-clinical students were to pair off for inter-examiner reliability studies. Doctors in group practice could team up, as might the solo-practitioner and the chiropractic assistant (or receptionist, depending on qualifications and the type of observation, measurement, or rating). Similarly, interns might conduct paired examinations with one another, or might supervise a small group of pre-clinical students who conduct paired, blinded observations/measurements of the interns' patients.

Research buddies could share responsibilities (e.g., literature search, method development, descriptive and inferential statistical analyses, writing the final report, seeking publication). A buddy system for learning spinal examination skills, with quantitative feedback via graphic and inferential statistical analyses, might promote skills development, and allow trainees to more carefully and critically examine the variety of assessment strategies in chiropractic. Students' appetites for rigorous, quantitative clinical data might be whetted through shared "hands on" experience and discussion. The trainee's attention could be research-oriented throughout the initial skills-acquisition phase of chiropractic clinical training.

Given the paucity of objectivity data for many routine clinical assessment techniques, quantitative evaluation of inter-examiner reliability by interns and doctors can easily provide important and original contributions to the literature in many cases. Reliability studies can also serve to shape and improve the clinical art. Further, by enabling practitioners to increase confidence in the objectivity of their examination processes, another impediment to chiropractic clinical trials can be reduced.

Conclusions and recommendations

Chiropractic clinical researchers have conducted inter-rater reliability studies for some kinds of measurements. However, strategies for evaluating the objectivity of measurement are not widely appreciated in the profession. As economic and interprofessional political pressures to scientifically evaluate healthcare methods grow, the chiropractic profession's need to collect and interpret objective measurements will increase.

Many of the assessment methods employed by chiropractors and others (physical therapists, orthopedists) have not been biometrically evaluated, and replication attempts have similarly been uncommon. Efforts to evaluate and improve the objectivity of chiropractic analysis and diagnosis should include measurement evaluation studies and objectivity sampling in clinical trials. Chiropractors should invest, individually and collectively, in greatly expanded training in clinical biometry.

Acknowledgements

I am grateful for the assistance provided by Dr. Grace E. Jacobs and Dr. Chuck Sawyer in the preparation of statistical materials for this paper.

Based on a paper presented at the Conference on Research and Education of the California Chiropractic Association and the Pacific Consortium for Chiropractic Research, June 13–14, 1987, Monterey, California.

References

1 Boline PD, Keating JC, Brist J, Denver G. Interexaminer reliability of palpatory evaluation of the lumbar spine. Am J Chiropractic Med 1988 (Mar); 1(1) 5–11.

2 Love RM, Brodeur RR. Inter- and intra-examiner reliability of motion palpation for the thoracolumbar spine. J. Manip Physiol Ther 1987; 10(1): 1–4.

3 DeBoer KF, Harmon R, Tuttle CD, Wallace H. Reliability study of detection of somatic dysfunctions in the cervical spine. J Manip Physiol Ther 1985; 8(1): 9–16.

4 Mior SA, King RS, McGregor M, Bernard M. Intra and interexaminer reliability of motion palpation in the cervical spine. JCCA 1985; 29(4): 195–8.

5 DeBoer KF, Harmon RO, Savoie S, Tuttle CD. Inter- and intra-examiner reliability of leg-length differential measurement: a preliminary study. J Manip Physiol Ther 1983; 6(2): 61–6.

6 Addington FA. Reliability and objectivity of anatometer, supine leg length test, thermoscribe II, and dermo-therm-o-graph measurements. Upper Cervical Monograph 1983; 3(6): 8–11.

7 Sigler DC, Howe JW. Inter- and intra-examiner reliability of the upper cervical x-ray marking system. J Manip Physiol Ther 1985; 8(2): 75–80.

8 Phillips RB, Frymoyer JW, MacPherson BV, Newburg AH. Low back pain: a radiographic enigma. J Manip Physiol Ther 1986; 9(3): 183–7.

9 Keeley J, Mayer TG, Cox R, Gatchel RJ, Smith J, Mooney V. Quantification of lumbar function, Part 5: reliability of range-of-motion measures in the sagittal plane and in vivo torsion rotation measurement technique. Spine 1986; 11(1): 31–5.

10 Mayerson NH, Milano RA. Goniometric measurement reliability in physical medicine. Archives Phy Med Rehab 1984; 65: 92–4.

11 Fitzgerald GK, Wynveen KJ, Rheault W, Rothschild B. Objective assessment with establishment of normal values for lumbar spinal range of motion. Phys Ther 1983; 63(11): 1776–81.

12 Defibaugh JJ. Measurement of head motion. Part II. An experimental study of head motion in adult males. J Am Phys Ther Assoc 1964; 44(3): 163–8.

13 Zachman Z, Traina AD, Keating JC, Bolles S, Porter LK. Interexaminer reliability and concurrent validity of two instruments for the measurement of cervical ranges of motion. J Manip Physiol Ther, accepted for publication 2/88.

14 DeBoer KF, Harmon R, Chambers R, Swank L. Inter- and intra-examiner reliability study of paraspinal infrared temperature measurements in normal students. Res Forum 1985b; 2(1): 4–12.

15 Vernon H. An assessment of the intra- and inter-reliability of the posturometer. J Manip Physiol Ther 1983; 6(2): 57–60.

16 Feinstein, R. A., Clinical epidemiology: The architecture of clinical research. Philadelphia: WB Saunders, 1985.

17 Rosner B. Fundamentals of biostatistics, Second Edition. Boston: PWS Publishers, 1986.

18 Reynolds PMG. Measurement of spinal mobility: a comparison of three methods. Rheumatology and Rehabilitation 1975; 14: 180–5.

19 Frost M, Stuckey S, Smalley LA, Dorman G. Reliability of measuring trunk motion in centimeters. Phys Ther 1982; 62(10): 1431–7.

20 Boone DC, Azen SP, Lin CM, Spence C, Baron C, Lee L. Reliability of goniometric measurements. Phys Ther 1978; 58(11): 1355–60.

21 Kelly MB. A review of the observational data-collection and reliability procedures in the Journal of Applied Behavior Analysis. J Appl Behavior Anal 1977; 10(1): 97–101.

22 Schnelle JF, Traughber B, Morgan DB, Embry JE, Binion AF, Coleman A. Management of geriatric incontinence in nursing homes. J Appl Behavior Anal 1983; 16(2): 235–41.

23 Kazdin AC. Artifact, bias, and complexity of assessment: the ABCs of reliability. J Appl Behavior Anal 1977; 10: 141–50.

24 Kratchowill TR, Wetzel RJ. Observer agreement, credibility, and judgement: some considerations in presenting observer agreement data. J Appl Behavior Anal 1977; 10(1): 133–9.

25 Campbell DP, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

26 Keating JC, Seville J, Meeker WC, Lonczak RS, Quitoriano LA, Dydo M, Leibel DP. Intrasubject experimental designs in osteopathic medicine: Applications in clinical practice. J Am Osteopath Assoc 1985; 85(3): 192–3.

27 Keating JC, Giljum K, Menke JM, Lonczak RS, Meeker WC. Toward an experimental chiropractic: time-series designs. J Manip Physiol Ther 1985; 8(4): 229–38.

28 Bruning JL, Kintz BL. Computational handbook of statistics. Glenview: Scott, Foresman and Co., 1987.

29 Keating JC, Boline PD. Letter to the Editor. J Manip Physiol Ther 1988; 11(1): 53–6.

30 Bergstrom E, Courtis G. An inter- and intraexaminer reliability study of motion palpation of the lumbar spine in lateral flexion in the seated position. European Journal of Chiropractic 1986; 34: 121–41.

31 Hartmann DP. Considerations in the choice of interobserver reliability estimates. J Appl Behavior Anal 1977; 10(1): 103–16.

32 Maclure M, Willett WG. Misinterpretation and misuse of the Kappa statistic. Am J Epid 1987; 126(2): 161–9.

33 Jackson BL, Barker W, Bentz J, Gambale AG. Inter- and intra-examiner reliability of the upper cervical x-ray marking system: a second look. J Manip Physiol Ther 1987; 10(4): 157–63.